The Information Retrieval Series 43

Tetsuya Sakai · Douglas W. Oard · Noriko Kando *Editors*

Evaluating Information Retrieval and Access Tasks

NTCIR's Legacy of Research Impact

OPEN ACCESS

Springer

The Information Retrieval Series

Volume 43

Information Retrieval (IR) deals with access to and search in mostly unstructured information, in text, audio, and/or video, either from one large file or spread over separate and diverse sources, in static storage devices as well as on streaming data. It is part of both computer and information science, and uses techniques from e.g. mathematics, statistics, machine learning, database management, or computational linguistics. Information Retrieval is often at the core of networked applications, web-based data management, or large-scale data analysis.

The Information Retrieval Series presents monographs, edited collections, and advanced text books on topics of interest for researchers in academia and industry alike. Its focus is on the timely publication of state-of-the-art results at the forefront of research and on theoretical foundations necessary to develop a deeper understanding of methods and approaches.

More information about this series at http://www.springer.com/series/6128

Tetsuya Sakai · Douglas W. Oard ·
Noriko Kando
Editors

Evaluating Information Retrieval and Access Tasks

NTCIR's Legacy of Research Impact

Springer

Editors
Tetsuya Sakai
Department of Computer
Science and Engineering
Waseda University
Tokyo, Japan

Douglas W. Oard
College of Information Studies
University of Maryland
College Park, MD, USA

Noriko Kando
Information-Society Research Division
National Institute of Informatics
Tokyo, Japan

This open access book is funded by the National Institute of Informatics, Japan.

ISSN 1871-7500 ISSN 2730-6836 (electronic)
The Information Retrieval Series
ISBN 978-981-15-5553-4 ISBN 978-981-15-5554-1 (eBook)
https://doi.org/10.1007/978-981-15-5554-1

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Foreword

It has been a privilege of my career to be involved in many information retrieval evaluation campaigns. For a range of reasons, my involvement with NTCIR has been the longest and the most enjoyable.

Principal among the reasons is Noriko Kando. Evaluation campaigns are sustained by the unyielding enthusiasm of a core organizer. Kando-san has devotedly innovated this campaign from its very start. Thanks to her, what you will consistently see in the chapters of this book is a sequence of tasks or tracks that were ahead of their time. NTCIR was the first to explore patent search, first to incorporate life logging data, and first to examine retrieval of mathematical formulas.

NTCIR has innovated in the methodologies used to measure a shared task: the visualization and summarization tasks, for example, require a quite different approach to evaluation than you would see in many other campaigns. Thanks to Sakai-san's diligent creation, many of the chapters will describe assessment with novel measures. NTCIR was the first to use graded evaluation.

Diversity of excellent evaluation research is what I know I will see when I attend NTCIR at the NII building in Chiyoda-ku. Such a range of innovations can only come from a team of outstanding collaborators: you can see from the diversity of chapter authors just how many have contributed their ideas and hard work to NTCIR.

Dedication to quality is another reason for my regular visits to Tokyo. Such is the commitment of Kando-san and her team that on the evening marking the end of each NTCIR conference (a time when normal organizers just want to sleep) the team meet up in the NII Tower to discuss what worked well, what didn't, and how to improve. Oard-san's thoughtful advice is often to be found there. At that meeting less than a few hours after NTCIR has completed, the next campaign is being planned.

The work described in this book charts the progression of the academic field of information retrieval research from a rather limited library focused research topic to a rich multi-faceted study of information access of all forms of content. It has been my honor to be a part of this campaign and I look forward to what rich new topics it will tackle in the future, at its sesquiennial pace.

Melbourne, Australia Mark Sanderson

Preface

The NTCIR-1 Conference took place in 1999. Back then, NTCIR stood for *NACSIS Test Collection for Information Retrieval systems*. Ever since, NTCIR has grown in size, broadened its scope, and evolved; now we know it as *NII Testbeds and Community for Information access Research*. We editors of this book would like to thank everyone who has been involved in NTCIR in the past two decades or so, and in particular the following people, for making this book happen.

- The chapter authors: Akiko Aizawa, Rami Albatal, Kuang-Hua Chen, Duc-Tien Dang-Nguyen, Zhicheng Dou, Atsushi Fujii, Takahiro Fukushima, Isao Goto, Cathal Gurrin, Graham Healy, Tsutomu Hirao, Frank Hopfgartner, Makoto Iwayama, Hideo Joho, Noriko Kando, Makoto P. Kato, Tsuneaki Kato, Kazuaki Kishida, Michael Kohlhase, Yiqun Liu, Cheng Luo, Teruko Mitamura, Hidetsugu Nanba, Eric Nyberg, Douglas W. Oard, Manabu Okumura, Tetsuya Sakai, Mark Sanderson, Yohei Seki, Ruihua Song, Masaharu Yoshioka, Min Zhang, and Liting Zhou;
- The chapter reviewers: Martin Braschler, Wolfgang Hurst, Nattiya Kanhabua, Stefano Mizzaro, Tatsunori Mori, Ian Soboroff, Damiano Spina, Takehiro Yamamoto, and Richard Zanibbi;
- Those who offered constructive comments on the early drafts of the chapters that were publicly available online;
- Past and present NTCIR general chairs, PC chairs, EVIA chairs, organizing committee members, and staff;
- Past and present NTCIR task organizers and participants, and last but not least;
- Springer's Mio Sugino for her support and perseverance.

This is the first book on NTCIR. A copy of it will be given to all NTCIR-15 participants in December 2020. It has been a long journey, but the journey continues. Stay safe and healthy.

Tokyo, Japan — Tetsuya Sakai
College Park, MD, USA — Douglas W. Oard
Tokyo, Japan — Noriko Kando
April 2020

Contents

1 **Graded Relevance** 1
Tetsuya Sakai

2 **Experiments on Cross-Language Information Retrieval Using Comparable Corpora of Chinese, Japanese, and Korean Languages** 21
Kazuaki Kishida and Kuang-hua Chen

3 **Text Summarization Challenge: An Evaluation Program for Text Summarization** 39
Hidetsugu Nanba, Tsutomu Hirao, Takahiro Fukushima, and Manabu Okumura

4 **Challenges in Patent Information Retrieval** 49
Makoto Iwayama, Atsushi Fujii, and Hidetsugu Nanba

5 **Multi-modal Summarization** 71
Tsuneaki Kato

6 **Opinion Analysis Corpora Across Languages** 83
Yohei Seki

7 **Patent Translation** 97
Isao Goto

8 **Component-Based Evaluation for Question Answering** 109
Teruko Mitamura and Eric Nyberg

9 **Temporal Information Access** 127
Masaharu Yoshioka and Hideo Joho

10 **SogouQ: The First Large-Scale Test Collection with Click Streams Used in a Shared-Task Evaluation** 143
Ruihua Song, Min Zhang, Cheng Luo, Tetsuya Sakai, Yiqun Liu, and Zhicheng Dou

11 Evaluation of Information Access with Smartphones 151
Makoto P. Kato

12 Mathematical Information Retrieval 169
Akiko Aizawa and Michael Kohlhase

13 Experiments in Lifelog Organisation and Retrieval at NTCIR 187
Cathal Gurrin, Hideo Joho, Frank Hopfgartner, Liting Zhou, Rami Albatal, Graham Healy, and Duc-Tien Dang Nguyen

14 The Future of Information Retrieval Evaluation 205
Douglas W. Oard

Index 217

Acronyms

ACLIA	Advanced Cross-Lingual Information Retrieval and Question Answering
AI	Artificial Intelligence
AKG	Actionable Knowledge Graph
AP	Average Precision
BLIR	Bilingual Information Retrieval
CENTRE	CLEF NTCIR TREC Reproducibility
CLEF	Conference and Labs of the Evaluation Forum
CLIR	Cross-Language Information Retrieval
CLPR	Cross-Lingual Patent Retrieval
CoNLL	Conference on Computational Natural Language Learning
CQA	Community Question Answering
DARPA	Defense Advanced Research Projects Agency
DCG	Discounted Cumulative Gain
DUC	Document Understanding Conference
ELRA	European Language Resources Association
ERR	Expected Reciprocal Rank
FIRE	Forum for Information Retrieval Evaluation
GALE	Global Autonomous Language Exploitation
GeoTime	Geographic and Temporal Information Retrieval
GIR	Geographic Information Retrieval
IA	Information Access
INEX	Initiative for Evaluation of XML Retrieval
IPC	International Patent Classification
IR	Information Retrieval
IR4QA	Information Retrieval for Question Answering
IREX	Information Retrieval and Extraction Exercise
IRF	Information Retrieval Facility
iUnit	Information Unit
IWSLT	International Workshop on Spoken Language Translation

JIPA	Japan Intellectual Property Association
JPO	Japan Patent Office
LDC	Linguistic Data Consortium
LDC-IL	Linguistic Data Consortium for Indian Languages
LM	Language Modeling
MAP	Mean Average Precision
MediaEval	Benchmarking Initiative for Multimedia Evaluation
MIR	Mathematical Information Retrieval
MIREX	Music Information Retrieval Evaluation eXchange
MLIR	Multilingual Information Retrieval
MT	Machine Translation
MUC	Message Understanding Conferences
MuST	Multi-modal Summarization for Trend information
NACSIS	National Center for Science Information Systems
nCG	Normalized Cumulative Gain
NCP	Normalized Cumulative Precision
NCU	Normalized Cumulative Utility
nDCG	Normalized Discounted Cumulative Gain
NER	Named Entity Recognition
NII	National Institute of Informatics
NIST	National Institute of Standards and Technology
NMT	Neural Machine Translation
NTCIR	NACSIS/NII Test Collection for Information Retrieval systems/NII Testbeds and Community for Information access Research
OHSUMED	Oregon Health Sciences University's MEDLINE Data Collection
OOV	Out-Of-Vocabulary
PEE	Patent Examination Evaluation
POS	Part-Of-Speech
PRF	Pseudo Relevance Feedback
PRP	Probability Ranking Principle
QAC	Question Answering Challenge
QE	Query Expansion
RBMT	Rule-Based Machine Translation
RR	Reciprocal Rank
SemEval	Evaluation Exercises for the Semantic Evaluation of Text
SensEval	Evaluation Exercises for Word Sense Disambiguation
SERP	Search Engine Result Page
SMT	Statistical Machine Translation
STC	Short Text Conversation
TDT	Topic Detection and Tracking
TIDES	Translingual Information Detection, Extraction, and Summarization
TREC	Text REtrieval Conference
TRECVID	TREC Video Retrieval Evaluation
USPTO	United States Patent and Trademark Office
WAT	Workshop on Asian Translation

WIPO	World Intellectual Property Organization
WMT	Workshop on Statistical Machine Translation
WRR	Weighted Reciprocal Rank
WSD	Word Sense Disambiguation
WWW	We Want Web/World Wide Web

Chapter 1
Graded Relevance

Tetsuya Sakai

Abstract NTCIR was the first large-scale IR evaluation conference series to construct test collections with graded relevance assessments: the NTCIR-1 test collections from 1998 already featured *relevant* and *partially relevant* documents. In this chapter, I provide a survey on the use of graded relevance assessments and of graded relevance measures in the past NTCIR tasks which primarily tackled ranked retrieval. My survey shows that the majority of the past tasks fully utilised graded relevance by means of graded evaluation measures, but not all of them; interestingly, even a few relatively recent tasks chose to adhere to binary relevance measures. I conclude the chapter by a summary of my survey in table form.

1.1 Introduction

The evolution of NTCIR is quite different from that of TREC when it comes to how *relevance assessments* have been collected and utilised. In 1992, TREC started off with a high-recall task (i.e., the *adhoc* track), with *binary* relevance assessments (Harman 2005). Moreover, early TREC tracks heavily relied on evaluation measures based on binary relevance such as *11-point Average Precision*, *R-precision*, and (noninterpolated) *Average Precision*. It was in the TREC 2000 (a.k.a. TREC-9) Main Web task that 3-point *graded* relevance assessments were introduced, based on feedback from web search engine companies at that time Hawking and Craswell (2005, p. 204). Accordingly, this task also Järvelin and Kekäläinen (2000) adopted Discounted Cumulative Gain (DCG), to utilise the graded relevance assessments.

NTCIR has collected graded relevance assessments from the very beginning: the NTCIR-1 test collections from 1998 already featured *relevant* and *partially relevant* documents (Kando et al. 1999). Thus, while NTCIR borrowed many ideas from TREC when it was launched in the late 1990s, its policy regarding relevance assessments seems to have followed the paths of *Cranfield II* (which had 5-point relevance

T. Sakai (✉)
Waseda University, Shinjuku-ku Okubo 3-4-1 63-05-04, Tokyo 169-8555, Japan
e-mail: tetsuyasakai@acm.org

T. Sakai et al. (eds.), *Evaluating Information Retrieval and Access Tasks*,
The Information Retrieval Series 43,
https://doi.org/10.1007/978-981-15-5554-1_1

levels) Cleverdon et al. (1966, p. 21), Oregon Health Sciences University's MEDLINE Data Collection (OHSUMED) (which had 3-point relevance levels) (Hersh et al. 1994), as well as the first Japanese IR test collections *BMIR-J1* and *BMIR-J2* (which also had 3-point relevance levels) (Sakai et al. 1999).

Interestingly, with perhaps a notable exception of the aforementioned TREC 2000 Main Web Task, it is true for both TREC and NTCIR that the introduction of graded relevance assessments did not necessarily mean immediate adoption of *evaluation measures* that can utilise graded relevance. For example, while the TREC 2003–2005 robust tracks constructed adhoc IR test collections with 3-point graded relevance assessments, they adhered to binary relevance measures such as Average Precision (AP). Similarly, as I shall discuss in this chapter,[1] while almost all of the past IR tasks of NTCIR had graded relevance assessments, not all of them fully utilised them by means of graded relevance measures. This is the case despite the fact that a graded relevance measure called the *normalised sliding ratio* (NSR)[2] was proposed in 1968 (Pollock 1968), and was discussed in an 1997 book by Korfhage along with another graded relevance measure (Korfhage 1997, p.209).

1.2 Graded Relevance Assessments, Binary Relevance Measures

This section provides an overview of NTCIR ranked retrieval tasks that did not use graded relevance evaluation measures even though they had graded relevance assessments.

1.2.1 Early IR and CLIR Tasks (NTCIR-1 Through -5)

The Japanese IR and (Japanese-English) crosslingual tasks of NTCIR-1 (Kando et al. 1999) constructed test collections with 3-point relevance levels, but used binary relevance measures such as AP and *R-precision* by either treating the relevant and partially relevant documents as "relevant" or treating only the relevant documents as "relevant." However, it should be stressed at this point that using binary relevance measures with different relevance thresholds cannot serve as substitutes for a graded relevance measure that enables optimisation towards an ideal ranked list (i.e., a list of documents sorted in decreasing order of relevance levels). If partially relevant

[1] A 31-page, March 2019 version of this chapter is available on arxiv.org Sakai (2019). The arxiv version contains the definitions of the main graded relevance measures used at NTCIR, as well as details on how graded relevance levels were constructed from individual assessors' judgements for some of the tasks.

[2] NSR is actually what is now known as *normalised (nondiscounted) cumulative gain* (nCG): See Sakai (2019).

documents are ignored, a Search Engine Result Page (SERP) whose top l documents are all partially relevant and one whose top l documents are all nonrelevant can never be distinguished from each other; if relevant documents and partially relevant documents are all treated as relevant, a SERP whose top l documents are all relevant and one whose top l documents are all partially relevant can never be distinguished from each other.

The Japanese and English (monolingual and crosslingual) IR tasks of NTCIR-2 (Kando et al. 2001) constructed test collections with 4-point relevance levels. However, the organisers used binary relevance measures such as AP and R-precision with two different relevance thresholds. As for the Chinese monolingual and Chinese-English IR tasks of NTCIR-2 (Chen and Chen 2001), three judges independently judged each pooled document using 4-point relevance levels, and then a score was assigned to each relevance level. Finally, the scores were averaged across the three assessors. The organisers then applied two different thresholds to map the scores to binary *rigid relevance* and *relaxed relevance* data. For evaluating the runs, rigid and relaxed versions of recall-precision curves (RP curves) were used.

The NTCIR-3 CLIR (Cross-Language IR) task (Chen et al. 2002) was similar to the previous IR tasks: 4-point relevance levels were used, and two relevance thresholds were used. Finally, rigid and relaxed versions of AP were computed for each run. The NTCIR-4 and NTCIR-5 CLIR tasks (Kishida et al. 2004, 2005) adhered to the above practice.

All of the above tasks used the `trec_eval` program from TREC to compute binary relevance measures such as AP.

1.2.2 Patent (NTCIR-3 Through-6)

The NTCIR-3 Patent Retrieval task (Iwayama et al. 2003) was a news-to-patent *technical survey* search task, with 4-point relevance levels. RP curves were drawn based on *strict relevance* and *relaxed relevance*.

The main task of the NTCIR-4 Patent Retrieval task (Fujii et al. 2004) was a patent-to-patent *invalidity search* task. There were two types of relevant documents: A (a patent that can invalidate a given claim on its own) and B (a patent that can invalidate a given claim *only when used with one or more other patents*). For example, patents B_1 and B_2 may each be nonrelevant (as they cannot invalidate a claim individually), but if they are both retrieved, the pair should serve as one relevant document. At the evaluation step, rigid and relaxed APs were computed. Note that the above-relaxed evaluation has a limitation: recall the aforementioned example with B_1 and B_2, and consider a SERP that managed to return only one of them (say B_1). Relaxed evaluation rewards the system for returning B_1, even though this document alone does *not* invalidate the claim.

The Document Retrieval subtask of the NTCIR-5 Patent Retrieval task (Fujii et al. 2005) was similar to its predecessor, but the relevant documents were determined purely based on whether and how they were actually used by a patent examiner to

reject a patent application; no manual relevance assessments were conducted for this subtask. The graded relevance levels were defined as follows: A (a citation that was actually used on its own to reject a given patent application) and B (a citation that was actually used along with another one to reject a given patent application). As for the evaluation measure for Document Ranking, the organisers adhered to rigid and relaxed APs. In addition, the task organisers introduced a Passage Retrieval subtask by leveraging passage-level binary relevance assessments collected as in the NTCIR-4 Patent task: given a patent, systems were required to rank the passages from that same patent. As both single passages and groups of passages can potentially be relevant to the source patent (i.e., the passage(s) can serve as evidence to determine that the entire patent is relevant to a given claim), this poses a problem similar to the one discussed above with patents B_1 and B_2: for example, if two passages p_1, p_2 are relevant as a group but not individually, and if p_1 is ranked at i and p_2 is ranked at $i'(> i)$, how should the SERP of passage be evaluated? To address this, the task organisers introduced a binary relevance measure called the *Combinational Relevance Score* (CRS), which assumes that the user who scans the SERP must reach as far as i' to view both p_1 and p_2.[3]

The Japanese Document Retrieval subtask of the NTCIR-6 Patent Retrieval task (Fujii et al. 2007) had two different sets of graded relevance assessments; the first set ("Def0" with A and B documents) was defined in the same way as in NTCIR-5; the second set ("Def1") was automatically derived from Def0 based on the International Patent Classification (IPC) codes as follows: H (the set of IPC subclasses for this cited patent has no overlap with that of the input patent), A (the set of IPC subclasses for this cited patent has some overlap with that of the input patent), and B (the set of IPC subclasses for this cited patent is identical to that of the input patent. As for the English Document Retrieval subtask, the relevance levels were also automatically determined based on IPC codes, but only two types of relevant documents (A and B) were identified, as each USPTO patent is given only one IPC code. In both subtasks, AP was computed by considering different combinations of the above relevance levels.

1.2.3 SpokenDoc/SpokenQuery& Doc (NTCIR-9 Through -12)

The Spoken Document Retrieval (SDR) subtask of the NTCIR-9 SpokenDoc task (Akiba et al. 2011) had two "subsubtasks": *Lecture Retrieval* and *Passage Retrieval*, where a passage is any sequence of consecutive inter-pausal units. Passage-level relevance assessments were obtained on a 3-point scale, and it appears that the lecture-

[3] In fact, AP, Q or any measure from the NCU family (Sakai and Robertson 2008) can easily be extended to handle *combinational relevance* for Document Retrieval (See the above example with (B_1, B_2)) and for Passage Retrieval (See the above example with (p_1, p_2)): See Sakai (2019).

level (binary) relevance was deduced from them.[4] AP was used for evaluating Lecture Retrieval, whereas variants of AP, called *utterance-based* (M)AP, *pointwise* (M)AP, and *fractional* (M)AP were used for evaluating Passage Retrieval. These are all binary relevance measures. The NTCIR-10 SpokenDoc-2 Spoken Content Retrieval (SCR) subtask (Akiba et al. 2013) was similar to the SDR subtask at NTCIR-9, with Lecture Retrieval and Passage Retrieval subsubtasks. Lecture Retrieval used a revised version of the NTCIR-9 SpokenDoc topic set, and its gold data does not contain graded relevance assessments[5]; binary relevance AP was used for the evaluation. As for Passage Retrieval, a new topic set was devised, again with 3-point relevance levels.[6] The AP variants from the NTCIR-9 SDR task were used for the evaluation again.

The Slide Group Segment (SGS) Retrieval subsubtask of the NTCIR-11 SpokenQuery& Doc SCR subtask involved the ranking of predefined retrieval units (i.e., SGSs), unlike the Passage Retrieval subsubtask which allows any sequence of consecutive inter-pausal units as a retrieval unit. Three-point relevance levels were used to judge the SGSs: R (relevant), P (partially relevant), and I (nonrelevant). However, binary AP was used for the evaluation after collapsing the grades to binary. As for the passage-level relevance assessments, they were derived from the SGSs labelled R or P, and were considered binary; the three AP variants were used for this subsubtask again. Segment Retrieval was continued at the NTCIR-12 SpokenQuery&Doc-2 task, again with the same 3-point relevance levels and AP as the evaluation measure.

1.2.4 Math/MathIR (NTCIR-10 Through -12)

In the Math Retrieval subtask of the NTCIR-10 Math Task, retrieved mathematical formulae were judged on a 3-point scale. Up to two assessors judged each formula, and initially 5-point relevance scores were devised based on the results. For example, for formulae judged by one assessor, they were given 4 points if the judged label was relevant; for those judged by two assessors, they were given 4 points if both of them gave them the relevant label. Finally, the scores were mapped to a 3-point scale: Documents with scores 4 or 3 were treated as relevant; those with 2 or 1 were treated as partially relevant; those with 0 were treated as ronrelevant. However, at the evaluation step, only binary relevance measures such as AP and Precision were computed using `trec_eval`, after collapsing the grades to binary. Similarly, in the Math Retrieval subtask of the NTCIR-11 Math Task (Aizawa et al. 2014), two assessors independently judged each retrieved unit on a 3-point scale, and the

[4]The official test collection data of the NTCIR-9 SDR task (`evalsdr`) contains only passage-level gold data.

[5]This was verified by examining SpokenDoc2-formalrun-SCR-LECTURE-golden-20130129.xml in the SpokenDoc-2 test collection http://research.nii.ac.jp/ntcir/permission/ntcir-10/perm-en-SPOKENDOC.html.

[6]This was verified by examining http://SpokenDoc2-formalrun-SCR-PASSAGE-golden-20130215.xml in the SpokenDoc-2 test collection http://research.nii.ac.jp/ntcir/permission/ntcir-10/perm-en-SPOKENDOC.html.

final relevance levels were also on a 3-point scale. If the two assessor labels were relevant/relevant or relevant/partially relevant, the final grade was relevant; if the two labels were both nonrelevant, the final grade was nonrelevant; the other combinations were considered partially relevant. As for the evaluation measures, *bpref* (Buckley and Voorhees 2004; Sakai 2007; Sakai and Kando 2008) was computed along with AP and Precision using `trec_eval`.

The NTCIR-12 MathIR task was similar to the Math Retrieval subtask of the aforementioned Math tasks. Up to four assessors judged each retrieved unit using a 3-point scale, and the individual labels were consolidated to form the final 3-point scale assessments. As for the evaluation, only Precision was computed at several cutoffs using `trec_eval`.

The NTCIR-11 Math (Aizawa et al. 2014) and NTCIR-12 MathIR (Zanibbi et al. 2016) overview papers suggest that one reason for adhering to binary relevance measures is that `trec_eval` could not handle graded relevance. On the other hand, this may not be the only reason: in the MathIR overview paper, it is reported that the organisers chose Precision because it is "*simple to understand*" (Zanibbi et al. 2016). Thus, some researchers indeed *choose* to focus on evaluation with binary relevance measures, even in the NTCIR community where we have graded relevance data by default and a tool for computing graded relevance measures is known.[7]

1.3 Graded Relevance Assessments, Graded Relevance Measures

This section provides an overview of NTCIR ranked retrieval tasks that employed graded relevance evaluation measures to fully enjoy the benefit of having graded relevance assessments.

1.3.1 Web (NTCIR-3 Through-5)

The NTCIR-3 Web Retrieval task (Eguchi et al. 2003) was the first NTCIR task to use a graded relevance evaluation measure, namely, DCG.[8] Four-point relevance levels were used. In addition, assessors chose a very small number of "best" documents from the pools. To compute DCG, two different gain value settings were used: Rigid (3 for highly relevant, 2 for fairly relevant, 0 otherwise) and Relaxed (3 for highly

[7] `NTCIREVAL` has been available on the NTCIR website since 2010; its predecessor `ir4qa_eval` was released in 2008 (Sakai et al. 2008). Note also that TREC 2010 released https://trec.nist.gov/data/web/10/gdeval.pl for computing Normalised Discounted Cumulative Gain (nDCG) and Expected Reciprocal Rank (ERR).

[8] This was the DCG as originally defined by Järvelin and Kekäläinen (2000) with the logarithm base $b = 2$, which means that gain discounting is not applied to documents at ranks 1 and 2. See also Sect. 1.3.3.

relevant, 2 for fairly relevant, 1 for partially relevant, 0 otherwise). The organisers of the Web Retrieval task also defined a graded relevance evaluation measure called Weighted Reciprocal Rank (WRR), designed for navigational searches. However, what was actually used in the task was the binary relevance Reciprocal Rank (RR), with two different relevance thresholds. Therefore, this measure will be denoted "(W)RR" hereafter whenever graded relevance is not utilised. Other binary relevance measures including AP and R-precision were also used in this task. For a comparison of evaluation measures designed for navigational intents including RR, WRR, and P+, see Sakai (2007).

The NTCIR-4 WEB Informational Retrieval Task (Eguchi et al. 2004) was similar to its predecessor, with 4-point relevance levels; the evaluation measures were DCG, (W)RR, Precision, etc. On the other hand, the NTCIR-4 WEB Navigational Retrieval Task (Oyama et al. 2004), used 3-point relevance levels: A (relevant), B (partially relevant), and D (nonrelevant); the evaluation measures were DCG and (W)RR, and two gain values settings for DCG were used: $(A, B, D) = (3, 0, 0)$ and $(A, B, D) = (3, 2, 0)$.

The NTCIR-5 WEB task ran the Navigational Retrieval subtask, which is basically the same as its predecessor, with 3-point relevance levels and DCG and (W)RR as the evaluation measures. For computing DCG, three gain value settings were used: $(A, B, D) = (3, 0, 0)$, $(A, B, D) = (3, 2, 0)$, and $(A, B, D) = (3, 3, 0)$. Note that the first and the third settings reduce DCG to binary relevance measures.

1.3.2 CLIR (NTCIR-6)

At the NTCIR-6 CLIR task, 4-point relevance levels (S,A,B,C) were used and rigid and relaxed AP scores were computed using `trec_eval` as before. In addition, the organisers computed "*as a trial*" (Kishida et al. 2007) the following graded relevance measures using their own script: nDCG (as defined originally by Järvelin and Kekäläinen 2002), Q-measure (Sakai 2014; Sakai and Zeng 2019) (or "Q"), and Kishida's *generalised AP* (Kishida 2005). See Sakai (2007) for a comparison of these three graded relevance measures. The CLIR organisers developed a program to compute these graded relevance measures, with the gain value setting: $(S, A, B, C) = (3, 2, 1, 0)$.

1.3.3 ACLIA IR4QA (NTCIR-7 and -8)

At the NTCIR-7 Information Retrieval for Question Answering (IR4QA) task (Sakai et al. 2008), a predecessor of `NTCIREVAL` called `ir4qa_eval` was released (See Sect. 1.2.4). This tool was used to compute the Q-measure, the "Microsoft version" of nDCG (Sakai 2014), as well as the binary relevance AP. Microsoft nDCG (called `MSnDCG` in `NTCIREVAL`) fixes a problem with the original nDCG (See also

Sect. 1.3.1): in the original nDCG, if the logarithm base is set to (say) $b = 10$, then discounting is not applied from ranks 1 to 10. Hence, the ranks of the relevant documents within top 10 do not matter. Microsoft nDCG avoids this problem by using $1/\log(1 + r)$ as the discount factor for *every* rank r, but thereby loses the patience parameter b (Sakai 2014).[9] The relevance levels used were L2, L1, and L0. A *linear* gain value setting was used: $(L2, L1, L0) = (2, 1, 0)$. The NTCIR-8 IR4QA task (Sakai et al. 2010) used the same evaluation methodology as above.

1.3.4 GeoTime (NTCIR-8 and -9)

The NTCIR-8 GeoTime task (Gey et al. 2010), which dealt with adhoc IR given "when and where"-type topics, constructed test collections with the following graded relevance levels: Fully relevant (the document answers both the "when" and "where" aspects of the topic), Partially relevant—where (the document only answers the "where" aspect of the topic), and Partially relevant—when (the document only answers the "when" aspect of the topic). The evaluation tools from the IR4QA task were used to compute (Microsoft) nDCG, Q, and AP, with a gain value of 2 for each fully relevant document and a gain value of 1 for each partially relevant one (regardless of "when" or "where") for the two graded relevance measures.[10] The NTCIR-9 GeoTime task (Gey et al. 2011) used the same evaluation methodology as above.

1.3.5 CQA (NTCIR-8)

The NTCIR-8 Community Question Answering (CQA) task (Sakai et al. 2010) was an answer ranking task: given a question from Yahoo! Chiebukuro (Japanese Yahoo! Answers) and the answers posted in response to that question, rank the answers by answer quality. While the Best Answers (BAs) selected by the actual questioners were already available in the Chiebukuro data, additional *graded* relevance assessments were obtained offline to find *Good Answers* (GAs), by letting four assessors independently judge each posted answer. Each assessor labelled an answer as either A (high-quality), B (medium-quality), or C (low-quality), and hence 15 different label *patterns* were obtained: $AAAA, AAAB, \ldots, BCCC, CCCC$. In the official evaluation at NTCIR-8, these patterns were mapped to 4-point relevance levels: for example, $AAAA$ and $AAAB$ were mapped to $L3$-relevant, and $ACCC$, $BCCC$ and $CCCC$ were mapped to $L0$. In a separate study, the same data were mapped to 9-point relevance levels, by giving 2 points to an A and 1 point to a B and summing

[9] D♯-nDCG implemented in `NTCIREVAL` also builds on the Microsoft version of nDCG, not the original nDCG.

[10] While the GeoTime overview paper suggests that the above relevance levels were mapped to binary relevance, this was in fact not the case: see Sakai (2019).

up the scores for each pattern. Using the graded Good Answers data, three graded relevance measures were computed: normalised gain at $l = 1$ (nG@1),[11] nDCG, and Q. In addition, Hit at $l = 1$ was computed for both Best Answers and Good Answers data: this is a binary relevance measure which only cares whether the top-ranked item is relevant or not.

1.3.6 INTENT/IMine (NTCIR-9 Through 12)

The NTCIR-9 INTENT task overview paper (Song et al. 2011) was the first NTCIR overview to mention the use of the `NTCIREVAL` tool, which can compute various graded relevance measures for adhoc and diversified IR including Q, nDCG, and D♯-measures (Sakai and Zeng 2019). D♯-nDCG and its components I-rec and D-nDCG were used as the official evaluation measures. The Document Retrieval (DR) subtask of the INTENT task had intentwise graded relevance assessments on a 5-point scale. While the Subtopic Mining (SM) subtask of the INTENT task also used D♯-nDCG to evaluate ranked lists of subtopic strings, no graded relevance assessments were involved in the SM subtask since each subtopic string either belongs to an intent (i.e., a cluster of subtopic strings) or not. Hence, the SM subtask may be considered to be outside the scope of the present survey; but see Sakai (2019) for a discussion.

The NTCIR-10 INTENT task was basically the same as its predecessor, with 5-point intentwise relevance levels for the DR subtask and D♯-nDCG as the primary evaluation measure. However, as the intents came with informational/navigational tags, new measures called DIN-nDCG and P+Q (Sakai 2014) were used in addition to leverage this information.

The NTCIR-11 IMine task (Liu et al. 2014) was similar to the INTENT tasks, except that its SM subtask required participating systems to return a two-level hierarchy of subtopic strings. The SM subtask was evaluated using the H-measure, which combines (a) the accuracy of the hierarchy, (b) the D♯-nDCG score based on the ranking of the first-level subtopics, and (c) the D♯-nDCG score based on the ranking of the second-level subtopics. However, recall the above remark on the INTENT SM subtask: intentwise graded relevance does not come into play in this subtask. On the other hand, the IMine DR subtask was evaluated in a way similar to the INTENT DR tasks, with D♯-nDCG computed based on 4-point relevance levels: highly relevant, relevant, nonrelevant, and spam. The gain value setting used was: (2, 1, 0, 0).[12] The IMine task also introduced the TaskMine subtask, which requires systems to rank strings that represent subtasks of a given task (e.g., "*take diet pills*" in response to "*lose weight*."). This subtask involved graded relevance assessments. Each subtask string was judged independently by two assessors from the viewpoint of whether

[11] nG@1 is often referred to as nDCG@1; however, note that neither discounting nor cumulation is applied at rank 1.

[12] Kindly confirmed by task organisers Yiqun Liu and Cheng Luo in a private email communication (March 2019).

the subtask is effective for achieving the input task. A 4-point per-assessor relevance scale was used,[13] with weights (3, 2, 1, 0), and final relevance levels were given as the average of the two scores, which means that a 6-point relevance scheme was adopted. The averages were used verbatim as gain values: (3.0, 2.5, 2.0, 1.5, 1.0, 0). The evaluation measure used was nDCG, but duplicates (i.e., multiple strings representing the same subtask) were not rewarded.

The Query Understanding (QU) subtask of the NTCIR-12 IMine-2 Task (Yamamoto et al. 2016), a successor of the previous SM subtasks of INTENT/IMine, required systems to return a ranked list of (subtopic, vertical) pairs (e.g., ("*iPhone 6 photo*", Image), ("*iPhone 6 review*", Web)) for a given query. The official evaluation measure, called the QU-score, is a linear combination of D♯-nDCG (computed as in the INTENT SM subtasks) and the V-score which measures the appropriateness of the named vertical for each subtopic string. Despite the binary relevance nature of the subtopic mining aspect of the QU subtask, it deserves to be discussed in the present survey because the V-score part relies on graded relevance assessments. To be more specific, the V-score relies on the probabilities $\{Pr(v|i)\}$, for intents $\{i\}$ and verticals $\{v\}$, which are derived from 3-point scale relevance assessments: 2 (highly relevant), 1 (relevant), and 0 (nonrelevant). Hence the QU-score may be regarded as a graded relevance measure. The Vertical Incorporating (VI) subtask of the NTCIR-12 IMine-2 Task (Yamamoto et al. 2016) also used a version of D♯-nDCG to allow systems to embed verticals (e.g., Vertical-News, Vertical-Image) within a ranked list of document IDs for diversified search. More specifically, the organisers replaced the intentwise gain value $g_i(r)$ at rank r in the global gain formula (Sakai 2014) with $Pr(v(r)|i)g_i(r)$, where $v(r)$ is the *vertical type* ("Web," Vertical-News, Vertical-Image, etc.) of the document at rank r, and the vertical probability given an intent is obtained from 3-point scale relevance assessments as described above. As for the intentwise gain value $g_i(r)$, it was also on a 3-point scale for the Web documents: 2 for highly relevant, 1 for relevant, and 0 for nonrelevant documents. Moreover, if the document at r was a vertical, the gain value was set to 2. In addition, the VI subtask collected *topicwise* relevance assessments on a 4-point scale: highly relevant, relevant, nonrelevant, and spam. The gain values used were: (2, 1, 0, 0).[14] As the subtask had a set of very clear, single-intent topics among their full topic set, Microsoft nDCG (rather than D♯-nDCG) was used for these particular topics.

1.3.7 RecipeSearch (NTCIR-11)

While the official evaluation results of Adhoc Recipe Search subtask of the NTCIR-11 RecipeSearch Task (Yasukawa et al. 2014) were based on binary relevance, the

[13] While the overview (Sect. 4.3) says that a 3-point scale was used, this was in fact not the case: kindly confirmed by task organiser Takehiro Yamamoto in a private email communication (March 2019).

[14] Kindly confirmed by task organisers Yiqun Liu and Cheng Luo in a private email communication (March 2019).

organisers also explored evaluation based on graded relevance: they obtained graded relevance assessments on a 3-point scale for a subset (111 topics) of the full test topic set (500 topics).[15] Microsoft nDCG was used to leverage the above data with a linear gain value setting, along with the binary AP and RR.

1.3.8 Temporalia (NTCIR-11 and -12)

The Temporal Information Retrieval (TIR) subtask of the NTCIR-11 Temporalia Task collected relevance assessments on a 3-point scale. Each TIR topic contained a *past question*, *recency question*, *future question*, and an *atemporal question*; participating systems were required to produce a Search Engine Result Page (SERP) for each of the above four questions. This adhoc IR task used Precision and Microsoft nDCG as the official measures, and Q for reference.

While the Temporally Diversified Retrieval (TDR) subtask of the NTCIR-12 Temporalia-2 Task was similar to the above TIR subtask, it required systems to return a fifth SERP, which covers all of the above four temporal classes. That is, this fifth SERP is a diversified SERP, where the temporal classes can be regarded as different search intents for the same topic. The relevance assessment process followed the practice of the NTCIR-11 TIR task, and the SERPs for the four questions were evaluated using nDCG. As for the diversified SERPs, they were evaluated using α-nDCG (Clarke et al. 2008) and D♯-nDCG.

A linear gain value setting was used in both of the above subtasks.[16]

1.3.9 STC (NTCIR-12 Through -14)

The NTCIR-12 Short Text Conversation (STC) task (Shang et al. 2016) was a response retrieval task given a tweet (or a Chinese Weibo post). For both Chinese and Japanese subtasks, the response tweets were first labelled on a binary scale, for each of the following criteria: Coherence, Topical Relevance, Context Independence, and Non-repetitiveness. The final graded relevance levels were determined using the following mapping scheme:

if Coherent AND Topically Relevant
  if Context-independent AND Non-repetitive
    RelevanceLevel = $L2$
  else
    RelevanceLevel = $L1$
else

[15] While the overview paper says that a 4-point scale was used, this was in fact not the case: kindly confirmed by task organiser Michiko Yasukawa (March 2019) in a private email communication.

[16] Kindly confirmed by task organiser Hideo Joho in a private email communication (March 2019).

RelevanceLevel = $L0$.

Following the *quadratic* gain value setting often used for web search evaluation (Burges et al. 2005) and for computing ERR (Chapelle et al. 2009), the Chinese subtask organisers mapped the L2, L1, and L0 relevance levels to the following gain values: $2^2 - 1 = 3, 2^1 - 1 = 1, 2^0 - 1 = 0$; according to the present survey of NTCIR retrieval tasks, this is the only case where a quadratic gain value setting was used instead of the linear one. The evaluation measures used for this subtask were nG@1, P+, and normalised ERR (nERR). As for the Japanese subtask which used Japanese Twitter data, the same mapping scheme was applied, but the scores $((L2, L1, L0) = (2, 1, 0))$ from 10 assessors were averaged to determine the final gain values; a binary relevance, set-retrieval accuracy measure was used instead of P+, along with nG@1 and nERR.

The NTCIR-13 STC task (Shang et al. 2017) was similar to its predecessor, although systems were allowed to *generate* responses instead of retrieving existing tweets. In the Chinese subtask, 7-point relevance levels were obtained by summing up the assessor scores, and a linear gain value setting was used to compute nG@1, P+, and nERR. In addition, an alternative approach to consolidating the assessor scores was explored, by considering the fact that some tweets receive unanimous ratings while others do not even if they are the same in terms of the sum of assessor scores (Sakai 2017). The NTCIR-13 STC Japanese subtask used Yahoo! News Comments data instead of Japanese Twitter data. The evaluation method was similar to what was used in the previous Japanese subtask; see Sakai (2019) for more details.

Although the Chinese Emotional Conversation Generation (CECG) subtask of the NTCIR-14 STC subtask (Zhang and Huang 2019) is not exactly a ranked retrieval task, we discuss it here as it is a successor of the previous Chinese STC subtasks that utilises graded relevance measures. Given an input tweet *and an emotional category* such as Happiness and Sadness, participating systems for this subtask were required to return *one* generated response. A mapping scheme similar to the previous Chinese subtasks were used to form 3-point relevance levels. As for the evaluation measures, the relevance scores $(L2, L1, L0) = (2, 1, 0)$ of the returned responses were simply summed or averaged across the test topics.

1.3.10 WWW (NTCIR-13 and -14) and CENTRE (NTCIR-14)

The NTCIR-13 We Want Web (WWW) Task (Luo et al. 2017) was an adhoc web search task. For the Chinese subtask, three assessors independently judged each pooled web page on a 4-point scale: (3, 2, 1, 0); the scores were then summed up to form the final 10-point relevance levels. For the English subtask, two assessors independently judged each pooled web page on a different 4-point scale: highly relevant (2 points), relevant (1 point), nonrelevant (0 points), and error (0 points); the scores were then summed up to form the final 5-point relevance levels. In both

subtasks, linear gain value settings were used to compute (Microsoft) nDCG, Q (the cutoff version (Sakai 2014)), and nERR.

The NTCIR-14 WWW Task (Mao et al. 2019) was similar to its predecessor. The Chinese subtask used the following judgment criteria: highly relevant (3 points), relevant (2 points), marginally relevant (1 point), nonrelevant (0 points), garbled (0 points). Although three assessors judged each topic, the final relevance levels were obtained on a majority-vote basis rather than taking the sum; hence 4-point scale relevance levels were used this time. As for the English subtask, 5-point relevance levels were obtained by following the methodology of the NTCIR-13 English subtask. Both subtasks adhered to Microsoft nDCG, (cutoff-based) Q, and nERR with linear gain value settings.

The NTCIR-14 CLEF NTCIR TREC Reproducibility (CENTRE) task (Sakai et al. 2019) encouraged participants to replicate a pair of runs from the NTCIR-13 WWW English subtask and to reproduce a pair of runs from the TREC 2013 Web Track adhoc task (Collins-Thompson et al. 2014). Additional relevance assessments were conducted on top of the official NTCIR-13 WWW English test collection, by following the relevance assessment methodology of the WWW subtask. As for the evaluation of the TREC runs with the TREC 2013 Web Track adhoc test collection, the original 6-point scale relevance levels Navigational, Key, Highly relevant, Relevant, Nonrelevant, Junk were mapped to $L4$, $L3$, $L2$, $L1$, $L0$, $L0$, respectively. All runs involved in the CENTRE task were evaluated using Microsoft nDCG, (cutoff-based) Q, and nERR, with linear gain value settings.

1.3.11 AKG (NTCIR-13)

The NTCIR-13 Actionable Knowledge Graph (AKG) task (Blanco et al. 2017) had two subtasks: *Action Mining* (AM) and *Actionable Knowledge Graph Generation* (AKGG). Both of them involved graded relevance assessments and graded relevance measures. The AM subtask required systems to rank *actions* for a given *entity type* and an *entity instance*: for example, given "*Product*" and "*Final Fantasy VIII*," the ranked actions could contain "*play on Android*," "*buy new weapons*," etc. Two sets of relevance assessments were collected by means of crowd sourcing: the first set judged the verb parts of the actions ("*play*," "*buy*," etc.) whereas the second set judged the entire actions (verb plus modifier as exemplified above). Both sets of judgements were done based on 4-point relevance levels. The AKGG subtask required participants to rank entity *properties*: for example, given a quadruple (Query, Entity, Entity Types, Action) = ("*request funding*," "*funding*," "*thing, action*," "*request funding*"), systems might return "*Agent*," "*ServiceType*," "*Result*," etc. Relevance assessments were conducted by crowd workers on a 5-point scale. Both subtasks used nDCG and nERR for the evaluation; linear gain value settings were used.[17]

[17] Kindly confirmed by task organiser Hideo Joho in a private email communication (March 2019).

1.3.12 OpenLiveQ (NTCIR-13 and -14)

The NTCIR-13 OpenLiveQ task (Kato et al. 2017) required participants to rank Yahoo! Chiebukuro questions for a given query, and the offline evaluation part of this task involved ranked list evaluation with graded relevance. Five crowd workers independently judged a list of questions for query q under the following instructions: "Suppose you input q and received a set of questions as shown below. Please select all the questions that you would want to click." Thus, while the judgement is binary for each assessor, 6-point relevance levels were obtained based on the number of votes. (Microsoft) nDCG, Q, and ERR were computed using a linear gain value setting.

The NTCIR-14 OpenLiveQ-2 task (Kato et al. 2019) is similar to its predecessor, but this time the evaluation involved *unjudged* documents, as the relevance assessments from NTCIR-13 were reused but the target questions to be ranked were not identical to the NTCIR-13 version. The organisers therefore used *condensed-list* (Sakai 2014) versions of Q, (Microsoft) nDCG, and ERR. Also, for OpenLiveQ-2, the organisers switched their primary measure from nDCG to Q, as Q substantially outperformed nDCG (at $l = 5, 10, 20$) in terms of correlation with online (i.e., click-based) evaluation in their experiments (Kato et al. 2018).

1.4 Summary

Table 1.1 summarises Sect. 1.2; Table 1.2 summarises Sect. 1.3. It can be observed that (a) the majority of the past NTCIR ranked retrieval tasks utilised graded relevance measures; and that (b) even a few relatively recent tasks, namely, SpokenQuery& Doc and MathIR from NTCIR-12 held in 2016, refrained from using graded relevance measures. As was discussed in Sect. 1.2.1, researchers should be aware that binary relevance measures with different relevance thresholds (e.g., Relaxed AP and Rigid AP) cannot serve as substitutes for good graded relevance measures. *If* the optimal ranked output for a task is defined as one that sorts all relevant documents in decreasing order of relevance levels, then by definition, graded relevance measures should be used to evaluate and optimise the retrieval systems.

One additional remark regarding Tables 1.1 and 1.2 is that the NTCIR-5 CLIR overview paper (Kishida et al. 2007) was the last to report on RP curves; the RP curves completely disappeared from the NTCIR overviews after that. This may be because (a) interpolated precisions at different recall points (Sakai 2014) do not directly reflect user experience; and (b) graded relevance measures have become more popular than before.

Over the past decade or so, some researchers have pointed out a few disadvantages of using graded relevance, especially in the context of promoting *preference judgements* (e.g., Bashir et al. 2013; Carterette et al. 2008). Carterette et al. (2008) argue that (i) it is difficult to determine relevance grades in advance and to anticipate how the decision will affect evaluation; and (ii) having more grades means more

Table 1.1 NTCIR ranked retrieval tasks with graded relevance assessments and binary relevance measures. Note that the relevance levels for the Patent Retrieval tasks of NTCIR-4 to -6 exclude the "nonrelevant" level: the actual labels are shown here because they are not simply different degrees of relevance (See Sect. 1.2.2)

Task or subtask	NTCIR round (year)	Relevance levels	Main evaluation measures discussed in overview
Japanese and JEIR	1 (1999)	3	AP, R-precision, Precision, RP curves
JEIR	2 (2001)	4	AP, R-precision, Precision, Interpolated Precision, RP curves
Chinese and CEIR	2	4 per assessor	RP curves
CLIR	3–5(2002–2005)	4	AP, RP curves
Patent retrieval	3 (2002)	4	RP curves
Patent retrieval	4 (2004)	A,B	AP, RP curves
Patent retrieval	5 (2005)	A,B	CRS (for passage retrieval), AP
Patent retrieval	6 (2007)	A,B/H,A,B (Japanese) A,B (English)	AP AP
Spoken document/content retrieval	9–11(2011–2014)	3	AP and passage-level variants
SQ-SCR (SGS)	12 (2016)	3	AP
Math retrieval	10 (2013)	5 mapped to 3	AP, Precision
Math retrieval	11 (2014)	3	AP, Precision, Bpref
MathIR	12 (2016)	3	Precision

burden on the users. Regarding (i), while it is important to always check how our use of grades affects the evaluation outcome, in many cases relevance grades can be naturally defined based on individual assessors' labels; I argue that it is useful to preserve the raw judgements in the form of graded relevance rather than to collapse them to binary; see also the discussion below on label *distributions*. Regarding (ii), rich relevance grades can be obtained even if the individual judgements are binary or tertiary, as I have illustrated in this chapter. Moreover, while I agree that simple side-by-side preference judgements are useful (and can even be used for constructing graded relevance data), it should be pointed out that some of the approaches in the preference judgements domain require more complex judgement protocols than this, e.g., *graded* preference judgements (Carterette et al. 2008), and *contextual* preference judgements (Chandar and Carterette 2013; Golbus et al. 2014). Moreover, while I agree that utilising preference judgements is a promising avenue for future research, the *incompleteness* problem of preference judgements needs to be solved.

What lies beyond graded relevance then? Here is my personal view concerning offline evaluation (as opposed to online evaluation using click data etc.). Information

Table 1.2 NTCIR ranked retrieval tasks with graded relevance assessments and graded relevance measures. Binary relevance measures are shown in parentheses

Task or subtask	NTCIR round (year)	Relevance levels	Main evaluation measures discussed in overview
Web retrieval	3 (2003)	4 + best documents	DCG ((W)RR, AP, RP curves)
WEB informational	4 (2004)	4	DCG ((W)RR, Precision, RP curves)
WEB navigational		3	DCG, ((W)RR, UCS)
WEB navigational	5 (2005)	3	DCG, ((W)RR)
CLIR	6 (2007)	4	nDCG, Q, generalised AP (AP)
IR4QA	7–8 (2008–2010)	3	nDCG, Q (AP)
GeoTime	8–9(2010–2011)	3∗	nDCG, Q (AP)
CQA	8 (2010)	4(9) + best answers	GA-{nG@1, nDCG, Q}, (GA-Hit@1, BA-Hit@1) etc.
INTENT DR	9 (2011)	5	D♯-nDCG
INTENT DR	10 (2013)	5	D♯-nDCG, DIN-nDCG, P+Q
IMine DR	11 (2014)	4 incl. Spam	D♯-nDCG
IMine TaskMine	11	6	nDCG
IMine QU	12 (2016)	3 (vertical)	QU-score
IMine VI	12	3 (vertical) 3 (intentwise) 3 + Spam (topicwise)	D♯-nDCG, nDCG
RecipeSearch	11 (2014)	3(2)	nDCG (AP, RR)
Temporalia TIR	11	3	nDCG, Q, (Precision)
Temporalia TDR	12 (2016)	3	nDCG, α-nDCG, D♯-nDCG
STC Chinese	12	3	nG@1, P+, nERR
STC Chinese	13 (2017)	7(10)	nG@1, P+, nERR
STC Japanese	12–13(2016–2017)	3 per assessor	nG@1, nERR (Accuracy)
STC CECG	14 (2019)	3	Sum/average of relevance scores
WWW English	13–14(2017–2019)	5	nDCG, Q, nERR
WWW Chinese	13 (2017)	10	nDCG, Q, nERR
WWW Chinese	14	4	nDCG, Q, nERR
AKG	13 (2017)	4 (AM) / 5 (AKGG)	nDCG, nERR
OpenLiveQ	13–14(2017–2019)	6	nDCG, Q, ERR(with condensed lists at NTCIR-14)
CENTRE	14 (2019)	5	nDCG, Q, nERR

∗two types of partially relevant (*when* and *where*) counted as one level

Retrieval (IR) and Information Access (IA) tasks have diversified, and relevance assessments require more subjective and diverse views than before. We are no longer just talking about whether a scientific article is relevant to the researcher's question (as in Cranfield); we are also talking about whether a response of a chatbot is "relevant" response to the user's utterance, about whether a reply to a post on social media is "relevant," and so on. Graded relevance implies that there should be a *single* label for each item to be retrieved (e.g., "this document is highly relevant"), but these new tasks may require a *distribution* of labels reflecting different users's points of view. Hence, instead of collapsing this distribution to form a single label, methods to preserve the distribution of labels in the test collection may be useful, as was implemented at the Dialogue Breakdown Detection Challenge (Higashinaka et al. 2017). The Dialogue Quality (DQ) and Nugget Detection (ND) subtasks of the NTCIR-14 STC task were the very first of NTCIR efforts in that direction: they compared gold label distributions with systems' estimated distributions (Sakai 2018; Zeng et al. 2019). See also Maddalena et al. (2017) for a related idea.

Acknowledgements Many thanks to the editors (especially Doug Oard who was in charge of this chapter), reviewers (especially Damiano Spina who kindly reviewed this chapter), and all authors of this book, and to the present and past organisers and participants of the NTCIR tasks! In particular, I thank the following past task organisers for clarifications regarding their overview papers: Cheng Luo, Yiqun Liu, and Takehiro Yamamoto (IMine), Makoto P. Kato (OpenLiveQ), Michiko Yasukawa (RecipeSearch), and Hideo Joho (Temporalia and AKG tasks).

References

Aizawa A, Kohlhase M, Ounis I (2014) NTCIR-11 Math-2 task overview. In: Proceedings of NTCIR-11, pp 88–98

Akiba T, Nishizaki H, Aikawa K, Kawahara T, Matsui T (2011) Overview of the IR for spoken documents task in NTCIR-9 workshop. In: Proceedings of NTCIR-9, pp 223–235

Akiba T, Nishizaki H, Aikawa K, Hu X, Ito Y, Kawahara T, Nakagawa S, Nanjo H, Yamashita Y (2013) Overview of the NTCIR-10 SpokenDoc-2 task. In: Proceedings of NTCIR-10, pp 573–587

Bashir M, Anderton J, Wu J, Golbus PB, Pavlu V, Aslam JA (2013) A document rating system for preference judgements. In: Proceedings of ACM SIGIR 2013, pp 909–912

Blanco R, Joho H, Jatowt A, Yu H, Yamamoto S (2017) Overview of NTCIR-13 actionable knowledge graph (AKG) task. In: Proceedings of NTCIR-13, pp 340–345

Buckley C, Voorhees EM (2004) Retrieval evaluation with incomplete information. In: Proceedings of ACM SIGIR 2004, pp 25–32

Burges C, Shaked T, Renshaw E, Lazier A, Deeds M, Hamilton N, Hullender G (2005) Learning to rank using gradient descent. In: Proceedings of ICML 2005, pp 89–96

Carterette B, Bennett PN, Chickering DM, Dumais ST (2008) Here or there: preference judgments for relevance. In: Proceedings of ECIR 2008. LNCS, vol 4956, pp 16–27

Chandar P, Carterette B (2013) Preference based evaluation measures for novelty and diversity. In: Proceedings of ACM SIGIR 2013, pp 413–422

Chapelle O, Metzler D, Zhang Y, Grinspan P (2009) Expected reciprocal rank for graded relevance. In: Proceedings of ACM CIKM 2009, pp 621–630

Chen KH, Chen HH (2001) The Chinese text retrieval tasks of NTCIR workshop 2. In: Proceedings of NTCIR-2

Chen KH, Chen HH, Kando N, Kuriyama K, Lee S, Myaeng SH, Kishida K, Eguchi K, Kim H (2002) Overview of CLIR task at the third NTCIR workshop. In: Proceedings of NTCIR-3

Clarke CL, Kolla M, Cormack GV, Vechtomova O, Ashkan A, Büttcher S, MacKinnon I (2008) Novelty and diversity in information retrieval evaluation. In: Proceedings of ACM SIGIR 2008, pp 659–666

Cleverdon CW, Mills J, Keen EM (1966) Factors determining the performance of indexing systems; volume 1: Design. Technical report, The College of Aeronautics, Cranfield

Collins-Thompson K, Bennett P, Diaz F, Clarke CL, Voorhees EM (2014) TREC 2013 web track overview. In: Proceedings of TREC 2013

Eguchi K, Oyama K, Ishida E, Kando N, Kuriyama K (2003) Overview of the web retrieval task at the third NTCIR workshop. In: Proceedings of NTCIR-3

Eguchi K, Oyma K, Aizawa A, Ishikawa H (2004) Overview of the information retrieval task at NTCIR-4 WEB. In: Proceedings of NTCIR-4

Fujii A, Iwayama M, Kando N (2004) Overview of patent retrieval task at NTCIR-4. In: Proceedings of NTCIR-4

Fujii A, Iwayama M, Kando N (2005) Overview of patent retrieval task at NTCIR-5. In: Proceedings of NTCIR-5

Fujii A, Iwayama M, Kando N (2007) Overview of the patent retrieval task at the NTCIR-6 workshop. In: Proceedings of NTCIR-6, pp 359–365

Gey F, Larson R, Kando N, Machado J, Sakai T (2010) NTCIR-GeoTime overview: evaluating geographic and temporal search. In: Proceedings of NTCIR-8, pp 147–153

Gey F, Larson R, Machado J, Yoshioka M (2011) NTCIR9-GeoTime overview: evaluating geographic and temporal search: round 2. In: Proceedings of NTCIR-9, pp 9–17

Golbus PB, Zitouni I, Kim JY, Hassan A, Diaz F (2014) Contextual and dimensional relevance judgments for reusable SERP-level evaluation. In: Proceedings of WWW 2014, pp 131–142

Harman DK (2005) The TREC test collections. In: Voorhees EM, Harman DK (eds) TREC: experiment and evaluation in information retrieval, chap 2. The MIT Press, pp 21–52

Hawking D, Craswell N (2005) The very large collection and web tracks. In: Voorhees EM, Harman DK (eds) TREC: experiment and evaluation in information retrieval, chap 2. The MIT Press, pp 200–231

Hersh W, Buckley C, Leone T, Hickam D (1994) OHSUMED: an interactive retrieval evaluation and new large test collection for research. In: Proceedings of ACM SIGIR 1994, pp 192–201

Higashinaka R, Funakoshi K, Inaba M, Tsunomori Y, Takahashi T, Kaji N (2017) Overview of dialogue breakdown detection challenge 3. In: Proceedings of dialog system technology challenge 6 (DSTC6) workshop

Iwayama M, Fujii A, Kando N, Takano A (2003) Overview of patent retrieval task at NTCIR-3. In: Proceedings of NTCIR-3

Järvelin K, Kekäläinen J (2000) IR evaluation methods for retrieving highly relevant documents. In: Proceedings of ACM SIGIR 2000, pp 41–48

Järvelin K, Kekäläinen J (2002) Cumulated gain-based evaluation of IR techniques. ACM TOIS 20(4):422–446

Kando N, Kuriyama K, Nozue T, Eguchi K, Kato H, Hidaka S (1999) Overview of IR tasks at the first NTCIR workshop. In: Proceedings of NTCIR-1, pp 11–44

Kando N, Kuriyama K, Yoshioka M (2001) Overview of Japanese and English information retrieval tasks (JEIR) at the second NTCIR workshop. In: Proceedings of NTCIR-2

Kato MP, Yamamoto T, Manabe T, Nishida A, Fujita S (2017) Overview of the NTCIR-13 OpenLiveQ task. In: Proceedings of NTCIR-13, pp 85–90

Kato MP, Manabe T, Fujita S, Nishida A, Yamamoto T (2018) Challenges of multileaved comparison in practice: lessons from NTCIR-13 OpenLiveQ task. In: Proceedings of ACM CIKM 2018, pp 1515–1518

Kato MP, Nishida A, Manabe T, Fujita S, Yamamoto T (2019) Overview of the NTCIR-14 OpenLiveQ-2 task. In: Proceedings of NTCIR-14, pp 81–89

Kishida K (2005) Property of average precision and its generalization: an examination of evaluation indicator for information retrieval. Technical report NII-2005-014E. National Institute of Informatics
Kishida K, Chen KH, Lee S, Kuriyama K, Kando N, Chen HH, Myaeng SH, Eguchi K (2004) Overview of CLIR task at the fourth NTCIR workshop. In: Proceedings of NTCIR-4
Kishida K, Chen KH, Lee S, Kuriyama K, Kando N, Chen HH, Myaeng SH (2005) Overview of the CLIR task at the fifth NTCIR workshop (revised version). In: Proceedings of NTCIR-5
Kishida K, Chen KH, Lee S, Kuriyama K, Kando N, Chen HH (2007) Overview of the CLIR task at the sixth NTCIR workshop. In: Proceedings of NTCIR-6, pp 1–19
Korfhage RR (1997) Information storage and retrieval. Wiley, New Jersey
Liu Y, Song R, Zhang M, Dou Z, Yamamoto T, Kato M, Ohshima H, Zhou K (2014) Overview of the NTCIR-11 IMine task. In: Proceedings of NTCIR-11, pp 8–23
Luo C, Sakai T, Liu Y, Dou Z, Xiong C, Xu J (2017) Overview of the NTCIR-13 we want web task. In: Proceedings of NTCIR-13, pp 394–401
Maddalena E, Roitero K, Demartini G, Mizzaro S (2017) Considering assessor agreement in IR evaluation. In: Proceedings of ACM ICTIR 2017, pp 75–82
Mao J, Sakai T, Luo C, Xiao P, Liu Y, Dou Z (2019) Overview of the NTCIR-14 we want web task. In: Proceedings of NTCIR-14, pp 455–467
Oyama K, Eguchi K, Ishikawa H, Aizawa A (2004) Overview of the NTCIR-4 WEB navigational retrieval task 1. In: Proceedings of NTCIR-4
Pollock SM (1968) Measures for the comparison of information retrieval systems. Am Docum 19(4):387–397
Sakai T (2007a) Alternatives to bpref. In: Proceedings of ACM SIGIR 2007, pp 71–78
Sakai T (2007b) On penalising late arrival of relevant documents in information retrieval evaluation with graded relevance. In: Proceedings of EVIA 2007, pp 32–43
Sakai T (2007c) On the properties of evaluation metrics for finding one highly relevant document. IPSJ Digital Courier 3:643–660. https://doi.org/10.2197/ipsjdc.3.643
Sakai T (2014) Metrics, statistics, tests. In: PROMISE winter school 2013: bridging between information retrieval and databases. LNCS, vol 8173, pp 116–163
Sakai T (2017) Unanimity-aware gain for highly subjective assessments. In: Proceedings of EVIA 2017, pp 39–42
Sakai T (2018) Comparing two binned probability distributions for information access evaluation. In: Proceedings of ACM SIGIR 2018, pp 1073–1076
Sakai T (2019) Graded relevance assessments and graded relevance measures of NTCIR: a survey of the first twenty years. Technical report, https://arxiv.org/pdf/1903.11272
Sakai T, Kando N (2008) On information retrieval metrics designed for evaluation with incomplete relevance assessments. Inf Retriev 11:447–470
Sakai T, Robertson S (2008) Modelling a user population for designing information retrieval metrics. In: Proceedings of EVIA 2018, pp 30–41
Sakai T, Zeng Z (2019) Which diversity evaluation measures are "good"? In: Proceedings of ACM SIGIR 2019, pp 595–604
Sakai T, Kitani T, Ogawa Y, Ishikawa T, Kimoto H, Keshi I, Toyoura J, Fukushima T, Matsui K, Ueda Y, Tokunaga T, Tsuruoka H, Nakawatase H, Agata T, Kando N (1999) BMIR-J2: a test collection for evaluation of Japanese information retrieval systems 33(1):13–17
Sakai T, Kando N, Lin CJ, Mitamura T, Shima H, Ji D, Chen KH, Nyberg E (2008) Overview of the NTCIR-7 ACLIA IR4QA task. In: Proceedings of NTCIR-7, pp 77–114
Sakai T, Ishikawa D, Kando N (2010a) Overview of the NTCIR-8 community QA pilot task (part II): system evaluation. In: Proceedings of NTCIR-8, pp 433–457
Sakai T, Shima H, Kando N, Song R, Lin CJ, Mitamura T, Sugimoto M, Lee CW (2010b) Overview of NTCIR-8 ACLIA IR4QA. In: Proceedings of NTCIR-8, pp 63–93
Sakai T, Ferro N, Soboroff I, Zeng Z, Xiao P, Maistro M (2019) Overview of the NTCIR-14 CENTRE task. In: Proceedings of NTCIR-14, pp 494–509

Shang L, Sakai T, Lu Z, Li H, Higashinaka R, Miyao Y (2016) Overview of the NTCIR-12 short text conversation task. In: Proceedings of NTCIR-12, pp 473–484
Shang L, Sakai T, Li H, Higashinaka R, Miyao Y, Arase Y, Nomoto M (2017) Overview of the NTCIR-13 short text conversation task. In: Proceedings of NTCIR-13, pp 194–210
Song R, Zhang M, Sakai T, Kato MP, Liu Y, Sugimoto M, Wang Q, Orii N (2011) Overview of the NTCIR-9 INTENT task. In: Proceedings of NTCIR-9, pp 82–105
Yamamoto T, Liu Y, Zhang M, Dou Z, Zhou K, Markov I, Kato MP, Ohshima H, Fujita S (2016) Overview of the NTCIR-12 IMine-2 task. In: Proceedings of NTCIR-12, pp 8–26
Yasukawa M, Diaz F, Druck G, Tsukada N (2014) Overview of the NTCIR-11 Cooking recipe search task. In: Proceedings of NTCIR-11, pp 483–496
Zanibbi R, Aizawa A, Kohlhase M, Ounis I, Topić G, Davila K (2016) NTCIR-12 MathIR task overview. In: Proceedings of NTCIR-12, pp 299–308
Zeng Z, Kato S, Sakai T (2019) Overview of the NTCIR-14 short text conversation task: dialogue quality and nugget detection subtasks. In: Proceedings of NTCIR-14, pp 289–315
Zhang Y, Huang M (2019) Overview of the NTCIR-14 short text generation subtask: emotion generation challenge. In: Proceedings of NTCIR-14, pp 316–327

Chapter 2
Experiments on Cross-Language Information Retrieval Using Comparable Corpora of Chinese, Japanese, and Korean Languages

Kazuaki Kishida and Kuang-hua Chen

Abstract This paper describes research activities for exploring techniques of cross-language information retrieval (CLIR) during the NACSIS Test Collection for Information Retrieval/NII Testbeds and Community for Information access Research (NTCIR)-1 to NTCIR-6 evaluation cycles, which mainly focused on Chinese, Japanese, and Korean (CJK) languages. First, general procedures and techniques of CLIR are briefly reviewed. Second, document collections that were used for the research tasks and test collection construction for retrieval experiments are explained. Specifically, CLIR tasks from NTCIR-3 to NTCIR-6 utilized multilingual corpora consisting of newspaper articles that were published in Taiwan, Japan, and Korea during the same time periods. A set of articles can be considered a "pseudo" comparable corpus because many events or affairs are commonly covered across languages in the articles. Such comparable corpora are helpful for comparing the performance of CLIR between pairs of CJK and English. This comparison leads to deeper insights into CLIR techniques. NTCIR CLIR tasks have been built on the basis of test collections that incorporate such comparable corpora. We summarize the technical advances observed in these CLIR tasks at the end of the paper.

2.1 Introduction

A "comparable corpus" can be defined as multiple sets of documents, each in different languages, which approximately describe the same things or events. Unlike a parallel corpus, explicit alignments of words, sentences, paragraphs, or documents are not necessarily contained in the comparable corpus. In this sense, pairs of scientific abstracts written in Japanese and English that were used for retrieval experiments

K. Kishida (✉)
Keio University, Mita 2-15-45, Minato-ku, Tokyo 108-8345, Japan
e-mail: kz_kishida@keio.jp

K. Chen
National Taiwan University, No.1, Sec.4, Roosevelt Rd., Taipei 10617, Taiwan
e-mail: khchen@ntu.edu.tw

T. Sakai et al. (eds.), *Evaluating Information Retrieval and Access Tasks*,
The Information Retrieval Series 43,
https://doi.org/10.1007/978-981-15-5554-1_2

during the first and second *NACSIS Test Collection for Information Retrieval/NII Testbeds and Community for Information access Research (NTCIR)* evaluation cycles (i.e., NTCIR-1 and -2) as test documents can be considered document-linked comparable corpora.

Scholarly journals or conference proceedings published in Japan often ask authors to attach an English title and abstract to their Japanese paper to promote scientific communication. Such a set of Japanese and English titles and abstracts is a parallel corpus, in which explicit alignments of titles or abstracts may be included if the authors attempted to write the English title or abstract such that they were equivalent to those in Japanese. Even though all the authors did not necessarily do so, the set can be regarded as a comparable corpus at least.

In NTCIR-1, a corpus of such titles and abstracts was used for experiments of *cross-language information retrieval (CLIR)* in which English (E) documents were searched for Japanese (J) queries (i.e., a J to E bilingual search). Note that even if only a monolingual corpus in English is available, J to E bilingual searching can be tested by creating Japanese queries as search topics. However, Japanese and English comparable (or parallel) corpora allow us to compare results of J to E and E to J searching in a controlled setting, as the two target document sets in Japanese and English are topically similar. This type of comparison would play an important role in developing more sophisticated CLIR techniques. Actually, in NTCIR-2, a research task of E to J searching was added.

This policy of designing CLIR experiments based on comparable corpora had been maintained for NTCIR-3 to -6, in which CLIR between Chinese (C), Japanese (J), Korean (K), and English (E) was explored as one of the research tasks. More specifically, as target documents, NTCIR CLIR tasks used newspaper articles published in Taiwan, Japan, and Korea during the same time periods, which can be considered to be topically sufficiently comparable because they include many descriptions of common events and affairs occurring globally or locally in regions of East Asia. Actually, a comparison between pairs of CJKE languages based on such document sets largely contributed to the development of CLIR techniques between the CJKE languages even though the sets of the CJKE newspaper articles were "more loosely" comparable corpora than the sets of Japanese and English titles and abstracts in NTCIR-1 and -2.

This paper mainly describes research efforts of CLIR tasks from NTCIR-3 to -6. Specifically, construction of test collections based on so-called "pseudo comparable corpora" (i.e., time- and region-aligned newspaper article sets) and CLIR techniques that were explored by research groups participating in the NTCIR CLIR tasks are the focus. In addition, CLIR experiments in NTCIR-1 and -2 are briefly mentioned before reviewing the NTCIR CLIR tasks. The NTCIR-3 CLIR task started on September in 2001 and the NTCIR-6 CLIR task ended on May 2007. Therefore, readers can understand the technical development of CLIR among CJKE during the time period from a historical perspective.

2.2 Outline of Cross-Language Information Retrieval (CLIR)

Before describing research efforts of NTCIR CLIR tasks, this section gives a concise, general overview of CLIR operations. Grefenstette (1998), Oard and Diekema (1998), Nie (2010), and Peters et al. (2012) provide more in-depth coverage of CLIR. Note that this section is based on a review article (Kishida 2005), which includes an exhaustive reference list on CLIR techniques that this section describes.

2.2.1 CLIR Types and Techniques

Some form of CLIR is required when a search query and target documents are written in different languages. If only a single language is used in documents then the task is termed *bilingual information retrieval (BLIR)*. An example is J to E searching, in which only English documents are involved. In the case of *multilingual information retrieval (MLIR)*, the target set consists of documents in two or more languages. In the NTCIR CLIR tasks, the most difficult challenge was to search a set of documents in four languages (CJKE). Note that if a query is written in C then standard monolingual information retrieval (i.e., C to C searching) may be included as a part of MLIR on the CJKE documents. Monolingual IR was specifically referred to as single language IR (SLIR) in the NTCIR CLIR tasks. Therefore, NTCIR CLIR tasks had three subtasks: SLIR, BLIR and MLIR.

Generally, research efforts of CLIR can be traced back to a work by Gerald Salton in 1970 (Kishida 2005). Many researchers had attempted to develop CLIR techniques, particularly since the 1990s following popularization of the Internet. At that time, the main research task was to explore cross-lingual techniques for conventional ad hoc IR, which was also focused on by NTCIR CLIR tasks. However, it is possible to apply cross-lingual techniques to other applications related to ad hoc IR.

An important operation for CLIR is to translate a query and/or individual documents. If the query is perfectly translated into a language of the target documents via *machine translation (MT)* software then CLIR transforms back to normal monolingual IR. However, the translation is often incomplete because the queries are generally short and ambiguous (Oard and Diekema 1998). For example, when a query including only two single words "mercury earth" is entered into a search engine, the "mercury" in the source language has to be correctly translated into an equivalent that corresponds to a planet in the target language, not the chemical substance, in most cases. Sense disambiguation is often difficult because the queries may not contain sufficient contextual information for determining the correct meaning of each query term. To maintain the accuracy of the translation, it may be better to translate documents that are typically longer than the queries although document translation is more time-consuming in comparison to query translation. Another difficulty using

document translation is that index files of the IR system increase in size because translations have to be added as index terms.

Therefore, CLIR techniques typically consist of two main modules: (1) translation and (2) monolingual IR. Their effectiveness has an influence on the overall CLIR performance in the case that translation and monolingual searching independently work to generate a final search output, which would be a typical architecture of CLIR systems. However, there are IR models more sophisticatedly incorporating both modules. For example, *language modeling (LM)* can elegantly implement a CLIR operation by combining two conditional probabilities $p(s|t)$ and $p(t|d)$ during a process of computing document scores for ranked output where s and t denote a query term and a term in document d, respectively (Xu et al. 2001). Particularly, $p(s|t)$ is termed as translation probability.

2.2.2 Word Sense Disambiguation for CLIR

As previously exemplified by an instance of "mercury," *word sense disambiguation (WSD)* is important in CLIR. Typical methods for WSD in CLIR utilize (1) *part-of-speech (POS)* tags, (2) term co-occurrence statistics in the target document set, and (3) *pseudo relevance feedback (PRF)* techniques.

When POS tags are used, target terms having the same POS tags as the source term are selected from a set of candidates as final query terms. The candidate target terms can be easily obtained from a machine-readable bilingual dictionary.

In the case of utilizing term co-occurrence statistics in the target document set, the operation is more complicated. It is assumed that two translations t_1 and t_2 are extracted from a bilingual dictionary for a query term and that other translations u_1 and u_2 are similarly obtained for another term in the same query. If t_1 and u_1 are semantically correct translations in the context of the given query then it is expected that t_1 and u_1 co-occur more frequently in the target corpus than a pair of t_1 and u_2 and that of t_2 and u_1. Therefore, the co-occurrence frequencies aid in selecting final query terms in the target language, which is a basic assumption of the disambiguation method. When a large number of terms are included in an original source query, too many translations may be extracted from the dictionary. Because selection of final query terms is computationally expensive in such cases, some special techniques for solving the problem have been explored thus far (Kishida 2005).

Whereas the co-occurrence frequencies have to be computed before actual searching, such a type of preparatory work is not required for applying disambiguation techniques based on PRF. Instead, the searching operation is repeated during the process, which may be time-consuming in a real situation. That is, first, the target document collection is searched for a set of all translations that were obtained from a dictionary, and thereafter, final query terms are selected from the set of top-ranked documents (e.g., from the top 30 documents). Searching for the selected query terms is again repeated to obtain a final result.

Originally, the PRF attempts to expand a given query by adding some "significant" terms in the top-ranked documents to the query under an assumption that they also indicate an information need represented by the original query. The newly added terms may mainly contribute to enhancing the recall ratio. In the context of disambiguation for CLIR, it is expected that documents including a semantically correct combination of target terms (e.g., t_1 and u_1 in the aforementioned example) are at a higher position in a ranked list of the first search (Ballesteros and Croft 1997). As a result, terms that co-occur in the top-ranked documents tend to be selected, which has the same effect as using term co-occurrence statistics. Thus, the selection based on the top-ranked documents works incidentally as a system for disambiguation. Note that the co-occurrences in the top-ranked documents are limited to a local context of the original query, unlike term co-occurrence statistics in the entire document set. Final query terms are typically selected according to term weights that are calculated using a formula of standard PRF techniques (Kishida 2005).

2.2.3 Language Resources for CLIR

As mentioned previously, a typical language resource for implementing CLIR is a machine-readable bilingual dictionary or MT software. When both the dictionary and the software are not available for a given pair of source and target languages, it is possible to apply a pivot language approach. For example, even if a resource between Japanese and Swedish (S) is not found, J to English and English to S resources allow us to execute J to S bilingual searching, where English is a pivot language. More specifically, by translating each Japanese query term into English equivalents and converting them again to Swedish terms, a final Swedish query can be obtained. Thus, the resulting Swedish query can be used for retrieval of the Swedish documents. Because English is an international language, many language resources related to English are actually available.

In addition, parallel corpora play an important role in CLIR. Without a dictionary or MT software, CLIR can be executed by searching a parallel corpus for a query written in the source language. That is, because textual data that were found via searching have another part in the target language, it is possible to extract final query terms in the target language from the data. Additionally, a parallel corpus consisting of sentence alignments can be used for estimating translation probabilities, in which the well-known IBM Model 1 for statistical MT has often been applied. The list of the translation probabilities works as a bilingual dictionary, and is indispensable for LM-based CLIR (see Sect. 2.2.1).

Of course, standard language processing tools such as a POS tagger (or a morphological analyzer), a stemmer, and a named entity recognizer are also employed in CLIR.

2.3 Test Collections for CLIR from NTCIR-1 to NTCIR-6

A main contribution of the NTCIR CLIR tasks is to examine whether or not the CLIR techniques that were reviewed in the previous section can be applied to the CJK languages and to enhance the techniques by tailoring them to situations in which CJK languages are used. When the NTCIR CLIR task started, the Chinese language had been already explored in the *Text REtrieval Conference (TREC)* (Voorhees and Harman 2005). In contrast, a systematic and large-scale CLIR experiment related to the Japanese and Korean languages would be considered as an original NTCIR contribution. This section provides a simple overview of test collections on which various trial-and-error attempts were made in NTCIR from the very beginning.

2.3.1 Japanese-English Comparable Corpora in NTCIR-1 and NTCIR-2

As previously mentioned, a set of Japanese and English titles and abstracts in conference proceedings that were published by Japanese academic societies was a source of documents in NTCIR-1. More specifically, in total, 339,483 bibliographic records of conference papers were collected. Because the set included three types of records having (1) only Japanese abstracts, (2) only English abstracts, and (3) both Japanese and English abstracts, the set of Japanese documents (J collection) and the set of English documents (E collection) were constructed as a subset of the whole set (JE collection). Research groups participating in NTCIR-1 were able to use the three sets during IR experiments (Kando et al. 1999).

All search requests for the experiments (i.e., search topics) were written in Japanese (30 topics for training and 53 topics for evaluation). Therefore, it was possible for the participants to examine only J to E bilingual searching as CLIR experiments. The NTCIR-1 conference was held in September of 1999, which would be the first opportunity for discussing internationally CLIR issues related to Japanese language.

In NTCIR-2, by adding bibliographic records of some scientific reports published in Japan, the document sets were substantially extended. The English and Japanese versions of 49 search topics were prepared by the task organizers (Kando et al. 2001). The test collection allowed the participants to experiment in E to J and J to E bilingual searching and in J to JE and E to JE multilingual searching.

2.3.2 *Chinese-Japanese-Korean (CJK) Corpora from NTCIR-3 to NTCIR-6*

Based on the knowledge that was obtained by the efforts of NTCIR-1 and -2, more sophisticated CLIR experiments involving C, J, K, and E were started as an independent task beginning with NTCIR-3. In the CLIR tasks from NTCIR-3 to -6, newspaper articles that were collected from various news agencies in East Asia were employed as target documents. Each record of the articles included its headline and full text. Table 2.1 summarizes the document sets; the number of documents is indicated in Table 2.2. Note that the CLIR task of NTCIR-6 had two stages (i.e., stages 1 and 2). The purpose of stage 2 was to obtain a more reliable measurement of search performance. Newspaper articles published in 1998 and 1999 were basically used for experiments in NTCIR-3 and -4 whereas newspaper articles for NTCIR-5 and stage 1 in NTCIR-6 were from 2000 and 2001.

For some reason, only the Korean document set in NTCIR-3 consisted of newspaper articles in 1994. However, from NTCIR-4, newspaper articles matching time periods (i.e., 1998–99 and 2000–01) were provided as CJKE document sets for experiments (English documents were out of scope in NTCIR-6). As previously discussed, the sets can be considered as types of comparable corpora because the newspaper articles in the sets were commonly concerned with worldwide or East Asian events and affairs of the time, allowing a CLIR performance comparison between the pairs of CJKE languages partly because documents in the individual languages are topically homogeneous to some extent. Notably, the Chinese documents were represented by only traditional Chinese characters, not simplified ones.

A newspaper article is typically written for general audiences; its text is relatively plain and shorter in comparison to that of scientific or technical papers. There is no explicit structure in the text of newspaper articles except for a headline and paragraphs, which is different from XML documents having a more complex structure. Additionally, newspaper article records in NTCIR CLIR tasks did not include any

Table 2.1 Document sets used by NTCIR CLIR tasks[a]

	Period of tasks	Date of newspaper articles	Set for MLIR
NTCIR-3	2001–02	C, J, E: 1998–99, K:1994	CJ, CE, JE, CJE
NTCIR-4	2003–04	C, J, K, E: 1998–99	CJE, CJKE
NTCIR-5	2004–05	C, J, K, E: 2000–01	CJKE
NTCIR-6	2006–07		
Stage1		C,J,K:2000-01	CJK
Stage2		NTCIR-3, –4, –5 test collections[b]	

[a]Search topics in C, J, K, and E were created for the document sets
[b]In stage 2 of NTCIR-6, a cross-collection analysis was attempted

Table 2.2 Number of records in document sets in the NTCIR-3 to -6 CLIR tasks

Language	No. of records	Usage (denoted by the mark x)				
		–3	–4	–5	–6	
					Stage 1	Stage 2
1994						
Korean	Korea Economic Daily: 66,146	x				x
1998–99						
Chinese	UDN[a]+others: 381,375	x	x			x
Japanese	Mainich: 220,078	x	x			x
	Yomiuri: 373,558		x			x
Korean	Hankookilbo+Chosunilbo: 254,438		x			x
English	Mainichi Daily+EIRB[b]: 22,927	x	x			
	Xinhua+others: 324,449		x			
2000–01						
Chinese	UDN[a]+others: 901,446			x	x	x
Japanese	Mainichi+Yomiuri: 858,400			x	x	x
Korean	Hankookilbo+Chosunilbo: 220,374			x	x	x
English	Xinhua+others: 259,050			x		

[a]UDN: United Daily News
[b]EIRB: Taiwan News and China Times English News

topic keywords such as descriptors that are often assigned to bibliographic records of scientific papers. Today various types of documents are exploited for current research on IR or related areas, but the test collections using such newspaper articles still provide IR researchers a sound experimental setting for examination of fundamental techniques that underlie more complicated searches.

2.3.3 CJKE Test Collection Construction

Test collections incorporating the CJKE documents were constructed according to a traditional pooling method explored by TREC. In general, a test collection consists of three components: a document set, topic set (set of search requests), and answer set (result of relevance judgments). By employing the answer set, metrics for evaluating IR results such as precision or recall can be computed. When calculating the recall, it is required to determine all relevant documents included in the document set, which is typically impossible for large-scale document sets. Therefore, the pooling method was developed for using such large sets. Figure 2.1 shows an operational model for IR evaluation based on the pooling method.

First, a document set such as that shown in Table 2.1 is sent to participants in the tasks for implementing into their own IR systems. Then, task organizers deliver a topic set to participants and ask them to submit search results by the designated day.

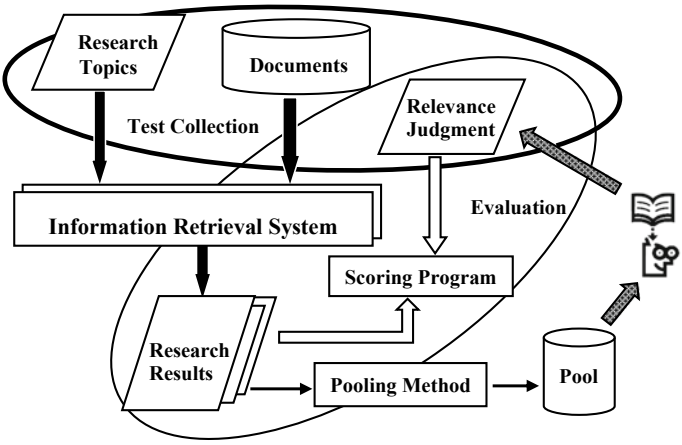


Fig. 2.1 Construction of test collection and evaluation

Under management of the task organizers, the degree of relevance is judged for each pair of a topic and a document included in the search results, by which an answer set is obtained. Finally, the search performance of the participating IR system is scored based on the answer set. By checking the scores, the advantages or disadvantages of IR theories or techniques are clarified. Because the relevance judgment is completed for pooled documents that are extracted from the search results that participants submitted, and not for the entire set of documents, this procedure for creating the answer set is termed the pooling method, which is an efficient means for constructing a large-scale test collection. Strictly speaking, scores of some evaluation metrics obtained from this procedure are only approximations because the entire set is not examined. However, a comparison of search effectiveness between the IR systems or models within the test collection is sufficiently feasible.

The organizers of the NTCIR CLIR tasks consisted of IR researchers in Taiwan, Japan, and Korea who collaboratively worked in designing the research tasks, creating the topics, managing the relevance judgment process, and evaluating the participating IR systems. The authors of this paper were members of the organizer group.

In our experience, it was difficult to create topics that were effective for measuring CLIR performance between the CJKE languages compared to a case of simple monolingual IR. The typical procedure for topic creation in the NTCIR CLIR tasks was as follows:

1. Topic candidates were created in Taiwan, Japan, and Korea, respectively, and were translated into English.
2. The English candidates were again translated into Chinese, Japanese, and Korean as necessary, and the task organizers preliminarily examined whether or not relevant documents were sufficiently included in the C, J, K, and E document sets.
3. Final topics were selected based on the preliminary examination result.

This complicated procedure was adopted for using topics commonly on all document sets in the CJKE languages, by which comparisons of search performance between bilingual searches of CJKE (e.g., between C to J and J to C searches that are related to processing of Chinese and Japanese texts) became easier.

Search topics created for NTCIR CLIR tasks can be approximately classified into two types: 1) event-based topics and 2) general concept-based topics. Event-based topics typically contain one or more proper nouns of a social event, geographic location, or person. An example is "Find reports on the G8 Okinawa Summit 2000" (ID 005 in NTCIR-5). If a CLIR system cannot find any corresponding translation of the proper noun during its process then the search performance is expected to be low. This is generally termed an *out-of-vocabulary (OOV)* problem.

Meanwhile, it may be relatively easier to find translations of a general concept, but CLIR systems often need to disambiguate translation candidates for the concept. For example, a correct translation in the context of the search topic often has to be selected from many terms listed in a bilingual dictionary (see Sect. 2.2.2). An instance of the general concept-based topics is "Find documents describing disasters thought to be caused by abnormal weather" (ID 044 in NTCIR-5). Even though "weather" has a relatively definite meaning, many translations are actually enumerated in an E to J dictionary and selection of final translations substantially affects CLIR performance. The task organizers considered a careful balance of the two topic types for allowing researchers to develop more effective systems. During the topic creation process, approximately 50 topics were included in each of the test collections for NTCIR-3 to -6, respectively.

Needless to say, jobs for pooling documents also are not easy. If the pool (i.e., a document set to be checked during a process of relevance judgment) is too large then it is impossible for an assessor to maintain consistent judgment for all documents. For avoiding this problem, the document pool size has to be appropriately adjusted when extracting top-ranked documents from the search results of each participant. This is a special matter of so-called pooling depth (Kuriyama et al. 2002).

Also, a system of relevance judgments developed by *National Institute of Informatics (NII)*, of which name was NACSIS at the time, was used for providing the assessors with a comfortable human-machine interface for the judgment task, which contributed to enhancing consistency and reliability of the judgment results. A windows-based assessment system created in Taiwan for the special purpose is explained by Chen (2002).

2.3.4 IR System Evaluation

During the process of relevance judgment, the assessors evaluated each document using a four-point scale: 1) highly relevant, 2) relevant, 3) partially relevant, and 4) irrelevant. The IR research field has a long history of studying relevance concepts and operational assessment of them. A multi-grade assessment based on the four-

point scale was adopted in ad hoc IR tasks beginning with NTCIR-1 after carefully examining discussions in the literature of relevance.

However, evaluation metrics based on a multi-grade assessment such as the *Normalized Discounted Cumulative Gain (nDCG)* were not yet popular at the time of NTCIR-1 to -6 and the main indicator for evaluating IR systems was *Mean Average Precision (MAP)*.[1] For calculating the average precision, the four-point scale has to be reduced to a binary scale. When "highly relevant" and "relevant" were considered to be relevant and the others to be irrelevant, it was specifically termed "rigid" relevance in the NTCIR CLIR tasks. If "partially relevant" was included in the relevant category then "relaxed" relevance was used. Therefore, in NTCIR CLIR tasks, two MAP scores were typically computed for a single search result based on rigid and relaxed relevance, respectively. Sakai (2020) summarizes evaluation metrics and methods in the overall NTCIR project.

2.4 CLIR Techniques in NTCIR

This section briefly summarizes typical techniques used in CLIR tasks from NTCIR-3 to -6. For knowing details of the techniques and systems, overviews of each task that were published at NTCIR conferences are helpful (Chen et al. 2002; Kishida et al. 2004, 2005, 2007). Lists of research groups participating in each task are also included in the overviews.

2.4.1 Monolingual Information Retrieval Techniques

IR systems of groups participating in NTCIR CLIR tasks typically have two independent components for 1) monolingual IR and 2) translation as explained in Sect. 2.2.1. Because computer processing of Chinese, Japanese, and Korean textual data had not yet been sufficiently developed at the time, NTCIR CLIR tasks also contributed to obtaining useful knowledge regarding CJK text processing for monolingual IR (or single language IR: SLIR). The resulting SLIR performance improvement can be considered as an achievement in the NTCIR CLIR tasks.

Particularly, sentences or phrases in the CJK texts have no explicit word boundary, which is a characteristic that is different from that of English texts (note that Korean texts include white spaces as a delimiter between phrasal units). To construct index files in SLIR systems for these languages, either

1. Word-based indexing, or
2. Overlapping character bigrams (i.e., n-grams when $n = 2$)

[1] Only in NTCIR-6, nDCG was used for evaluating CLIR performance as a trial.

were typically used in the NTCIR CLIR tasks. For word-based indexing, some groups employed morphological analyzers, whereas index terms were identified from texts by simply matching with entries of machine-readable dictionaries in some other systems.

Extracting character bigrams from texts are a characteristic during the indexing process of East Asian languages. Assume that a Japanese sentence is "ABCDE" where A, B, C, D, and E are a Japanese character, respectively. In the case of overlapping character bigrams, "AB," "BC," "CD," and "DE" are automatically selected as index terms. This was known as an effective method for processing texts that were represented by ideograms. Although character unigrams (i.e., n-grams when $n = 1$) are extracted from the target text in the current Internet search engines or some online public access catalog (OPAC) systems, $n = 2$ was used in NTCIR CLIR tasks.

By utilizing an index file constructed according to an indexing method, documents have to be ranked by the degree of relevance to each search query. The relevance degree is operationally estimated in the system based on a retrieval model. In NTCIR CLIR tasks, participant groups typically adopted some standard and well-known models such as the vector space mode (VSM), Okapi BM25, LM, INQUERY, PIRCS, or logistic regression model. In addition, *query expansion (QE)* by PRF or techniques using external resources (e.g., statistical thesauri based on term co-occurrence statistics or web pages) were incorporated for enhanced search performance. The retrieval models and QE techniques were originally developed in the USA or Europe mainly for English IR. NTCIR CLIR tasks provided good opportunities for systematically confirming their effectiveness for IR of CJK languages.

2.4.2 Bilingual Information Retrieval (BLIR) Techniques

Section 2.2 reviewed typical CLIR techniques, which were also utilized in NTCIR CLIR tasks. Dictionaries and MT software that were employed by participants in the NTCIR-4 CLIR task were extensively enumerated in Kishida et al. (2004).

Specifically, important problems to be solved for translation in CLIR among CJKE were as follows.

1. Query translation versus document translation: Most participating groups adopted a means of translating search topics (queries), whereas some explored "pseudo" document translation in which terms in target documents were simply replaced with equivalents in another language using a bilingual dictionary (i.e., not MT). Additionally, search performance may be improved by combining search results from both the query and document translations because it is possible that the probability of successful matching of terms between a topic and a relevant document increases. This technique was attempted by one group.
2. Pivot language approach: English was typically used as a pivot language for CLIR among CJK, whereas one group attempted bilingual searching via Japanese.

Selection of the pivot language depends on translation resources such as MT software.

3. OOV problem: As previously mentioned, when a term representing an important concept in a search topic is not included in the dictionaries for translation, search performance largely degrades. Some search topics in the NTCIR CLIR tasks contained names related to current events or affairs and they were not often covered if using only a standard bilingual dictionary (see Sect. 2.3.3). For solving this problem, some groups attempted to extract translations from web pages for the unknown term.
4. Automatic transliteration: In general, when a word in a foreign language is imported, transliteration is often used without semantically representing the word in its own language. For example, an English word "hotel" is transliterated into three *Katakana* characters corresponding phonetically to "ho," "te," and "ru" in Japanese. Although popular *Katakana* words are listed in standard bilingual dictionaries, an OOV problem occurs if this is not the case. At this time, an English word may be automatically converted into a *Katakana* word (and vice versa) via heuristic rules phonetically measuring the similarity between them (Fujii and Ishikawa 2001). This type of automatic transliteration was explored in the NTCIR CLIR tasks.
5. Conversion of *Kanji* character codes: An idea similar to automatic transliteration is automatic conversion of *Kanji* characters between Chinese and Japanese. In the NTCIR CLIR tasks, one group attempted to convert traditional Chinese characters encoded by the BIG5 character code into Japanese characters represented by Extended Unix Code-Japanese (EUC-JP).
6. Term disambiguation (or WSD): A typical method for term disambiguation was to use statistical information of term co-occurrences in the set of target documents. In addition, many CLIR systems incorporated a PRF process, which had an effect of increasing the rank of documents that included a combination of correct translations (see Sect. 2.2.2). Both methods do not require any external resource. In contrast, some external resources such as web pages or parallel corpora were also applied for term disambiguation by some groups. For example, one system attempted to select final query terms based on web pages that were extracted from a web category to which the search topic corresponded.

As a technique for improving CLIR performance, pre-translation PRF was explored in the NTCIR CLIR tasks. That is, if a corpus in the source language of an original query is available as an external resource, then a PRF operation on the external resource may result in a set of more useful query terms, termed the pre-translation PRF. After obtaining a "richer" representation of the original query in the source language using it, standard CLIR is executed based on the modified query. More common was to include PRF in the form of post-translation PRF after the retrieval process proper. A combination of pre- and post-translation PRF was used in some groups in the NTCIR CLIR tasks.

In addition, participants in the NTCIR CLIR tasks attempted to address other various challenges such as document re-ranking, QE via a statistical thesaurus, trans-

Table 2.3 Best MAP scores of SLIR and BLIR in NTCIR-6: Rigid relevance, DESC field[a]

	Documents (X)		
Search topics	Chinese	Japanese	Korean
Monolingual (baseline)	0.313 (100%)	0.325 (100%)	0.454 (100%)
BLIR			
Chinese (C to X search)	–	0.312 (95.8%)	N/A
Japanese (J to X search)	0.078 (24.7%)	–	0.287 (63.2%)
Korea (K to X search)	0.102 (32.6%)	0.267 (82.1%)	–
English (E to X search)	0.191 (61.0%)	0.307 (94.4%)	0.292 (64.3%)

[a] A short sentence describing each search topic was included in a <DESC> element of an XML file of the topics. The sentence was used as a query for executing searches in this table

lation probability estimation, the use of an ontology to enhance the effectiveness of mono- or cross-lingual IR. Although similar research efforts had already been completed in TREC or CLEF (Cross-Language Evaluation Forum, at that time), a special aspect of the NTCIR CLIR tasks was the larger differences in language types between English and CJK. For example, nobody would deny that "linguistic distance" between English and Japanese is greater than that between English and Swedish. The special characteristics of CJK as languages may have contributed to unique modification or refinement of CLIR techniques (e.g., automatic transliteration).

It is difficult to concisely present an overview of search performance attained by CLIR systems participating in CLIR tasks from NTCIR-3 to -6. Only the best performance of the SLIR and BLIR subtasks in NTCIR-6 is shown in Table 2.3 (Kishida et al. 2007), which provides the best MAP scores based on "rigid" relevance by each language combination. When comparing the MAP scores between monolingual searching (SLIR) and BLIR, it appears that BLIR to Japanese documents was more successful than to other languages because the percentages were 95.8% for C, 82.1% for K and 94.4% for E search topics. However, the percentage highly depends on the system performance of the research group participating in the task at the time; thus, Table 2.3 does not indicate any research finding based on a scientific examination. This table is only an example for superficially understanding an aspect of NTCIR CLIR tasks. Readers that are interested in the search runs of Table 2.3 can refer to Kishida et al. (2007) for more detail.

2.4.3 Multilingual Information Retrieval (MLIR) Techniques

Two types of MLIR strategies are most commonly used:

- (A) All documents and the query are translated into a single language (e.g., English), and then monolingual IR is executed thereafter and
- (B) BLIR is repeated for all pairs of document language and query language, and then all search results are finally merged into a single ranked document list.

Fewer research groups participated in MLIR subtasks compared to those in SLIR and BLIR subtasks, and most adopted the type B strategy. In the strategy, an important choice is how search results (actually, individual ranked lists by language pairs) are merged, which can be considered as a type of data fusion problem. The merging operation is also important for applications other than MLIR.

Typical merging methods in NTCIR CLIR tasks are as follows.

1. Round-robin merging: Documents are repeatedly selected from the top of each ranked list in a sequence.
2. Raw score merging: All documents are merged and re-ranked according to document scores calculated by an IR model.
3. Normalized score merging: Document scores that are calculated by an IR model are normalized before the documents are merged and re-ranked.

When applying these methods, there are some difficulties. For example, if the number of relevant documents included in the C, J, K, and E components is significantly different, then the difference makes the MLIR more difficult. In this situation, an "absolute" relevance probability that is effective over all languages may have to be estimated for each document to achieve better performance. Braschler (2004) discusses the other difficulties of MLIR. Actually, MAP scores of MLIR were typically lower than those of SLIR and BLIR in the NTCIR CLIR tasks.

2.5 Concluding Remarks

Research activity for exploring the cross-lingual ad hoc IR of newspaper articles in the NTCIR project ended at the CLIR task in NTCIR-6, for which the conference was held in May of 2007. Thereafter, during the 2010s, the Internet search engine performance remarkably improved, more easily allowing one to search Chinese, Japanese, and Korean documents in situations of monolingual IR. In addition, several excellent tools or resources for language processing have become available. Specifically, new technologies such as statistical machine translation or neural machine translation have drastically enhanced the effectiveness of MT.

The current state of monolingual IR and language processing has largely changed from the time of the NTCIR CLIR tasks. Experimental findings that were obtained from the tasks have contributed to such technological advances and aided researchers

in developing a more sophisticated CLIR system based on the current technologies of monolingual IR and language processing. In addition, the authors believe that experience of constructing test collections that consist of comparable corpora in NTCIR CLIR tasks is useful for further development of IR theories and techniques in multilingual environments.

Acknowledgements Many researchers in Taiwan, Japan, and Korea worked collaboratively as organizers in managing NTCIR CLIR tasks as follows: Hsin-Hsi Chen, Koji Eguchi, Noriko Kando, Kazuko Kuriyama, Hyeon Kim, Sukhoon Lee, and Sung Hyon Myaeng (as well as the authors of this paper). Additionally, in NTCIR-1 and -2, Toshihiko Nozue, Souichiro Hidaka, Hiroyuki Kato, and Masaharu Yoshioka also joined to organize the IR tasks. This paper attempted to summarize some aspects of valuable activities and efforts by the organizers and by all participants in the research tasks.

References

Ballesteros L, Croft WB (1997) Phrasal translation and query expansion techniques for cross-language information retrieval. In: Proceedings of the 20th annual international ACM SIGIR conference on research and development in information retrieval, pp 84–91

Braschler M (2004) Combination approaches for multilingual text retrieval. Inf Retr 7(1/2):183–204

Chen KH (2002) Evaluating Chinese text retrieval with multilingual queries. Knowl Organ 29(3/4):156–170

Chen KH, Chen HH, Kando N, Kuriyama K, Lee S, Myaeng SH, Kishida K, Eguchi K, Kim H (2002) Overview of CLIR task at the third NTCIR workshop. In: Proceedings of the Third NTCIR workshop on research in information retrieval, automatic text summarization and question answering

Fujii A, Ishikawa T (2001) Japanese/English cross-language information retrieval: Exploration of query translation and transliteration. Comput Human 35(4):389–420

Grefenstette G (1998) Cross-language information retrieval. Springer, Berlin

Kando N, Kuriyama K, Nozue T, Eguchi K, Kato H, Hidaka S (1999) Overview of IR tasks. In: Proceedings of the First NTCIR workshop on research in Japanese text retrieval and term recognition, pp 11–44

Kando N, Kuriyama K, Yoshioka M (2001) Overview of Japanese and English information retrieval tasks (JEIR) at the second NTCIR workshop. In: Proceedings of the second NTCIR workshop on research in Chinese and Japanese text retrieval and text summarization

Kishida K (2005) Technical issues of cross-language information retrieval: a review. Inf Process Manag 41(3):433–455

Kishida K, Chen KH, Lee S, Kuriyama K, Kando N, Chen HH, Myaeng SH, Eguchi K (2004) Overview of CLIR task at the fourth NTCIR workshop. In: Proceedings of the fourth NTCIR workshop on research in information access technologies: information retrieval, question answering and summarization

Kishida K, Chen KH, Lee S, Kuriyama K, Kando N, Chen HH, Myaeng SH (2005) Overview of CLIR task at the fifth NTCIR workshop. In: Proceedings of the fifth NTCIR workshop meeting on evaluation of information access technologies: information retrieval, question answering and cross-lingual information access

Kishida K, Chen KH, Lee S, Kuriyama K, Kando N, Chen HH (2007) Overview of CLIR task at the sixth NTCIR workshop. In: Proceedings of the 6th NTCIR workshop meeting on evaluation of information access technologies: information retrieval, question answering and cross-lingual information access

Kuriyama K, Kando N, Nozue T, Eguchi K (2002) Pooling for a large-scale test collection: an analysis of the search results from the first NTCIR workshop. Inf Retr 5(1):41–59
Nie JY (2010) Cross-language information retrieval. Morgan & Claypool Publishers
Oard DW, Diekema AR (1998) Cross-language information retrieval. Ann Rev Inf Sci Technol 33:223–256
Peters C, Braschler M, Clough P (eds) (2012) Multilingual information retrieval. Springer, Berlin
Sakai T (2020) Graded relevance. In: Sakai T, Oard DW, Kando N (eds) Evaluating information retrieval and access tasks. Springer, Singapore. The Information Retrieval Series, (in this book)
Voorhees E, Harman DK (eds) (2005) TREC: experiment and evaluation in information retrieval. MIT Press
Xu J, Weischedel R, Nguyen C (2001) Evaluating a probabilistic model for cross-lingual information retrieval. In: Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval, pp 105–110

Chapter 3
Text Summarization Challenge: An Evaluation Program for Text Summarization

Hidetsugu Nanba, Tsutomu Hirao, Takahiro Fukushima, and Manabu Okumura

Abstract In Japan, the Text Summarization Challenge (TSC), the first text summarization evaluation of its kind, was conducted in 2000–2001 as a part of the NTCIR (NII-NACSIS Test Collection for IR Systems) Workshop. The purpose of the workshop was to facilitate collecting and sharing text data for summarization by researchers in the field and to clarify the issues of evaluation measures for summarization of Japanese texts. After that, TSC has been held every 18 months as a part of the NTCIR project. In this chapter, we describe our TSC series, the data used, and the evaluation methods for each task, and the features of TSC evaluation.

3.1 What is Text Summarization?

The ever-growing amount of information forces us to read through a great number of documents in order to extract relevant information from them. To cope with this situation, research on text summarization has attracted much attention recently, producing many studies in this field.[1]

[1]Many survey papers are now available on text summarization, e.g., Gambhir and Gupta (2017), Allahyari et al. (2017), Nazari and Mahdavi (2019).

H. Nanba
Faculty of Science and Engineering, Chuo University, Tokyo, Japan
e-mail: nanba@kc.chuo-u.ac.jp

T. Hirao
NTT Communication Science Laboratories, Kyoto, Japan
e-mail: tsutomu.hirao.kp@hco.ntt.co.jp

T. Fukushima
Department of International Liberal Arts, Otemon Gakuin University, Ibaraki, Japan
e-mail: fukusima@otemon.ac.jp

M. Okumura (✉)
Tokyo Institute of Technology, Tokyo, Japan
e-mail: oku@pi.titech.ac.jp

T. Sakai et al. (eds.), *Evaluating Information Retrieval and Access Tasks*,
The Information Retrieval Series 43,
https://doi.org/10.1007/978-981-15-5554-1_3

As research on text summarization is a hot topic in Natural Language Processing (NLP), we also see the needs to discuss and clarify issues of how to evaluate text summarization systems. In Japan, the Text Summarization Challenge (TSC), the first text summarization evaluation of its kind, was conducted in 1999–2000 as a part of the NTCIR (NII-NACSIS Test Collection for IR Systems) Workshop. The aim of TSC was to facilitate collecting and sharing text data for summarization by researchers in the field and to clarify the issues of evaluation measures for summarization of Japanese texts.

Since that time, TSC[2] was held twice more, every 18 months, as a part of the NTCIR project. Multiple document summarization as one of the tasks was included for the first time at the TSC2 in 2002.

As we mention in Sect. 3.5, the contributions of our TSC can be considered as follows:

- We proposed a new evaluation method, evaluation by revision, that evaluates summaries by measuring the degree of revisions of the system results.
- We proposed a new evaluation method for multiple documents summarization that enables us to measure the effectiveness of redundant sentence reduction in the systems.

In the following sections, we first introduce the types of summarization and the evaluation methods in general. Then, we describe our TSC series, the data used, and the evaluation methods for each task. Finally, we summarize the contributions of the TSC evaluations.

3.2 Various Types of Summaries

Text summarization is a task of producing a shorter text from the source, while keeping the information content of the source. Summaries are the results of such a task. Perhaps, one of the most widely used summaries in the world today is the snippets that Web search engines display for each Web page. Sparck Jones (1999) discussed several ways to classify summaries. The following three factors are considered to be important for text summarization research:

Input factors: text length, genre, and single versus multiple documents,
Purpose factors: who the user is, and the purpose of summarization,
Output factors: running text or headed text, etc.

Summaries can be classified with respect to the number of the source texts (single document versus multiple document summarization), and with respect to whether they are tailored to particular users. Early research in summarization was primarily based on single-document summarization, in which systems produced a summary from a single-source document. However, another task has been later introduced into

[2] http://www.lr.pi.titech.ac.jp/tsc/index-en.html.

text summarization, that is based on multiple source documents. In multi-document summarization, several documents sharing a similar topic are taken as the input. The task of multi-document summarization can be considered more difficult than the single-document one, because the systems would need to remove any redundancies across multiple documents and then make the contents from multiple documents into a coherent summary.

If summaries are targeted for specific users, they are called user focused, and if they are intended for users in general, they are called generic. Query-focused summaries are another name for user focused summaries. In query-based summarization, the summary is generated by selecting sentences that correspond to the user's query (Tombros and Sanderson 1998). Sentences that are relevant to the query have a higher chance to be extracted for the final summary. In terms of summarization purpose, summaries can be either indicative or informative. Users can make use of indicative summaries before referring to the source, e.g., to judge relevance of the source text. On the other hand, users may use summaries in place of the source text (informative summaries). The snippets of Web search engines are a good example of indicative and query-focused summaries.

As pointed out by Mani and Maybury (1999), summaries can be also classified into extracts and abstracts, depending on how they are composed. Conventional text summarization systems produce summaries by using sentences or paragraphs as a basic unit, giving them a degree of importance, sorting them based on the importance, and gathering the important sentences. In short, summaries that are constructed of a set of important sentences extracted from the source text are called extracts. In contrast, summaries that may contain newly produced texts are called abstracts. Therefore, abstractive summarization can be much more complex than extractive summarization.

3.3 Evaluation Metrics for Text Summarization

Evaluation methods for text summarization can be largely divided into two categories: intrinsic and extrinsic. The quality of summaries can be judged directly based on some norms; typically, ideal summaries are produced by hand, or important sentences are selected by hand. Then, the quality of summaries is evaluated by comparing them with the human-produced summaries (intrinsic evaluation). The quality of a summary can also be judged by measuring how it influences the achievement of some other task (extrinsic evaluation). Mani and Maybury (1999) stated such tasks can be question-answering, reading comprehension, as well as relevance judgement of a document to a certain topic indicated by a query.

Relevance judgement: determines whether it is possible to judge whether the presented document is relevant to a user's topic, that can be indicated by her query, by reading the summary.

Reading comprehension: determines whether it is possible to correctly complete a multiple-choice test after reading the summary.

There are two measures for intrinsic evaluation: Quality and informativeness (Gambhir and Gupta 2017). The first measure checks the summary for grammatical errors, redundant information, and structural coherence. Here, the linguistic aspects of the summary are considered. In the Document Understanding Conference (DUC) and Text Analysis Conference (TAC), five questions based on linguistic quality are employed for evaluating summaries, which are non-redundancy, focus, grammaticality, referential clarity, and structure and coherence. Human assessors evaluate summaries manually by assigning a score to the summary, on a five-point scale.

For intrinsically evaluating the informativeness of a summary, the most popular metrics are precision, recall, and F-measure; they measure the overlap between human-made summaries and automatically generated machine-made summaries.

Precision: determines what fraction of the sentences selected by the system are correct.

Recall: determines what proportion of the sentences chosen by humans are selected by the system.

F-measure: is computed by combining recall and precision.

3.4 Text Summarization Evaluation Campaigns Before TSC

The first conference where text summarization systems were evaluated was held at the end of the 90's and was named the TIPSTER Text Summarization Evaluation (SUMMAC) (Mani and Maybury 1999). At that time, text summaries were evaluated using two extrinsic and one intrinsic methods. Two main extrinsic evaluation tasks were defined: adhoc and categorization. In the adhoc task, the focus was on indicative summaries which were tailored to a particular topic, and they were used for relevance judgement. In the categorization task, the evaluation sought to find out whether a generic summary could effectively present enough information to allow people to quickly and correctly categorize a document. The final task, a question-answering task, involved an intrinsic evaluation, where a topic-related summary for a document was evaluated in terms of its "informativeness".

Another important conference for text summarization was DUC, which was held every year from 2001 to 2007 (Gambhir and Gupta 2017). All editions of this conference contained newswire documents. Initially, in DUC-2001 and DUC-2002, the tasks involved generic summarization of single and multiple documents; they later extended to query-based summarization of multiple documents in DUC-2003. In DUC-2004, topic-based single and multi-document cross-lingual summaries were evaluated.

3.5 TSC: Our Challenge

Another evaluation program, NTCIR, formed a series of three Text Summarization Challenge (TSC) workshops—TSC1 in NTCIR-2 from 2000 to 2001, TSC2 in NTCIR-3 from 2001 to 2002, and TSC3 in NTCIR-4 from 2003 to 2004. These workshops incorporated summarization tasks for Japanese texts. The evaluation was done using both extrinsic and intrinsic evaluation methods.

3.5.1 TSC1

In TSC1, newspaper articles were used, and two tasks for a single article with intrinsic and extrinsic evaluations were performed (Fukushima and Okumura 2001; Nanba and Okumura 2002). We used newspaper articles from the Mainichi newspaper database of 1994, 1995, and 1998. The first task (Task A) was to produce summaries (extracts and free summaries) for intrinsic evaluation. We used recall, precision, and F-measure for evaluation of the extracts, and content-based as well as subjective methods for the evaluation of free summaries. The second task (Task B) was to produce summaries for the information retrieval task. The measures for evaluation were recall, precision, and F-measure for the correctness of the task, as well as the time that it takes to carry out the task. We also prepared human-produced summaries for the evaluation. In terms of genre, we used editorials and business news articles in the TSC1 dry-run evaluation, and editorials and articles on social issues in the formal run evaluation. As shareable data, we gathered summaries not only for the TSC evaluation but also for the researchers to share. By spring 2001, we collected summaries of 180 newspaper articles. For each article, we had the following seven types of summaries: important sentences (10, 30, 50%), summaries created by extracting important parts in sentences (20, 40%), and free summaries (20, 40%).

The basic evaluation design of TSC1 was similar to that of SUMMAC. The differences were as follows:

- As the intrinsic evaluation in Task A, we used a ranking method in subjective evaluation for four different summaries (baseline system results, system results, and two kinds of human summaries).
- Task B was basically the same as one of the SUMMAC extrinsic evaluations (the adhoc task), except the documents were in Japanese.

The following points were some of the features of TSC1. For Task A, we used several summarization rates and prepared the texts of various lengths and genres to use for evaluations. Their lengths varied at 600, 900, 1200, and 2400 characters, and the genres included business news, social issues, as well as editorials. As for Task A, because it was difficult to perform intrinsic evaluation on informative summaries, we presented the evaluation results as materials for discussions, at NTCIR workshop 2.

3.5.2 TSC2

TSC2 had two tasks (Okumura et al. 2003): single-document summarization (Task A) and multi-document summarization (Task B). In Task A, we asked the participants to produce summaries in plain text to be compared with human-prepared summaries from single texts. This task was the same as Task A in TSC1. In Task B, more than one (multiple) texts were summarized for the task. Given a set of texts, which has been manually gathered for a pre-defined topic, the participants produced summaries of the set in plain text format. The information that was used to produce the document set such as queries and summarization lengths were also given to the participants.

We used newspaper articles of the Mainichi newspaper database from 1998 and 1999. As the gold standard (human prepared summaries), we prepared the following types of summaries:

Extract-type summaries: We asked annotators, captioners who were well experienced in summarization, to select important sentences from each article.

Abstract-type summaries: We asked the annotators to summarize the original articles in two ways. First, to choose important parts of the sentences in extract-type summaries. Second, to summarize the original articles freely without worrying about sentence boundaries and trying to obtain the main ideas of the articles. Both types of abstract-type summaries were used for Task A. Both extract-type and abstract-type summaries were made from single articles.

Summaries from more than one article: Given a set of newspaper articles that has been selected based on a certain topic, the annotators produced free summaries (short and long summaries) for the set. Topics varied from a kidnapping case to the Y2K problem.

We used summaries prepared by humans for evaluation. The same two intrinsic evaluation methods were used for both tasks. They were evaluated by ranking the summaries and by measuring the degree of revisions.

Evaluation by ranking: This is basically the same method as the one we used for Task A in TSC1 (subjective evaluation). We asked human judges, who are experienced in producing summaries, to evaluate and rank the system summaries from two points of views:

1. Content: How much the system summary covers the important content of the original article?
2. Readability: How readable the system summary is?

Evaluation by revision: It was a newly introduced evaluation method in TSC2 to evaluate summaries by measuring the degree of revisions of the system results. The judges read the original texts and revised the system summaries in terms of content and readability. The revisions were made by only three editing operations (insertion, deletion, and replacement). The degree of the human revisions, which we call “edit distance”, was computed from the number of revised characters divided by the number of characters in the original summary. As a baseline for

Task A, human-produced summaries, as well as lead-method results, were used. Also, as a baseline for Task B, human-produced summaries, lead-method results, and the results based on the Stein method (Stein et al. 1999) were used. The lead-method extracts a few first sentences of news articles. The procedure of the Stein method is roughly as follows:

1. Produce a summary for each document.
2. Group the summaries into several clusters. The number of clusters is adjusted to be less than the half of the number of the documents.
3. Choose the most representative summary as the summary of the cluster.
4. Compute the similarity among the clusters and output the representative summaries in such order that the similarity of neighboring summaries is high.

We compared the evaluation by revision with the ranking evaluation, which is a manual method used in both TSC1 and TSC2. To investigate how well the evaluation measure recognizes slight differences in the quality of the summaries, we calculated the percentage of cases where the order of edit distance of two summaries matched the order of their ranks given by the ranking evaluation by checking the score from 0 to 1 at 0.1 intervals. As a result, we found that the evaluation by revision is effective for recognizing slight differences between computer-produced summaries (Nanba and Okumura 2004).

3.5.3 TSC3

In a single document, there are few sentences with the same content. In contrast, in multiple documents with multiple sources, there are many sentences that convey the same content with different words and phrases, or even with identical sentences. Thus, a text summarization system needs to recognize such redundant sentences and reduce this redundancy in the output summary.

However, we have no ways of measuring the effectiveness of such methods of reducing redundancy in the corpora for DUC and TSC2. The gold standard in TSC2 was given as abstracts (free summaries) with the number of characters less than a fixed number. It was therefore difficult to use for repeated or automatic evaluation and for the extraction of important sentences. Moreover, in DUC, where most of the gold standard was abstracts with the number of words less than a fixed number, the situation was the same as in TSC2. At DUC 2002, extracts (important sentences) were used, and this allowed us to evaluate sentence extraction. However, it was not possible to measure the effectiveness of redundant sentence reduction because the corpus was not annotated to show sentences with the same content.

Because many of the current summarization systems for multiple documents were based on sentence extraction, in TSC3, we assumed that the process of multiple

document summarization should consists of the following three steps. We produced a corpus for evaluating the system at each of these three steps[3] (Hirao et al. 2004).

Step 1 Extract important sentences from a given set of documents,
Step 2 Minimize redundant sentences from the results of Step 1,
Step 3 Rewrite the results of Step 2 to reduce the size of the summary to the specified number of characters or less.

We have annotated not only the important sentences in the document set, but also those among them that have the same content. These are the corpora for Steps 1 and 2. We have prepared human-produced free summaries (abstracts) for Step 3. We constructed extracts and abstracts of thirty sets of documents drawn from the Mainichi and Yomiuri newspapers published between 1998 to 1999, each of which was related to a certain topic.

In TSC3, because we had the gold standard (a set of correct important sentences) for Steps 1 and 2, we conducted automatic evaluation using a scoring program. We adopted *intrinsic* evaluation by human judges for Step 3. Therefore, we used the following *intrinsic* and *extrinsic* evaluation. The *intrinsic* metrics were "Precision", "Coverage", and "Weighted Coverage." The *extrinsic* metric was "Pseudo Question-Answering," i.e., whether a summary has an "answer" to the question or not. The evaluation was inspired by the question-answering task in SUMMAC. Please refer to Hirao et al. (2004) for more details of the intrinsic metrics.

3.6 Text Summarization Evaluation Campaigns After TSC

In DUC-2005 and DUC-2006, multi-document query-based summaries were evaluated whereas in DUC-2007, multi-document update query-based summaries were evaluated. These conferences also provided standard corpora of documents and gold summaries.

After 2007, DUC was succeeded by TAC, in which summarization tracks were presented (Gambhir and Gupta 2017). The 2008 summarization track consisted of two tasks: update task and opinion pilot. The update summarization task aimed to produce a short summary (around 100 words) from a collection of news articles, assuming that the user has already gone through a collection of previous articles. The opinion pilot task aimed to produce summaries of opinions from blogs. The 2009 summarization track had two tasks: update summarization, which was the same as in 2008, and Automatically Evaluating Summaries of Peers (AESOP). AESOP was a new task that was introduced in 2009; AESOP computes a summary's score with respect to a particular metric that is related to the summary's content, such as overall responsiveness and pyramid scores. The 2010 summarization track had two tasks: guided summarization and AESOP. The guided summarization task aimed

[3]This is based on a general idea of a summarization system and is not intended to impose any conditions on a summarization system.

to generate a 100-word summary from a collection of 10 news articles pertaining to a specific topic; each topic belongs to a previously defined category. The 2011 summarization track consisted of three tasks: guided summarization, AESOP, and multilingual pilot.

3.7 Future Perspectives

We described our TSC series, the data used, and the evaluation methods for each task, and the features of TSC evaluation. As we mentioned in Sect. 3.5, the contributions of our TSC can be considered as follows:

- We proposed a new evaluation method, evaluation by revision, that evaluates summaries by measuring the degree of revisions of the system results.
- We proposed a new evaluation method for multiple document summarization that enables us to measure the effectiveness of redundant sentence reduction in the systems.

More than 15 years have passed since our last evaluation challenge. Today, the text summarization field has changed a lot in that a huge amount of summarization data is now available in the field and neural models have prevailed and dominated the field. While we now have a variety of large summarization datasets such as Gigaword Corpus, New York Times Annotated Corpus, CNN/Daily Mail dataset, and NEWSROOM dataset (Grusky et al. 2018), it becomes difficult to compare systems on the datasets, against our expectations, because we do not necessarily have a standard dataset to compare them with. Even for the same dataset, the performance might change depending on differently sampled test data. Therefore, we can say that the current evaluation of summarization systems might not necessarily be reliable. In the future, we should construct a good standard dataset, against which we could compare summarization systems. For this purpose, it is necessary to investigate the properties of a variety of datasets that will enable us to sample test data to create a good evaluation dataset.

References

Allahyari M, Pouriyeh S, Assef M, Safaei S, Trippe ED, Gutierrez JB, Kochut K (2017) Text summarization techniques: a brief survey. Int J Adv Comput Sci Appl (IJACSA) 8(10)

Fukushima T, Okumura M (2001) Text summarization challenge: text summarization evaluation in Japan. In: Proceedings of NAACL workshop on automatic summarization, pp 51–59

Gambhir M, Gupta V (2017) Recent automatic text summarization techniques: a survey. Artif Intell Rev 47(1):1–66

Grusky M, Naaman M, Artzi Y (2018) Newsroom: a dataset of 1.3 million summaries with diverse extractive strategies. In: Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: human language technologies, vol. 1 (long papers), pp 708–719

Hirao T, Fukusima T, Okumura M, Nobata C, Nanba H (2004) Corpus and evaluation measures for multiple document summarization with multiple sources. In: Proceedings of the 20th international conference on computational linguistics, pp 535–541

Mani I, Maybury M (eds) (1999) Advances in automatic text summarization. MIT Press, Cambridge

Nanba H, Okumura M (2002) Some examinations of intrinsic methods for summary evaluation based on the text summarization challenge (TSC). In: Proceedings of the 3rd international conference on language resources and evaluation (LREC2002), pp 739–746

Nanba H, Okumura M (2004) Comparison of some automatic and manual methods for summary evaluation based on the text summarization challenge 2. In: Proceedings of the 4th international conference on language resources and evaluation, pp 1029–1032

Nazari N, Mahdavi MA (2019) A survey on automatic text summarization. J AI Data Min 7(1):121–135. https://doi.org/10.22044/JADM.2018.6139.1726

Okumura M, Fukushima T, Nanba H (2003) Text summarization challenge 2 text summarization evaluation at NTCIR workshop 3. In: Proceedings of the HLT/NAACL 2003 workshop on text summarization, pp 49–56

Sparck Jones K (1999) Automatic summarizing: factors and directions. In: Mani I, Maybury M (eds) Advances in automatic text summarization. MIT Press, Cambridge, pp 1–12. http://xxx.lanl.gov/ps/cmp-lg/9805011

Stein G, Strzalkowski T, Wise G (1999) Summarizing multiple documents using text extraction and interactive clustering. In: Proceedings of the PACLING'99, pp 200–208

Tombros A, Sanderson M (1998) Advantages of query biased summaries in information retrieval. In: Proceedings of the 21st annual international ACM-SIGIR conference on research and development in information retrieval, pp 2–10

Chapter 4
Challenges in Patent Information Retrieval

Makoto Iwayama, Atsushi Fujii, and Hidetsugu Nanba

Abstract We organized tasks on patent information retrieval during the decade from NTCIR-3 to NTCIR-8. All of the tasks were ones that reflected real needs of professional patent searchers and used large numbers of patent documents. This chapter describes the designs of the tasks, the details of the test collections, and the challenges addressed in the research field of patent information retrieval.

4.1 Introduction

A patent for an invention is a grant for the inventor to exclusively exploit the invention in the limited term in return to disclosing it to the public. The invention is described in a document called a patent application (also called a patent specification or an application document), which is composed of an abstract and sections describing the scope of the invention (the claims), the problems to be solved, the embodiments of the invention, etc. The patent application is filed with the patent office. The date of filing is called the filing date or application date. After the filing, the patent office examines the patent application, and if the invention is judged to be novel, in other words, one which has no prior art, a patent is granted for it.

As the economy grows worldwide, the number of patent applications and grants has also grown. The World Intellectual Property Organization (WIPO) announced

M. Iwayama (✉)
Hitachi Ltd. Research & Development Group, 1-280, Higashi-Koigakubo, Kokubunji-shi, Tokyo 185-8601, Japan
e-mail: makoto.iwayama.nw@hitachi.com

A. Fujii
School of Computing, Tokyo Institute of Technology, W8-62, 2-21-1 Ookayama, Meguro-ku, Tokyo 152-8552, Japan
e-mail: fujii@cs.titech.ac.jp

H. Nanba
Department of Industrial and Systems Engineering, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan
e-mail: nanba@kc.chuo-u.ac.jp

T. Sakai et al. (eds.), *Evaluating Information Retrieval and Access Tasks*,
The Information Retrieval Series 43,
https://doi.org/10.1007/978-981-15-5554-1_4

that the number of patent applications in 2017 had exceeded three million. In Japan, about three-hundred-thousand patent applications are filed every year. Since patent applications are highly technical and their length tends to be long, the task of searching for patent applications poses many issues in relation to information retrieval; similar situations are searching for technical papers or searching for legal documents.

In this chapter, we introduce the challenges aimed at addressing the issues of patent information retrieval.[1] These challenges were formulated as tasks performed in NTCIR workshops from 2001 to 2010. The NTCIR tasks were designed on the basis of actual patent-related work involving a large number of patent applications. The remainder of this chapter is organized as follows. Section 4.2 briefly introduces the NTCIR tasks. Section 4.3 describes the tasks in detail, including the search topics, document collections, submissions, relevance judgements, evaluation measures, and participants. Finally, Sect. 4.4 summarizes NTCIR's contributions to the research activities on patent information retrieval.

4.2 Overview of NTCIR Tasks

4.2.1 Technology Survey

Managers, researchers, and developers often want to know whether there are existing inventions related to the products they are planning to develop. This situation is similar to when researchers survey research papers before embarking on new research.

To satisfy this information need, they have to conduct a "technology survey" that involves searching for relevant patent applications published so far. Here the query might not be described in patent-specific terms, because the searcher is not always familiar with the procedure for searching for patents. Moreover, the notion of relevance is not patent-specific. Patent applications are treated like technological articles such as research papers.

4.2.2 Invalidity Search

After inventing a new method, device, material, etc., the inventor describes the invention in a patent application and sends it to the patent office. The patent office then examines the application to see if there is prior art which invalidates the invention by searching for patent applications filed before the filing date. This is called a "invalidity search" or "prior art search". Invalidity searches are also conducted by applicants themselves, because they should be confident of their inventions being granted patents before they make their applications.

[1] Readers who are interested in patent machine translation can refer to Chap. 7.

The invalidity search is patent-specific work. A searcher should be able to understand the components of an invention in accordance with the claims described in the application. Relevance is assessed based on the novelty or the invalidity of the invention. The searcher compares each component of the invention with portions of each retrieved document to see if they describe the same invention. If there is no prior art which can invalidate the novelty of the invention, a patent is granted; otherwise, the application is rejected. In most cases of rejection, several instances of prior art are cited, each of which corresponds to a component of the described invention.

4.2.3 Classification

Classification codes are extensively used when searching to narrow down the relevant applications. The patent office assigns each patent application appropriate classification codes before it is published. Human experts have to expend much effort to make this assignment, and for this reason, a (semi-)automatic method is desired.

The most popular classification codes for patents are the International Patent Classification (IPC) codes which are used worldwide. The Japan Patent Office (JPO) additionally uses and maintains a list of F-terms (File forming terms). F-terms are facet-oriented classification codes, and a patent application is classified from a variety of facets (viewpoints) such as objective, application, structure, purpose, and means.

In NTCIR, patent applications are automatically classified with F-terms in accordance with the behavior of human experts who perform their classification work in two steps. The first step is the theme (topic) classification, assigning a patent application to technological themes. Each theme corresponds to a group of IPC "sub-classes". The number of themes is about 2,500. The second step is F-term classification, i.e., assigning F-terms to an application that has already been assigned themes in the first step. Although the total number of F-terms is huge, over 300,000, the number of F-terms within each theme is relatively small, about 130 on average.

4.2.4 Mining

For a researcher in a field with high industrial relevance, analyzing research papers and patents has become an important aspect of assessing the scope of their field. The JPO creates patent application technical trend surveys for fields in which the development of technologies is expected, or fields to which social attention is being paid. However, it is costly and quite time-consuming.

In NTCIR, we aimed to construct a technical trend map from research papers and patents in a specific field. For the construction of the map, we focused on the elemental (underlying) technologies used in a particular field and their effects. Knowledge of the history and effects of the elemental technologies used in a particular field is

important for grasping the outline of technical trends in the field. Therefore, we designed the task to extract elemental technologies and their effects from research papers and patents.

4.3 Outline of the NTCIR Tasks

4.3.1 Technology Survey Task: NTCIR-3

Table 4.1 summarizes the test collections of the NTCIR tasks for patent retrieval.

A technology survey task was performed at only NTCIR-3 (Iwayama et al. 2003). Since this task was our first attempt to handle a practical number of patent applications, we designed the task to be as close as possible to the ad hoc retrieval tasks in TREC, except that the targeting documents were patent applications.

To launch the task, we obtained the cooperation of members of the Japan Intellectual Property Association (JIPA), who are experts in patent searches. Each JIPA member belongs to the intellectual property division in the company he or she works for. We collaborated with them in designing our tasks, constructing search topics, collecting relevant documents, evaluating the submitted results, and many other ways. Our collaboration continued through to NTCIR-4, and this was a major reason for the success of our challenges.

4.3.1.1 Search Topics

The technology survey task assumed a situation where a searcher is interested in a technology, for example, a "blue light-emitting diode", described in a newspaper

Table 4.1 NTCIR test collections for patent retrieval

Task		NTCIR-3	NTCIR-4		NTCIR-5	NTCIR-6	
		Technology survey	Invalidity search				
			Main	Additional		Main	English
Search topics		31(+6) ja,en,zh-CN,zh-TW	34(+7) ja,en,zh-CN	69 ja	1189 ja	1685 ja	2221(+1000) en
Document collections	Patent applications	1998–1999 ja	1993–1997 ja		1993–2002 ja		1993–2000 en
	Abstracts	1995–1999 ja,en	1993–1997 en		1993–2002 en		–
Relevance judgments		Manual		Citation			

article. The JIPA members constructed 31 search topics from newspaper articles which, in most cases, were selected from the topics they were working on in their daily jobs. Each search topic contained a title, headline, and text of the article that triggered the request for information, a description and a narrative of the topic, a set of concepts (keywords) related to the topic, and a supplement with more information about the relevance. All the search topics were translated into English, Korean, and Chinese (simplified and traditional).

4.3.1.2 Document Collections

The main documents for the retrieval were Japanese full texts of (unexamined) patent applications published in 1998 and 1999. The number of documents was 697,262. We also released abstracts of patent applications published over the 1995–1999 period, in Japanese and in English. The English abstracts were translations from the Japanese ones. The number of documents was 1,706,154 for the Japanese abstracts and 1,701,339 for the English abstracts. Some of the Japanese abstracts did not have corresponding English abstracts.

4.3.1.3 Submissions

Each participant submitted at least one run that used only the newspaper articles and supplements on the given search topics. In addition, we recommended that they submit ad hoc runs that used the descriptions and the narratives. For each search topic, a ranked list of at most 1000 patent applications was submitted in decreasing order of relevance score.

4.3.1.4 Relevance Judgments

The relevance of the technology survey is not patent-specific; that is, it is not based on the novelty of the invention, but rather on the relatedness of the search topic to the patent application.

The relevance was assessed by JIPA members in two steps. In the first step, the JIPA member who created the topic collected relevant documents on the topic before its release. Here although the members were allowed to use any search tools, almost all of them used Boolean ones, despite the fact that the organizers had provided a rank-based search system. In the second step, after the participants submitted their results for a topic, the JIPA member who created the topic judged the relevance of the unseen documents in the pool collected from the top-ranked submitted documents.

4.3.1.5 Evaluation

The submitted runs were evaluated by comparing recall/precision trade-off curves and values of mean average precision (MAP).

4.3.1.6 Participants

Eight groups submitted 36 runs. The top-performing run was from Ricoh (Itoh et al. 2003). They focused on re-weighting terms based on their statistics in the different collections (patent applications vs. newspaper articles). For example, the query term "president" in a newspaper article might not be effective for retrieving relevant patent applications. However the inverse document frequency (IDF)-based weighting gives this term a large weight, because it occurs rarely in patent applications. Their approach, called "term distillation", involved multiplying the weights in the query (i.e., newspaper articles) and the target documents (i.e., patent applications) to select effective terms from a query newspaper article.

4.3.2 Invalidity Search Task: NTCIR-4, NTCIR-5, and NTCIR-6

Having gained experience in the technology survey task at NTCIR-3, we moved on to invalidity search, which is a patent-specific search. Invalidity search tasks were performed in NTCIR-4 (Fujii et al. 2004), NTCIR-5 (Fujii et al. 2005), and NTCIR-6 (Fujii et al. 2007b).

4.3.2.1 Search Topics

Invalidity searches are searches, for instances, of prior art that could invalidate a patent application. Here, the patent application itself becomes a search topic. As search topics in NTCIR-4, JIPA members selected 34 Japanese patent applications that had been rejected by JPO. We called this set "NTCIR-4 main topics".

Regarding the NTCIR4 main topics, relevant documents were thoroughly collected by the JIPA members (see Sect. 4.3.2.4 for the details of this collection procedure). However, we found that the number of relevant documents in invalidity searches was small compared with the existing test collections for information retrieval. Consequently, evaluations made on a small number of topics could potentially be inaccurate. To increase the number of search topics, JIPA members selected an additional 69 search topics from other rejected patent applications. Here, relevant documents were only the citations reported by JPO. We called this set "NTCIR-4 additional topics". Note that the major difference between these two sets relates to

the completeness of relevant documents. We will discuss the effect of this issue on the evaluation in Sect. 4.3.2.6.

The NTCIR-4 main topics set had two more resources with which to make additional evaluations. First, each claim was translated into English and simplified Chinese for evaluating cross-language retrieval. Second, each Japanese claim had annotations for the components of the invention. Components were, for example, parts of a machine or substances of a chemical compound. Some participants used component information in the task.

In both NTCIR-5 and NTCIR-6, the organizers increased the number of search topics by following the same method used to create NTCIR-4 additional topics. Accordingly, the relevant documents for these topics were only citations. The number of search topics was 1,189 in NTCIR-5 and 1,685 in NTCIR-6. We called these topic sets "NTCIR-5 main topics" and "NTCIR-6 main topics".

NTCIR-5 included a passage retrieval task as a sub-task of the invalidity search task. Since patent applications are lengthy, it is useful to point out significant fragments ("passages") in a relevant application. In the passage retrieval task, a relevant application retrieved from a search topic was given, and the purpose was to identify the relevant passages in the relevant application. We used 378 relevant applications obtained from 34 search topics of NTCIR-4 main topics plus another 6 that had been used in the dry run in NTCIR-4.

NTCIR-6 involved an invalidity search task on English patent applications (called the "English retrieval task"). The design of the task was the same as the Japanese one. Each search topic was a patent application published by the United States Patent and Trademark Office (USPTO) in 2000 or 2001. We collected 3,221 search topics (1,000 for the dry run and 2,221 for the formal run) from those satisfying the two conditions; first, at least 20 citations are listed, and second, at least 90% of the citations are included in the target document collection. These citations were relevant documents.

4.3.2.2 Document Collections

In NTCIR-4, the document collection for the target of searching consisted of 5 years' worth of Japanese (unexamined) patent applications published from 1993 to 1997. The number of documents totaled 1.7 million. We additionally released English abstracts that were translations of the Japanese abstracts in these applications.

In NTCIR-5 and NTCIR-6, the document collections (both Japanese patent applications and English abstracts) were enlarged to include those published over the 10 year period from 1993 to 2002. The number of documents in each collection was 3.5 million.

In the English retrieval task at NTCIR-6, the document collection was patent applications published from USPTO over the period from 1993 to 2000. The number of documents totaled about 1 million.

4.3.2.3 Submissions

Although the full texts of the patent applications were provided as search topics, each participant was requested to submit a result which only used the claims and the filing dates in the search topics. Participants could submit additional results, in which they could use any information in the search topics. The number of documents retrieved for each search topic was 1,000 at maximum, and these documents were submitted in decreasing order of relevance score.

To assess effectiveness across different sets of search topics, each participant was requested to submit a set of results from all the Japanese main topics released so far. For example, in NTCIR-6, each participant had to submit runs using NTCIR-4 main topics and NTCIR-5 main topics in addition to NTCIR-6 main topics.

In the passage retrieval task in NTCIR-5, each participant was requested to sort all passages in each of the given relevant applications according to the degree to which a passage provided grounds to judge if the application was relevant.

4.3.2.4 Relevance Judgments

In invalidity searches, the most reliable relevant documents are ones cited by the patent office when rejecting patent applications. However, we were not confident that using only citations would be enough to evaluate the participating systems from the standpoint of recall. Therefore, in NTCIR-4, we exhaustively collected relevant documents by performing the same two steps that were used in the technology survey task of NTCIR-3.

First, the JIPA members who created the search topics (NTCIR-4 main topics) performed manual searches to collect as many relevant documents as possible. Citations from the topic applications were included among the relevant documents. We allowed the JIPA members to use any system or resource to find relevant documents. In this way, we would obtain a relevant document set under the circumstances of their daily patent searches. Most members used Boolean searches, which to this day remains the most popular method used in invalidity searches. Second, after the participants submitted their runs, the JIPA members judged the relevance for the unseen documents in the pool collected from the top-ranked documents in each run. Here, one promising result was that the participating systems could find a relatively large number of relevant documents which were neither citations nor relevant documents found by the JIPA members in the first step.

In NTCIR-5 and NTCIR-6, we used only citations as relevant documents, mainly because we could not cooperate with expert searchers.

Relevance was automatically graded as relevant, partially relevant, or irrelevant. A document that could solely be used to reject an application was regarded as relevant. A document that could be used with other documents to reject an application was regarded as partially relevant. Other documents were regarded as irrelevant.

In NTCIR-6, we tried an alternative definition of the relevance grade, one based on the observation that if a search topic and its relevant application have the same IPC

codes, systems could easily retrieve the relevant application by using IPC codes as filters. We divided the relevant documents into three classes according to the number of shared IPC codes between the topic and a relevant document and compared the submitted runs on the basis of the classes (Fujii et al. 2007b).

In the passage retrieval task of NTCIR-5, we reused the search topics in NTCIR-4 and all the relevant passages had been collected in NTCIR-4 by JIPA members. Relevance was graded as follows. If a single passage could be grounds to judge the target document as relevant or partially relevant, this passage was judged to be relevant. If a group of passages could be grounds, each passage in the group was judged as partially relevant. Otherwise, the passage was judged as irrelevant.

4.3.2.5 Evaluation

We used MAP for the evaluation measure in all the invalidity search tasks. In the passage retrieval task, we additionally used the averaged passage rank at which an assessor obtains sufficient grounds to judge whether a target document is relevant or partially relevant, when the assessor checked the passages in the top-ranked to bottom-ranked target documents.

4.3.2.6 Participants

Eight groups participated in the invalidity search tasks of NTCIR-4, ten in NTCIR-5, and five in NTCIR-6. The passage retrieval task of NTCIR-5 had four groups, while the English retrieval task of NTCIR-6 had five groups. In this section, we introduce only those groups who participated in most of the main tasks, i.e., document-level invalidity searches using Japanese topics.

Hitachi submitted runs to all of the invalidity search tasks (Mase et al. 2004, 2005; Mase and Iwayama 2007). From NTCIR-4 to NTCIR-6, they tried various methods, for example, using stop words, filtering by IPC codes, term re-weighting, or using the claim's structure. The methods were composed of two-step searches. The first step was a recall-oriented search, and the second step was a re-ranking of the documents retrieved by the first step to improve precision.

NTT Data participated in NTCIR-4 (Konishi et al. 2004) and NTCIR-5 (Konishi 2005). They expanded the query terms with keywords selected from the "detailed descriptions of the invention" ("embodiments") section. First, they decomposed a topic claim into components of the invention by using pattern-matching rules. Next, they identified descriptions that explain each component by using another set of pattern-matching rules. Lastly, they added keywords in the descriptions to the query terms.

The University of Tsukuba participated in all of the invalidity search tasks (Fujii and Ishikawa 2004, 2005; Fujii 2007). They automatically decomposed a topic claim into components and searched the components independently. Then, they integrated the results. Query terms were also extracted from related passages automatically

identified in the topic document. Retrieved documents which did not share the IPC codes of the query application were filtered out. They observed that the IPC filtering was more effective in NTCIR-5 main topics than in NTCIR-4 main topics (Fujii and Ishikawa 2005; Fujii 2007). This difference might have been due to the nature of the relevance of the two sets. The relevant documents for NTCIR-4 main topics were manually collected, while those for NTCIR-5 main topics were only citations. If we imagine that searches at a patent office often rely on metadata (IPC codes), we could further assume that citations by the patent office might be retrieved by the IPC filtering. This hypothesis became a motivation for NTCIR-6 to divide up the relevant documents according to the number of shared IPC codes with a search topic (Fujii et al. 2007b).

Ricoh used the IPC codes for both filtering and pseudo-relevance feedback (Itoh 2004, 2005). In the latter usage, they first retrieved documents and extracted IPC codes from the top-ranked documents; then, they filtered out the retrieved documents which did not share any of the extracted IPC codes.

4.3.3 Patent Classification Task: NTCIR-5, NTCIR-6

4.3.3.1 Data Collections

Patent classification tasks were performed in NTCIR-5 (Iwayama et al. 2005) and NTCIR-6 (Fujii et al. 2007b). Table 4.2 summarizes the test collections of the NTCIR patent classification tasks.

The training documents in NTCIR-5 and NTCIR-6 consisted of Japanese (unexamined) patent applications published during 1993–1997 and their English abstracts. Themes and F-terms for these documents were also released. As for test documents in NTCIR-5, 2,008 documents were released for the theme classification task, while 500 were released for the F-term classification task. The documents were selected from Japanese (unexamined) patent applications published in 1998 and 1999. Five themes were selected in the F-term classification task.

In NTCIR-6, only the F-term classification task was performed. We increased the number of themes to 108, and the test documents to 21,606.

Table 4.2 NTCIR test collections for patent classification

Task		NTCIR-5		NTCIR-6
		Theme	F-term	F-term
Test documents		2,008	2,562 (5 themes)	21,606 (108 themes)
Training documents	Patent applications	1993–2002 ja		
	Abstracts	1993–2002 en		

4.3.3.2 Submissions

Each participant in NTCIR-5 submitted a ranked list of themes (at maximum 100) for each test document in the theme classification task and a ranked list of F-terms (at maximum 200) for each test document in the F-term classification task. Note that the participants were given the themes of each test document in the F-term classification task.

4.3.3.3 Evaluation

MAP and F-measure were used in the evaluation. To calculate the F-measure, participants were requested to submit a confident set of themes or F-terms for each test document.

4.3.3.4 Participants

Four groups submitted results to the theme classification task in NTCIR-4. Theme classification is similar to classifying patent applications into IPC sub-classes; k-Nearest Neighbor (k-NN) and naïve Bayes classifiers were popular methods, and the participants used these methods in the task (Kim et al. 2005; Tashiro et al. 2005).

Three groups participated in the F-term classification task in NTCIR-5 and six in NTCIR-6. Some groups used support vector machine (SVM) (Tashiro et al. 2005; Li et al. 2007) in addition to k-NN (Murata et al. 2005) and naïve Bayes (Fujino and Isozaki 2007) classifiers. The results suggested that feature selection had a greater influence on classification effectiveness than the choice of classifier. Since patent applications have several components including abstract, claim, technological field, purpose, embodiments, etc., we have many options for which components should be used as the source of features.

4.3.4 Patent Mining Task: NTCIR-7, NTCIR-8

Patent mining tasks were performed in NTCIR-7 (Nanba et al. 2008) and NTCIR-8 (Nanba et al. 2010). Table 4.3 summarizes the test collections of the patent mining tasks.

The purpose of the patent mining task was to create technical trend maps from a set of research papers and patents. Table 4.4 shows an example of a technical trend map. In this map, research papers and patents are classified in terms of elemental technologies and their effects.

Table 4.3 NTCIR test collections for patent mining

Task		NTCIR-7	NTCIR-8	
		Clasification	Clasification	Map creation
Test document		879	549	200
Training documents	Patent applications	1993–2002 ja, en		
	Abstracts of research papers	1988–1999 ja, en		

Table 4.4 Example of a technical trend map created from a set of research papers and patents

	Effect 1	Effect 2	Effect 3
Technology 1	[AA 1993] [US Pat. XX/XXX]		[BB 2002]
Technology 2	[CC 2000]		
Technology 3		[US Pat. YY/YYYY]	[US Pat. ZZ/ZZZZ] [JP Pat. WW/WWWW]

Two steps were used to create a technical trend map:

- (Step 1) For a given field, collect research papers and patents written in various languages.
- (Step 2) Extract elemental technologies and their effects from the documents collected in Step 1 and classify the documents in terms of the elemental technologies and their effects. Example of elemental technologies and their effects will be shown in Sect. 3.4.4.

Two subtasks were conducted in each step:

- Classify research paper abstracts.
- Create a technical trend map.

We describe the details of these subtasks below.

4.3.4.1 Research Paper Classification Subtask

The goal of this subtask was to classify research paper abstracts in accordance with the IPC system, which is a standard hierarchical patent classification system used around the world. One or more IPC codes are manually assigned to each patent, aiming for effective patent retrieval.

Task
This task involved assigning one or more IPC codes at the subclass, main group, and subgroup levels to a given topic, expressed in terms of the title and abstract of a research paper.

The following tasks were conducted.

- Japanese: classification of Japanese research papers using patent data written in Japanese.
- English: classification of English research papers using patent data written in English.

Data Collection
We created English and Japanese topics (titles and abstracts) and their correct classifications (IPC codes extracted from patents). On average, 1.6, 1.9, and 2.4 IPC codes were assigned at the subclass, main group, and subgroup levels, respectively, to each topic. In NTCIR-7, we randomly assigned 97 topics to the dry run and the remaining 879 topics to the formal run. In NTCIR-8, we assigned 95 topics to the dry run and the remaining 549 topics to the formal run. The dry run data were provided to the participating teams as training data for the formal run. Patents with IPC codes were also provided as additional training data.

Submission
Participating teams were asked to submit one or more runs, each of which contained ranked lists of IPC codes for each topic.

Evaluation
MAP, recall, and precision were used in the evaluation.

Participants
In NTCIR-7, we had 24 participating systems for the Japanese subtask, 20 for the English subtask, and five for the cross-lingual subtask. As far as the number of groups is concerned, we had 12 participating groups from universities and companies. In NTCIR-8, there were 71 participating systems for the Japanese subtask, 24 for the English subtask, and nine for the cross-lingual subtask. There were six participating groups.

Most participating teams employed the k-Nearest Neighbor (k-NN) method, which is a comparatively easy way of dealing with a large number of categories, because the classification is based only on extracting similar examples, with no training process being required. Furthermore, the k-NN method is itself a ranking, which enables it to be applied directly to the IPC code ranking. In NTCIR-7, Xiao et al. (2008) used the k-NN framework and various similarity calculation methods and re-ranking methods were examined.

4.3.4.2 Technical Trend Map Creation Subtask

Task
This task was conducted in NTCIR-8. The goal of this subtask was to extract expressions of elemental technologies and their effects from research papers and patents. We defined the tag set for this subtask as follows:

- TECHNOLOGY including algorithms, tools, materials, and data used in each study or invention.
- EFFECT including pairs of ATTRIBUTE and VALUE tags.
- ATTRIBUTE and VALUE including effects of a technology that can be expressed by a pair comprising an attribute and a value.

For example, suppose that the sentence "Through closed-loop feedback control, the system could minimize the power loss." is given to a system. In this case, the system was expected to output the following tagged sentence: "Through <TECHNOLOGY>closed-loop feedback control</TECHNOLOGY>, the system could <EFFECT><VALUE>minimize</VALUE> the <ATTRIBUTE>power loss</ATTRIBUTE></EFFECT>."

The following tasks were conducted:

- Japanese: extraction of technologies and their effects from research papers and patents written in Japanese.
- English: extraction of technologies and their effects from research papers and patents written in English.

Data Collection
Sets of topics with manually assigned TECHNOLOGY, EFFECT, ATTRIBUTE, and VALUE tags were used for the training and evaluation. Here, we asked a human subject to assign these tags to the following four text types:

- Japanese research papers (500 abstracts)
- Japanese patents (500 abstracts)
- English research papers (500 abstracts)
- English patents (500 abstracts)

Then, for each text type, We randomly selected 50 texts for the dry run and 200 texts for the formal run. We provided the remaining 250 texts to the participating teams as training data.

Submission
The teams were asked to submit texts with automatically annotated tags.

Evaluation
Recall, precision, and F-measure were used in the evaluation.

Participants
In NTCIR-8, there were 27 participating systems for the Japanese subtask and 13 for the English subtask. There were nine participating teams of universities and companies. For example, Nishiyama et al. (2010) used a system that applied a domain-adaptation method on both research papers and patents and confirmed its effectiveness.

4.4 Contributions

This section chronologically summarizes NTCIR's contributions to research activities on patent information retrieval. Figure 4.1 shows an overview.

4.4.1 Preliminary Workshop

In 2000, the Workshop on Patent Retrieval was co-located with the ACM SIGIR Conference on Research and Development in Information Retrieval (Kando and Leong 2000). This was the first opportunity for researchers and practitioners associated with patent retrieval to exchange knowledge and experience. The outcome of this workshop motivated researchers to foster research and development in patent retrieval by developing large test collections.

4.4.2 Technology Survey

Following the workshop in 2000, NTCIR-3 was organized as the first evaluation workshop focusing on patent information retrieval (2001–2002). The task was a technology survey.

Since patent offices publish patent applications in public, information retrieval, and natural language processing researchers can use them as a resource. The test collection constructed for NTCIR-3 was unique in that it contained not only patent applications but also search topics and their relevant documents; these were created

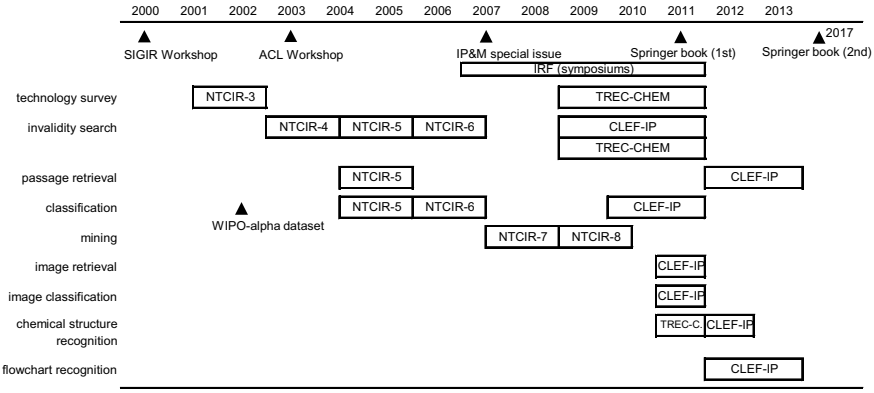


Fig. 4.1 History of research activities on patent information processing

and assessed by human experts. It was the first test collection for patent information retrieval with a large number of documents.

Here we should note that other workshops included technology survey tasks similar to the one performed in NTCIR-3, including the TREC-CHEM tracks in 2009 (Lupu et al. 2009), 2010 (Lupu et al. 2010) and 2011 (Lupu et al. 2011a). These tasks focused on research and development in the chemical domain, in which patent information plays important role.

4.4.3 Collaboration with Patent Experts

The organizers of NTCIR-3 and NTCIR-4 collaborated with patent experts, who were JIPA members, in constructing the test collections. The JIPA members created the search topics and they also collected and assessed relevant documents. The organizers and the JIPA members met once a month to discuss the task design. The participants and the JIPA members also shared knowledge and experiences at round-table meetings and tutorials. These activities helped to build bridges between information retrieval researchers and patent searchers.

4.4.4 Invalidity Search

NTCIR-4 (2003–2004) was the first workshop to include an invalidity search. Invalidity search is truly patent-specific work, and the organizers carefully designed the task with JIPA members.

In NTCIR-4, we examined the issue of whether it was possible to use only citations as relevant documents when evaluating submitted runs. While the NTCIR-4 collection included an exhaustive collection of relevant documents, the NTCIR-5 and NTCIR-6 collections had only citations. Moreover, since the topics and the documents were the same in the NTCIR4, NTCIR-5, and NTCIR-6 collections, researchers can compare their retrieval methods under the different ways of identifying relevant documents.

Invalidity search tasks were continuously organized in CLEF-IP in 2009 (Roda et al. 2009), 2010 (Piroi and Tait 2010) and 2011 (Piroi et al. 2011), and TREC-CHEM in 2009 (Lupu et al. 2009), 2010 (Lupu et al. 2010) and 2011 (Lupu et al. 2011a), under the name “prior art search task”.

The passage retrieval task in NTCIR-5 (2004–2005) was the first attempt to evaluate handling of passages in invalidity search. A passage retrieval task was revisited in CLEF-IP in 2012 (Piroi et al. 2012) and 2013 (Piroi et al. 2013) in a more challenging setting. In NTCIR, a relevant document to a search topic was given and the purpose was to find relevant passages in the given relevant document. On the other hand, in CLEF-IP, relevant passages were directly retrieved based on the claims in the search topic.

4.4.5 Patent Classification

The WIPO-alpha collection, released in 2002, was the first test collection for patent classification. It consisted of 75,250 English patent documents labeled with IPC codes. Many research papers on patent classification have used WIPO-alpha (Fall et al. 2003).

The NTCIR collections released by the classification tasks (2004–2007) were not for IPC, but for the classification codes used in the JPO, i.e., F-terms. F-terms re-classify a specific technical field of IPC from a variety of viewpoints, such as purpose, means, function, and effect.

The CLEF-IP classification tasks in 2010 (Piroi and Tait 2010) and 2011 (Piroi et al. 2011) released test collections on IPC codes; these were larger than the WIPO-alpha collection. The released documents totaled 2.6 million in 2010 and 3.5 million in 2011.

4.4.6 Mining

Patent mining tasks were performed in NTCIR-7 and NTCIR-8, and similar tasks were conducted in subsequent research (Gupta and Manning 2011; Tateisi et al. 2016). Gupta and Manning (2011) proposed a method to assign FOCUS (an article's main contribution), DOMAIN (an article's application domain), and TECHNIQUE (a method or a tool used in an article) tags to abstracts in the ACL Anthology[2] for the purpose of identifying technical trends. Tateisi et al. (2016) constructed a corpus for analyzing the semantic structures of research articles in the computer science domain.

Since February 2019, JDream III[3] has provided a new service for retrieving research papers using IPC codes. This service assigns IPC codes at the main group level to each research paper by using Nanba's method (Nanba 2008), which is based on the k-NN method.

4.4.7 Workshops and Publications

The organizers of the NTCIR tasks organized the ACL Workshop on Patent Corpus Processing in 2003 and edited a special issue on patent processing in Information Processing & Management in 2007 (Fujii et al. 2007a).

The Information Retrieval Facility (IRF), which was a not-for-profit research institution based in Vienna, Austria, organized a series of symposia between 2007 and 2011 to explore reasons for the knowledge gap between information retrieval

[2] https://www.aclweb.org/anthology/.

[3] https://jdream3.com/.

researchers and patent search specialists. The symposia were followed by publication of two editions of a book in 2011 (Lupu et al. 2011b) and 2017 (Lupu et al. 2017) introducing studies by information retrieval researchers and patent experts.

These activities contributed to the research trends in the communities of information retrieval and natural language processing.

4.4.8 CLEF-IP and TREC-CHEM

The NTCIR project ended with NTCIR-8 (2009–2010), and it left behind several unaddressed issues. Firstly, while NTCIR-3 and NTCIR-4 released multi-lingual search topics in English, Korean and Chinese, as well as English abstracts over the course of ten years and NTCIR-6 included an English retrieval task using patent applications published by USPTO, the workshops focused on Japanese and the multi-lingual resources were not widely used. This meant there were no serious evaluations of multi-lingual or cross-lingual patent retrieval. Secondly, the tasks ignored images, formulas, and chemical structures, despite the fact that these are important pieces of information for judging relevance in some domains.

The above issues that the NTCIR project did not address were investigated in CLEF-IP (2009–2013) and TREC-CHEM (2009–2011). Both were annual evaluation workshops (campaigns) on patent information retrieval. CLEF-IP had tasks for prior art search, passage retrieval, and patent classification. The tasks were similar to the NTCIR tasks, but most resources were from the European Patent Office, covering English, French, and German; hence, the CLEF-IP tasks were inherently multi/cross-lingual. In addition, CLEF-IP performed completely new tasks, including ones on image-based retrieval (Piroi et al. 2011), image classification (Piroi et al. 2011), flowchart/structure recognition (Piroi et al. 2012, 2013), and chemical structure recognition (Piroi et al. 2012). TREC-CHEM also had tasks for prior art search and technology survey. The TREC-CHEM tasks were challenging and focused on the chemical domain, which has many formulae and images. The image-to-structure task (Lupu et al. 2011a) in TREC-CHEM was the first one to include chemical structure recognition. TREC-CHEM used resources from USPTO.

References

Fall C, Torcsvari A, Benzineb K, Karetka G (2003) Automated categorization in the international patent classification. ACM SIGIR Forum 37(1):10–25

Fujii A (2007) Integrating content and citation information for the NTCIR-6 patent retrieval task. In: Proceedings of the 6th NTCIR workshop meeting on evaluation of information access technologies: information retrieval, question answering and cross-lingual information access

Fujii A, Ishikawa T (2004) Document structure analysis in associative patent retrieval. In: Proceedings of the 4th NTCIR workshop on research in information access technologies information retrieval, question answering and summarization

Fujii A, Ishikawa T (2005) Document structure analysis for the NTCIR-5 patent retrieval task. In: Proceedings of the 5th NTCIR workshop meeting on evaluation of information access technologies: information retrieval, question answering and cross-lingual information access

Fujii A, Iwayama M, Kando N (2004) Overview of patent retrieval task at NTCIR-4. In: Proceedings of the 5th NTCIR workshop on research in information access technologies information retrieval, question answering and summarization

Fujii A, Iwayama M, Kando N (2005) Overview of patent retrieval task at NTCIR-5. In: Proceedings of the 5th NTCIR workshop meeting on evaluation of information access technologies: information retrieval, question answering and cross-lingual information access

Fujii A, Iwayama M, Kando N (2007a) Introduction to the special issue on patent processing. Inf Process Manag 45(3):1149–1153

Fujii A, Iwayama M, Kando N (2007b) Overview of the patent retrieval task at the NTCIR-6 workshop. In: Proceedings of the 6th NTCIR workshop meeting on evaluation of information access technologies: information retrieval, question answering and cross-lingual information access

Fujino A, Isozaki H (2007) Multi-label patent classification at NTT communication science laboratories. In: Proceedings of the 6th NTCIR workshop meeting on evaluation of information access technologies: information retrieval, question answering and cross-lingual information access

Gupta S, Manning GD (2011) Analyzing the dynamics of research by extracting key aspects of scientific papers. In: Proceedings of the 5th international joint conference on natural language processing

Itoh H (2004) NTCIR-4 patent retrieval experiments at RICOH. In: Proceedings of the 4th NTCIR workshop on research in information access technologies information retrieval, question answering and summarization

Itoh H (2005) NTCIR-5 patent retrieval experiments at RICOH. In: Proceedings of the 5th NTCIR workshop meeting on evaluation of information access technologies: information retrieval, question answering and cross-lingual information access

Itoh H, Mano H, Ogawa Y (2003) Term distillation for cross-DB retrieval. In: Proceedings of the 3rd NTCIR workshop on research in information retrieval, automatic text summarization and question answering

Iwayama M, Fujii A, Fujii A, Takano A (2003) Overview of patent retrieval task at NTCIR-3. In: Proceedings of the 3rd NTCIR workshop on research in information retrieval, automatic text summarization and question answering

Iwayama M, Fujii A, Kando N (2005) Overview of classification subtask at NTCIR-5 patent retrieval task. In: Proceedings of the 5th NTCIR workshop meeting on evaluation of information access technologies: information retrieval, question answering and cross-lingual information access

Kando N, Leong M (2000) Workshop on patent retrieval SIGIR 2000 workshop report. ACM SIGIR Forum 34(1):28–30

Kim JH, Huang JX, Jung HY, Choi KS (2005) Patent document retrieval and classification at KAIST. In: Proceedings of the 5th NTCIR workshop meeting on evaluation of information access technologies: information retrieval, question answering and cross-lingual information access

Konishi K (2005) Query terms extraction from patent document for invalidity search. In: Proceedings of the 5th NTCIR workshop meeting on evaluation of information access technologies: information retrieval, question answering and cross-lingual information access

Konishi K, Kitauchi A, Takaki T (2004) Invalidity patent search system of NTT DATA. In: Proceedings of the 5th NTCIR workshop on research in information access technologies information retrieval, question answering and summarization

Li Y, Bontcheva K, Cunningham H (2007) SVM based learning system for F-term patent classification. In: Proceedings of the 6th NTCIR workshop meeting on evaluation of information access technologies: information retrieval, question answering and cross-lingual information access

Lupu M, Piroi F, Huang X, Zhu J, Tait J (2009) Overview of the TREC 2009 chemical IR track. In: Proceedings of TREC 2009

Lupu M, Tait J, Huang J, Zhu J (2010) TREC-CHEM 2010: Notebook report. In: Proceedings of TREC 2010

Lupu M, Gurulingappa H, Filippov I, Jiashu Z, Fluck J, Zimmermann M, Huang J, Tait J (2011a) Overview of the TREC 2011 chemical IR track. In: Proceedings of TREC 2011

Lupu M, Mayer K, Tait J, Trippe AJ (eds) (2011b) Current challenges in patent information retrieval. Springer, Berlin

Lupu M, Mayer K, Kando N, Trippe AJ (eds) (2017) Current challenges in patent information retrieval,2nd edn. Springer, Berlin

Mase H, Iwayama M (2007) NTCIR-6 patent retrieval experiments at hitachi. In: Proceedings of the 6th NTCIR workshop meeting on evaluation of information access technologies: information retrieval, question answering and cross-lingual information access

Mase H, Matsubayashi T, Ogawa Y, Iwayama M, Oshio T (2004) Two-stage patent retrieval method considering claim structure. In: Proceedings of the 4th NTCIR workshop on research in information access technologies information retrieval, question answering and summarization

Mase H, Matsubayashi T, Ogawa Y, Yayoi T, Sato Y, Iwayama M (2005) NTCIR-5 patent retrieval experiments at hitachi. In: Proceedings of the 5th NTCIR workshop meeting on evaluation of information access technologies: information retrieval, question answering and cross-lingual information access

Murata M, Kanamaru T, Shirado T, Isahara H (2005) Using the k nearest neighbor method and BM25 in the patent document categorization subtask at NTCIR-5. In: Proceedings of the 5th NTCIR workshop meeting on evaluation of information access technologies: information retrieval, question answering and cross-lingual information access

Nanba H (2008) Hiroshima City University at NTCIR-7 patent mining task. In: Proceedings of the 7th NTCIR workshop meeting on evaluation of information access technologies: information retrieval, question answering and cross-lingual information access

Nanba H, Fujii A, Iwayama M, Hashimoto T (2008) Overview of the patent mining task at the NTCIR-7 workshop. In: Proceedings of the 7th NTCIR workshop meeting on evaluation of information access technologies: information retrieval, question answering and cross-lingual information access

Nanba H, Fujii A, Iwayama M, Hashimoto T (2010) Overview of the patent mining task at the NTCIR-8 workshop. In: Proceedings of the 8th NTCIR workshop meeting on evaluation of information access technologies: information retrieval, question answering and cross-lingual information access

Nishiyama R, Tsuboi Y, Unno Y, Takeuchi H (2010) Feature-rich information extraction for the technical trend map creation. In: Proceedings of the 8th NTCIR workshop meeting on evaluation of information access technologies: information retrieval, question answering and cross-lingual information access

Piroi F, Tait J (2010) CLEF-IP 2010: retrieval experiments in the intellectual property domain. In: Working notes for CLEF-2010 conference

Piroi F, Lupu M, Hanbury A, Zenz V (2011) CLEF-IP 2011: retrieval experiments in the intellectual property domain. In: Working notes for CLEF-2011 conference

Piroi F, Lupu M, Hanbury A, Sexton A, Magdy W, Filippov I (2012) CLEF-IP 2012: retrieval experiments in the intellectual property domain. In: Working notes for CLEF-2012 conference

Piroi F, Lupu M, Hanbury A (2013) Overview of CLEF-IP 2013 lab. In: CLEF 2013 Proceedings of the 4th international conference on information access evaluation. Multilinguality, multimodality, and visualization

Roda G, Tait J, Piroi F, Zenz V (2009) CLEF-IP 2009: Retrieval experiments in the intellectual property domain. In: Multilingual information access evaluation I. Text retrieval experiments 10th workshop of the cross-language evaluation forum, CLEF 2009

Tashiro T, Rikitoku M, Nakagawa T (2005) Justsystem at NTCIR-5 patent classification. In: Proceedings of the 5th NTCIR workshop meeting on evaluation of information access technologies: information retrieval, question answering and cross-lingual information access

Tateisi Y, Ohta T, Miyao Y, Pyysalo S, Aizawa A (2016) Typed entity and relation annotation on computer science papers. In: Proceedings of the 10th international conference on language resources and evaluation (LREC 2016)

Xiao T, Cao F, Li T, Song G, Zhou K, Zhu J, Wang H (2008) Knn and re-ranking models for english patent mining at NTCIR-7. In: Proceedings of the 7th NTCIR workshop meeting on evaluation of information access technologies: information retrieval, question answering and cross-lingual information access

Chapter 5
Multi-modal Summarization

Tsuneaki Kato

Abstract Multi-modal summarization is a technology that provides users with abridgments of topics of interest. Such abridgments consist of organized text and informative graphics. These summarizations have two roles. One is to assist the users to review and understand their topics of interest. The other is to guide users both visually and verbally in their exploratory search. To establish this technology, it was necessary to integrate several research streams. These included information access, information extraction, and information visualization; all of these technologies had been developing rapidly since the beginning of the twenty-first century. MuST was a workshop, the main theme of which was research on multi-modal summarization of trend information. It was not an evaluation workshop and did not present the participants with a specific task, because at the time when the workshop was conducted, multi-modal summarization was merely an agglomeration of yet-to-be-developed technologies that had not yet been fully synthesized. Rather than sharing a task, the MuST workshop shared a data set. Making an annotated corpus shared as its unifying force, the workshop encouraged cooperative and competitive researches on trend information. Several innovations emerged from the workshop. These covered trend information extraction, visualization as information access interface and as data analysis method, linguistic summary generation from charts, and trend mining.

5.1 Background

By the beginning of the twenty-first century, information access technologies had changed and diversified. What was being accessed had changed from entire documents to passages within documents, and thence to the information itself. Question answering, the motto of which was to return information itself rather than pages or documents, had already progressed to managing simple factoid questions, and was

T. Kato (✉)
The University of Tokyo, Meguro-ku Komaba 3-8-1, Tokyo 153-8902, Japan
e-mail: kato@boz.c.u-tokyo.ac.jp

T. Sakai et al. (eds.), *Evaluating Information Retrieval and Access Tasks*,
The Information Retrieval Series 43,
https://doi.org/10.1007/978-981-15-5554-1_5

expected to reply to increasingly complicated queries such as those that included causes and definitions.

Access methods had also changed. Exploratory and interactive search was being emphasized. Information gathering was no longer a one-shot interaction through which users described their interest precisely and in return obtained adequate relevant feedback; instead, the process had become continuous, wherein users browsed information that was gathered according to general descriptions and then identified aspects regarding which they need more detailed information. Through this process, users interactively accumulated information while simultaneously expanding their area of exploration.

Methods for displaying the information so obtained had also advanced from simple ranked lists to information visualization. Some visualization techniques helped users to represent their information requests visually, others helped them to interactively analyze and interpret the results. Such information visualization techniques for information access were new and had different characteristics from those for scientific visualization.

Information was no longer simply collected or retrieved. Advances now allowed it to be compiled and synthesized using information extraction and multi-document summarization, which were techniques that had matured during that period.

Some of the research fields, such as exploratory search and information visualization, that adopted such changes in that era closely interacted with each other. This was, however, not the case for many other fields. Although one could find a limited implementation of some aspects (Ahmad et al. 2004), at that time, it was not envisioned that anything similar to the recent disaster informatics system would arise; this system synthetically processes both numeric data and linguistic data, such as documents, and summarizes and visualizes that data according to the users' requirements. There was, however, an expectation that interactions among, and fusions of, those research fields would bring about a number of fundamental innovations.

5.2 Applications Envisioned

These anticipated fusions could take many forms. One form could lead to a sophisticated question-answering system for responding to queries such as "How have oil and gasoline prices changed this year?" or "How bad were the typhoons last year?" The system would achieve this by compiling text and statistical data and then generating combinations of succinct text and information graphics. More advanced applications of such systems may include patent or research-map generation, which would show and explain the trends of patent applications or the publication of scholarly papers. These potential developments were subsequently pursued in another NTCIR workshop, which is briefly mentioned in Sect. 5.4.

This mechanism, which we termed *multi-modal summarization*, can be regarded as an effort to expand text summarization. While text summarization extracts important content from a body of real-world text and presents it in a condensed form, multi-

modal summarization also processes non-linguistic information such as numerical data and information graphics. Whereas multimedia presentation generation (Fasciano and Lapalme 1996; Roth and Mattis 1990 for example), which had been actively studied at the end of the last century, aimed to generate multimedia presentations from media-independent semantic representations; multi-modal summarization does not presume the existence of such well-formed semantic representations and grapples with the enormous amount of unstructured and uncoordinated information available in the real world.

Another form of fusion supports interactive and exploratory search. It interprets and guides users' queries linguistically and visually, progressing from the abstract to the concrete and thence to the specific. For example, initially, one may be interested in the annual movement of the oil price but later become interested in the change at a specific point in time, and finally, decide to investigate the cause and effect of that change. It also supports users' analysis of a series of events by showing various data from several viewpoints. The occurrence of typhoons is plotted on a geographic space and time scale and then linked to data on resultant damage and its associated verbal descriptions. At least two characteristics are required for such systems to be effective. Firstly, a framework is needed that seamlessly supports users throughout the information access process, from browsing an outline or summary to subsequent elaboration or specificity and to acquiring accurate information. Secondly, linguistic and non-linguistic information could be cooperatively employed in this process. Information need not be limited to text but may include non-linguistic information such as a series of numerical values. Non-linguistic modes could be utilized even during presentation, which would then lead onward to multi-modal presentation and information visualization.

The term, multi-modal summarization, is also used for the second technology, though the name does not adequately emphasize the significance of interactivity and relationship to exploratory search. These technologies share the name because these techniques have a common core that compiles useful and relevant information and presents it to users utilizing multiple modes, including text and visuals.

5.3 Multi-modal Summarization on Trend Information

The MuST was a workshop on multi-modal summarization focused on trend information (Kato et al. 2005, 2007a, b, 2008). Why did we focus on trend information? It was because a trend, which is a general tendency in the way a situation is changing or developing,[1] is based on temporal statistical data and can be obtained by synthetically summarizing it, but not by simple enumeration. Trends are the first answers to users' questions such as "How has the game machine industry performed since 2006?", "How have oil and gasoline prices changed this year?", and "How bad were the typhoons last year?" Each answer to those questions can be considered a summary

[1] From Longman Advanced American Dictionary.

of all the information that users are interested in and a starting point for interactive and explorative information access.

The information from which trends are composed and the process of identifying trends have several interesting features. First, to obtain trends, it is necessary to compile information spanning a specific and extensive period. As they include significant redundancies, such compilations must be synthetic and well organized. Secondly, trends usually contain summaries of non-linguistic information, for example, statistical information such as time-series data and geometric data. Some statistics such as political party approval ratings and companies' market share of a given product type are more complicated and have other dimensions. Each dimension could be an axis representing those statistics and bring different summarization methods. Thirdly, not only information such as reports on changes in statistical data, but also their interpretation, analysis of causes, and forecasts of impacts are important and should be included when defining trends.

As trend compilation requires sophisticated processes for handling complex and diverse information, it is an important research subject for multi-modal summarization aimed at supporting interactive and explorative information access.

5.3.1 Objective

The objective of the MuST workshop was to create an agora or arena where researchers from the several fields mentioned above could interact. The workshop prioritized trend analysis as its common theme because trends have interesting characteristics that are suitable as the starting point for exploratory search and as a subject for analysis. The MuST workshop promoted both cooperative and competitive research on trend information. It was not an evaluation workshop and thus identified neither a specific task nor evaluation measures.[2] For many, the workshop was motivated by a common evaluation. Sometimes the objective of the workshop was to enable large-scale evaluation, which required to employ the pooling method. It is beneficial to evaluate technologies on the common ground using standard measures. That, however, is only possible when technologies have matured or when they are focused on common objectives. Research on multi-modal summarization consists of many kernels of technologies still in development and not synthesized yet. Accordingly, each research group had its specific focus. In that situation, neither a common evaluation nor shared tasks were possible or stimulating. That is why we did not conduct an evaluation-oriented workshop. We needed another motivation to make the workshop cooperative and competitive, yet still, allow the participants to focus on their interests.

The MuST workshop was conducted a bit earlier than the IEEE VAST shared-task evaluation (IEEE symposium on VAST 2006). Although both were concerned

[2] In its third cycle, however, some evaluation tasks were set. Those tasks were considered as shared building blocks common to trend information summarization.

with visualization technology, they were different in nature. MuST addressed various problems, rather than a substantial single problem such as the one that IEEE VAST undertook. Rather, the policy of MuST was similar to that of the interactive track held in TREC 6 (Dumais and Belkin 2005), in which, through a common experiment, the participants conducted their own studies; such individual studies are more productive than a joint evaluation. During the MuST workshop, many technologies reflecting each participant's interests were examined. Although they would be associated with each other later in the process, initially, they did not have the same goal.

5.3.2 *Data Set as a Unifying Force*

Instead of a common topic for evaluation, a data set provided a unifying force for the MuST workshop. The use of a shared resource, which motivated researchers to participate and to conduct several research missions, was the major characteristic of the workshop. The resources that were shared, *the MuST data set*, included the materials to be processed, the intermediate results acting as the organizational hub, and the eventual output design.

The core of the data set is annotated newspaper articles concerning statistics and a wide variety of topics.[3] The topics were drawn from disparate social and economic domains, such as the oil industry, the personal-computer market, and car production; groups of events such as earthquakes and typhoons; and organizations such as Sony Corp. Linguistic descriptions of statistics and reports on events in articles were identified and annotated, as trends would be extracted from them. For example, trends in the personal-computer industry included statistics on shipment volume, shipment value, and market share of major manufacturers. Typhoon trends consisted of a review of typhoon-related events, such as their formation, landfalls, and related damage statistics.

Examples of English texts to which the annotation schema was applied are shown in Fig. 5.1, instead of the real data, which is in Japanese. Sentences mentioning selected statistics or events are annotated as `unit` elements. From the text of an `unit` element, phrases mentioning the name of the statistic (`name` element), the value of the statistic (`val` element), the relative values, which are associated with the statistic but are not the value itself (`rel` element), dates (`date` element), and other parameters (`par` element) are identified and annotated.

The annotation of the MuST data set represents the intermediate result of semantic and pragmatic analysis tuned to statistical and/or event information. In the summarization, extraction and analysis of important sentences are followed by rephrasing and sentence construction to eliminate redundancy and maintain consistency. Annotation corresponds to the output of extraction and analysis and the input to rephrasing and sentence generation. Using the terminology of the information extraction field,

[3] Articles of the Japanese Mainichi newspapers from 1998 and 1999 were used.

<unit stat="nationwide average of pump price of gasoline"><del type="src">Based on the July 6th report announced by the Oil Information Center,</del> <name part="head">the price of gasoline (one liter, regular)</name>, based on the research conducted <date gra="week" abs="19990617">this week</date>, reached a <name part="foot">national average</name> of <val>92 yen</val>, <rel type="diff">1 yen</rel> higher than <date gra="week" abs="19990610">last week</date>'s <name part="head">average price</name></unit>.

<unit stat="Dubai oil price"><name>The oil price (Dubai Oil)</name> has kept dropping since its <rel type="ord">peak</rel> <date gra="month" abs="199710">last October</date>, of <val>around $20</val> <name part="foot">per barrel</name>, and fell to <val>$12.50</val>. in <date gra="ten-days" abs="19980121">late January</date></unit>. <unit stat="Dubai oil price">After <date gra="ten-days" abs="19980121">that</date>, ø<ins type="name">Dubai oil price</ins> rose temporarily, <del type="rsn">because of tension due to the Iraq situation,</del> but has been struggling recently at <val>around $10</val><del type="other">, due to oversupply "</del></unit>.

<unit event="typhoon landfall">Medium-strength <par>typhoon No. 10</par> struck <par>Makuraszaki-shi, Kagoshima</par>, at <date gra="hour" abs="19980917">about 4:30 pm on the 17th</date>, and will strike in <par>the vicinity of Shukumo-shi, Kochi</par> <date gra="hour" abs="19980917">the same night</date></unit>.

<unit stat="domestic shipment volume"><name>The domestic shipment volume </name>for <date gra="half-year" abs="199804">the first half of the year</date> was <val>4,391,000</val>, which is <rel type="prop">34%</rel> higher than for <date gra="half-year" abs="199704">the same period last year</date> and marked <rel type="ord">the highest level</rel> for <name part="foot">the half-year range</foot></unit>.

Fig. 5.1 Examples of MuST data annotations on English text

this annotation completes named-entity recognition and temporal-expression analysis. For researchers who are interested in sentence extraction or text processing on named-entity recognition and temporal-expression analysis, annotation can be referred to as the gold standard of their process. It can also be used as training data if they take a machine learning approach. For researchers interested in rephrasing, sentence generation, and information visualization, annotation can be used as input data in which several fundamental analyses are already completed. In extreme cases, studies on information visualization from the text could be conducted without text processing. In this sense, the annotated articles behave as a hub for multi-modal summarization.

Multi-modal summarization requires several component technologies that are dispersed across many research fields. This makes it difficult to construct an integrated system. By using this data set, nevertheless, the participants can address their own subjects of interest. This is especially important for those studying elemental technologies. Moreover, participants from different communities can discuss their interests with each other using the data set as common ground and can contemplate how their studies or their modules fit into the framework. Of course, researchers

having the same interest can use the data set as material for objective evaluation. To encourage and foster research through such interchanges was the objective of sharing this research resource and of the MuST workshop.

5.3.3 Outcome

Many research themes were pursued in the MuST workshop and several technologies emerged from it. These include extraction of statistics from texts as materials for trend summarization; visualization of statistical information extracted and/or collected; generation of text that explains statistical information; and trend mining that is a version of text mining, and attempt to find and visualize trends from huge document sets.

5.3.3.1 Trend Information Extraction

Information extraction on statistics from the text was a major sub-problem of trend summarization. Many participants had addressed this problem, which is the reason that this theme was pursued in the evaluation-workshop style at the final cycle of the MuST workshop.

The simplest form of information extraction is to obtain as many tuples as possible of three elements; the name of a statistic, the date, and a value for the statistic on that date, an example of which looks like this; (Dubai oil price, 1998/12/21, $12.50). That triplet constitutes points plotted on the chart depicting the changes or trends of a given statistical category. Many complicated problems would remain even if the date and numeral expressions could be extracted using techniques of named-entity recognition. Those difficulties are epitomized in the first passage shown in Fig. 5.1, "the price of gasoline (one liter, regular), . . ., reached a national average of 92 yen, 1 yen higher than last week's average price."

First, the names of statistics are long and complex; they are frequently abbreviated and may be expressed in more than one way. These are usually expressed as a noun phrase, but sometimes split into many phrases. That is the case in this example in which the name of the statistic discussed is a national average of pump price of gasoline (one liter, regular). A method to handle such complex names of statistics was proposed. It deconstructs statistic names into their components and categorizes those characteristics and functions. To identify the name in its entirety, the method first identifies each component by text-chunking and then assembles those components into one name (Mori et al. 2008).

Second, not all numerical expressions directly describe the statistical values. Some of them are comparative or relative expressions. In this example, "1 yen" is not a gasoline price itself but the differential of two prices. Such relative expressions must be distinguished from direct expressions of statistical values. On the other hand, using such comparisons, an additional triplet instance of the gasoline price, "last

week," and "91 yen" could be obtained. Methods were proposed for distinguishing those expressions and using them to obtain additional triplet instances.

Besides, relative or context-dependent time expressions such as "last week" and cases where more than one statistic is mentioned in a sentence raised problems that are still to be solved.

Other research paid attention to extraction of information beyond simple triplets. Qualitative expressions, such as "peak" and "keep dropping" in the second passage in Fig. 5.1, were used cooperatively with numerical data representations for trend summarization. Descriptions of causes of events described such as "because of the tension of the Iraq situation" are useful for understanding context. Techniques were proposed for extracting and using such descriptions for summarization and visualization purposes.

5.3.3.2 Visualization

The interactivity of visualization was a major feature identified as an objective in the MuST workshop. Interactivity allows for interactive and exploratory search. Techniques were proposed that would assist users to analyze trends from various viewpoints and provide response mechanisms for new requests that emerged from such analysis.

Figure 5.2a shows an example of visualization as information access interface (Matsushita et al. 2004). A line chart was used as an information access interface. The chart as a whole represents the changes of a statistic of interest. The data points and segments are connected to the article that describes those statistics. Users can easily go back and forth between the chart and the articles as they are interconnected. This is a technique known as brushing (Scherr 2008). In another visualization shown in Fig. 5.2b, the line chart is augmented by schematic shapes that represent qualitative changes extracted from articles, such as "rebounding" and "continuing to increase" (Matsushita and Kato 2006). This chart can also be interconnected with textual materials. This is a typical example of multi-modal summarization.

For data analysis using visualization, a framework named a visualization cube was proposed (Takama and Yamada 2009, 2010). Events such as earthquakes which are characterized by time and geographical locations have their features represented as a cube, which allows the systematic manipulation of visual representation according to changes in the user's viewpoint. That is, a user can, through intuitive operations, freely place earthquakes of interest on a topographical map or on a timeline. Figure 5.3 schematically shows this operation. Statistics can be handled similarly. Each statistic corresponds to one cube and the cubes can be stacked upon each other. This operation corresponds to drawing a stacked bar chart. Changing the granularity of the chart or focusing on a specific data range are also defined as operations of particular cubes. Thus, it is a visualized version of an OLAP cube (Codd et al. 1993) used in online data analysis.

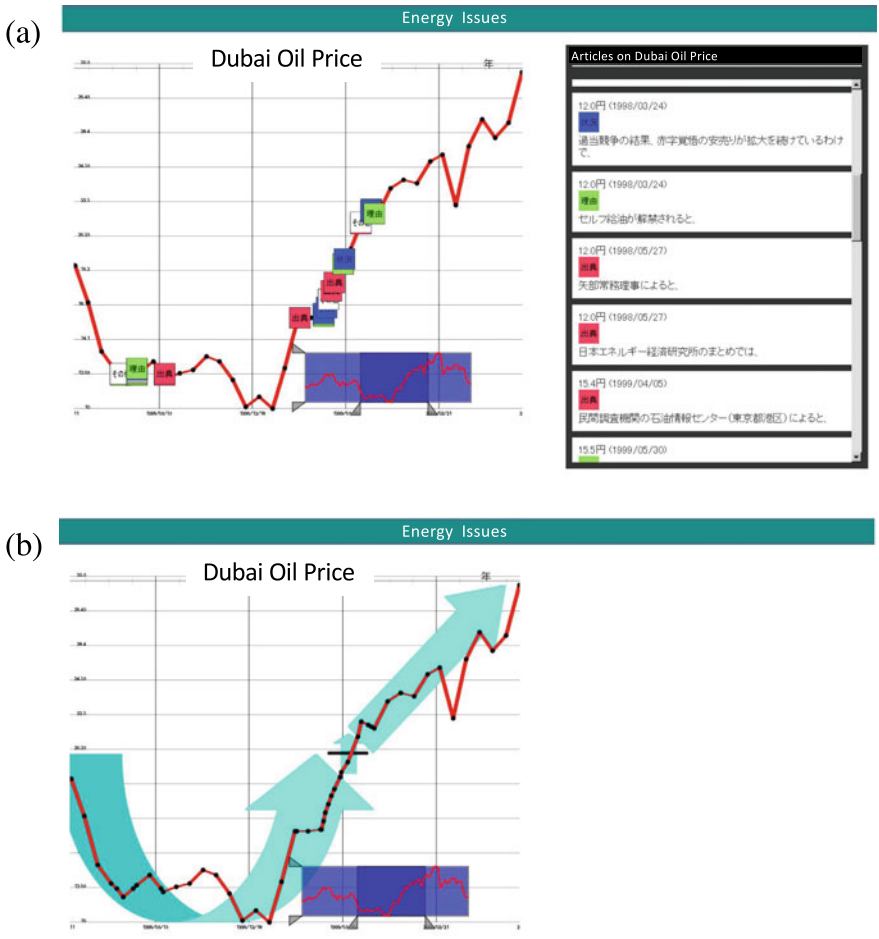


Fig. 5.2 Two examples of visualization for information access interface from Matsushita et al. (2004), Matsushita and Kato (2006)

5.3.3.3 Linguistic Summary Generation from Charts

Summarization can be done using linguistic expressions. A typical approach is to redact long documents into succinct phrases. In multi-modal summarization, series of numbers, tables, and charts can be verbalized. This makes it possible for complex numerical dynamics to be expressed in a short descriptive phrase such as "wild gyration."

This method was proposed for generating paragraph-length documents to explain a line chart of a given set of statistics. (Kobayashi et al. 2007; Kobayashi and Okumura 2008). The method for determining such content is critical. The chart is segmented, and a description of the relevant values and a description of the shape of the segments

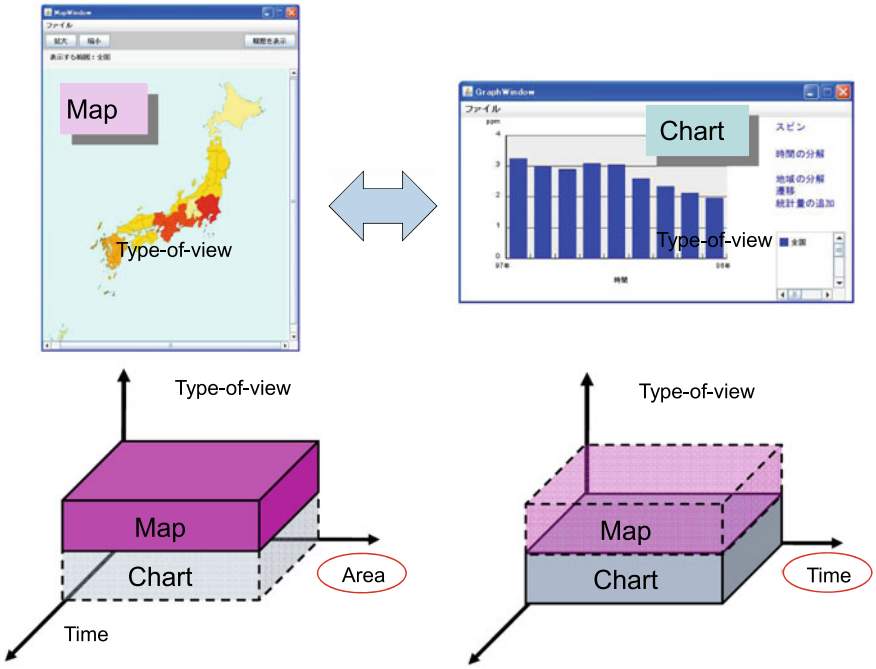


Fig. 5.3 Visualization for data analysis from Takama and Yamada (2009)

are decided and then appropriately linked to the content. The sets of two types of texts, those for describing values and those for shapes, are stored and used in the system as linguistic knowledge that is drawn from the corpus of real-life human explanations.

5.3.3.4 Trend Mining

Some trend summarizations can be conducted with a broader perspective via a version of text mining, which we termed as trend mining, that reveals current trends. Keywords, such as names of statistics, are linked to relevant topics. The observation that certain keywords appear frequently in documents reveals a trend that specific subjects are topical. Moreover, the co-occurrence pattern of those keywords suggests their relationship. One proposal visualized the relationship of statistical terms by calculating their co-occurrence frequencies. Such patterns are characteristic of events and phenomena in the real world (Kawai et al. 2008). The dynamic network established in this way allows users to review the structure of complex and global problems. Reviewing this, the user can discover the structure of a given problem and other useful related factors, thus facilitating access to accurate information about it.

5.4 Implication

The MuST workshop was conducted from 2005 to 2008 at the NTCIR-5, 6, and 7 workshops. It was a pilot task at first, and then became a core task with an evaluation subtask. Research activities on multi-modal summarization and trends went beyond these workshops. For five years, since 2006, special theme sessions were held at annual conferences of the Japan Society for Artificial Intelligence (JSAI). These focused on information compilation (Kato and Matsushita 2006), which aimed at using multi-modal summarization as an interface for interactive information access. It was emphasized that linguistic and non-linguistic information should be managed and utilized seamlessly. In 2009, a special interest group of the same name was launched by the JSAI. In 2012, it was renamed to Interactive Information Access and Visual Mining, and its activities have continued to the present (SIG-AM 2020).

In the NTCIR workshops, at NTCIR-8, an evaluation task was conducted on interactive information access using visual information (Kato et al. 2011). The patent information mining task in NTCIR-8 also handled text data and numerical data and extracted some trends observed in patent information (Nanba et al. 2010).

It is doubtful whether the MuST workshop itself had any direct influence on subsequent research trends. The workshop, however, contributed to advancing research on information access. Explanatory search has since become a key research area. Visual interfaces are an important component of such research. The MuST workshop was a significant catalyst in these developments.

Acknowledgements The author thanks Mitsunori Matsushita and Noriko Kando, co-organizers of the MuST workshop, for their contribution to organizing the workshop. The author also thanks all the participants of the workshop for their valuable research efforts on multi-modal summarization.

References

Ahmad S, de Oliveria PCF, Ahmad K (2004) Summarization of multimodal information. In: Proceedings of LREC-2004, pp 1049–1052

Codd EF, Codd SB, Salley CT (1993) Providing OLAP to user-analysts: an IT mandate. Technical report, E. F. Codd and Associates

Dumais ST, Belkin NJ (2005) The TREC interactive tracks: putting the user into search. In: Voorhees EM, Harman DK (eds) TREC experiment and evaluation in information retrieval. The MIT Press, Cambridge, pp 123–152

Fasciano M, Lapalme G (1996) Postgraphe: a system for the generation of statistical graphics and text. In: Proceedings of 8th international workshop on natural language generation, pp 51–60

IEEE symposium on VAST (visual analytics science and technology) 2006 (2006). http://www.cs.umd.edu/hcil/VASTcontest06/

Kato T, Matsushita M, Kando N (2005) MuST: a workshop on multimodal summarization for trend information. In: Proceedings of the 5th NTCIR workshop meeting on evaluation of information access technologies, pp 556–563

Kato T, Matsushita M, Kando N (2007a) Expansion of multimodal summarization for trend information –report on the first and second cycles of the MuST workshop–. In: Proceedings of the 6th NTCIR workshop meeting, pp 235–242

Kato T, Matsushita M, Kando N (2007b) Fostering multi-modal summarization for trend information. In: Proceedings of KES2007, pp 377–386
Kato T, Matsushita M, Kando N (2008) Overview of MusT at the NTCIR-7 workshop –challenges to multi-modal summarization for trend information–. In: Proceedings of the 7th NTCIR workshop meeting, pp 475–488
Kato T, Matsushita M (2006) Toward information compilation (in Japanese). In: Proceedings of the 20th annual conference of the Japan society for artificial intelligence, 1D3-2
Kato T, Matsushita M, Joho H (2011) Overview of the VisEx task at NTCIR-9. In: Proceedings of the 9th NTCIR workshop meeting, pp 526–532
Kawai H, Kunieda K, Yamada K, Saito H, Tsuchida M, Mizuguchi H (2008) Visualization for statistical term network in newspaper. In: Proceedings of the 7th NTCIR workshop meeting, pp 549–554
Kobayashi I, Watanabe C, Okumura N (2007) Intelligent information presentation based on collaboration between 2D chart and text-with an example of Nikkei stock average text and its 2D charts presentation (in Japanese). Trans Inform Proc Soc Jpn 48(3):1058–1070
Kobayashi I, Okumura N (2008) Text generation for explaining the behavior of 2D charts: with an example of stock price trends. In: Proceedings of the 7th NTCIR workshop meeting, pp 515–519
Matsushita M, Nakakoji K, Yamamoto Y, Kato T (2004) InTREND: an interactive tool for reflective data exploration through natural discourse. In: Proceedings of KES2004, vol 2, pp 148–155
Matsushita M, Kato T (2006) Statistical chart generation from multiple documents based on numerical data supplement and chart shape suggestion (in Japanese). Trans Jpn Soc Fuzzy Theory Intell Inform 18(5):721–734
Mori T, Fujioka A, Murata I (2008) Automated extraction of statistical expressions from text for information compilation (in Japanese). Trans Jpn Soc Artif Intell 23(5):310–318
Nanba H, Fujii A, Iwayama M, Hashimoto T (2010) Overview of the patent mining task at the NTCIR-8 workshop. In: Proceedings of the 8th NTCIR workshop meeting, pp 293–302
Roth SF, Mattis J (1990) Data characterization for intelligent graphic presentation. In: Proceedings of conference on human factors in computing systems, pp 193–200
Scherr M (2008) Multiple and coordinated views in information visualization. In: Trends in information visualization, vol. 38, pp 1–8
SIG-AM (special interest group on interactive information access and visual mining) (in Japanese) (2020). https://must.c.u-tokyo.ac.jp/sigam/
Takama Y, Yamada T (2009) Visualization cube: modeling interaction for exploratory data analysis of spatiotemporal trend information. In: Proceedings of IWI2009 (WI-IAT2009), pp 1–4
Takama Y, Yamada T (2010) Application of visualization cube to analysis of interaction pattern in exploratory data analysis of spatiotemporal trend information. In: Proceedings of ISCIIA2010, pp 85–93

Chapter 6
Opinion Analysis Corpora Across Languages

Yohei Seki

Abstract At *NTCIR-6*, *7*, and *8*, we included a new multilingual opinion analysis task (*MOAT*) that involved Japanese, English, and Chinese newspapers. This was the first task that compared the performance of sentiment retrieval strategies with common subtasks across languages. In this paper, we introduce the research question posed by *NTCIR MOAT* and present what has been achieved to date. We then describe the types of tasks and research that have involved our test collection both previously and in current research. Finally, we summarize our contributions and discuss future research directions.

6.1 Introduction

Sentiment analysis (sometimes called "opinion mining") is a research topic that has been actively discussed and developed for some 20 years, particularly in the fields of natural language processing (NLP) and information retrieval (IR) (Pang and Lee 2008). In this paper, we introduce the multilingual opinion analysis task (*MOAT*) (Seki et al. 2010, 2008, 2007), which was included in *NTCIR-6*, *7*, and *8* (2006–2010). We then discuss the role and novelty of the task in sentiment analysis research.

Sentiment analysis research began in 2002 (Pang et al. 2002; Turney 2002; Wiebe et al. 2002). Various frameworks for classifying documents in terms of positivity or negativity that use either supervised learning (Pang et al. 2002) or unsupervised learning (Turney 2002) have been proposed. In parallel, many researchers started to build opinion corpora based on newspaper articles (Wiebe et al. 2002) for multi-perspective question answering (*MPQA*). Other early research work was published at the *AAAI 2004 Spring Symposium: Exploring Attitude and Affect in Text: Theories and Applications* (Shanahan et al. 2006).

Y. Seki (✉)
University of Tsukuba, 1-2 Kasuga, Tsukuba, Ibaraki 305-8550, Japan
e-mail: yohei@slis.tsukuba.ac.jp

T. Sakai et al. (eds.), *Evaluating Information Retrieval and Access Tasks*,
The Information Retrieval Series 43,
https://doi.org/10.1007/978-981-15-5554-1_6

At the Text Retrieval Conference (TREC) in 2006, a new "Blog Track" was introduced, and was continued until 2010.[1] The original organizers released the TREC Blogs06 Collection (Macdonald and Ounis 2006), for which there have been 100,649 blog posts (excluding duplicate documents) and over 3.2 million permalinks. This dataset was used for the opinion finding (blog post) retrieval task in the *TREC 2006 Blog Track* and for the polarity opinion finding (blog post) retrieval task in the *TREC 2007 Blog Track*. In addition, the *MPQA* opinion corpus from the University of Pittsburgh (Wiebe et al. 2005), which defines a framework for opinion annotation using multiple assessors, has been released.

Building on this previous work, we introduced our opinion analysis task at *NTCIR-6* in 2006. The novel aspects of the *NTCIR MOAT* task can be summarized as follows:

1. We have released an opinion annotation corpus for evaluation workshops. The annotation units include opinionatedness, topic relevance, polarity, opinion holder (from *NTCIR-6*), and opinion target (from *NTCIR-7*).
2. We have provided a multilingual opinion corpus that includes material in English, Chinese, and Japanese.
3. The topic set in the evaluation corpus is shared across languages.

In Sect. 6.2, we give details of the *NTCIR MOAT* design to clarify its novel features and suggest an opinion corpus annotation strategy for evaluation workshops. In Sect. 6.3, we explain the evolution of opinion analysis research since the introduction of *MOAT*. Finally, in Sect. 6.4, we conclude our remarks and discuss future research directions.

6.2 NTCIR MOAT

6.2.1 Overview

NTCIR MOAT was held at NTCIR-6 (Seki et al. 2007), NTCIR-7 (Seki et al. 2008), and NTCIR-8 (Seki et al. 2010). The task definition evolved through the three sessions, as shown in Table 6.1.

The goal of the task is to form a bridge between element technologies such as opinion/polarity sentence classification or opinion holder/target phrase recognition to an application such as (opinion) IR or question answering. The target languages include English, Chinese (both Traditional and Simplified), and Japanese, and the topic set for IR or question answering is shared across languages. We have prepared a document set relevant to the topics retrieved from newspaper articles published in each target language, and have evaluated the system using these document sets annotated with multiple assessors.

[1] http://trec.nist.gov/data/blog.html.

Table 6.1 *MOAT* progress during *NTCIR-6, 7, & 8*

	NTCIR-6	*NTCIR-7*	*NTCIR-8*
Target Language	English, Japanese, Traditional Chinese		
	–	+Simplified Chinese	
Subtasks	Opinionated, Relevance, Polarity, Holder	+Target	+Cross-lingual
Annotation Unit	Sentence	Opinion Clause	
Focused Application	Information Retrieval	Q&A (*ACLIA*[a])	Opinion Q&A
Target Corpora	Mainichi, Yomiuri, CIRB, Xinhua English, Hong Kong Standard, etc.	+Xinhua Chinese	+NYT, UDN
(Period)	1998–2001		2002–2005

[a]http://research.nii.ac.jp/ntcir/permission/ntcir-7/perm-ja-ACLIA.html

6.2.2 Research Questions at NTCIR MOAT

Many researchers have focused on a resourceless approach to sentiment analysis (Elming et al. 2014; Le et al. 2016). Blitzer et al. (2007) proposed a domain adaptation approach for sentiment classification. Wan (2009) addressed the Chinese sentiment classification problem by using English sentiment corpora on the Internet. This type of research can be categorized as a semi-supervised approach to opinion/sentiment analysis that aims to solve the resource problem by using small labeled and large unlabeled datasets. We recognize that addressing language resource problems in sentiment analysis for nonnative languages is an important research area. Alternatively, applications such as the *Europe Media Monitor* (*EMM*) News Explorer[2] provide an excellent service by including viewpoints from different countries. We also understand that providing these varied opinions from different countries offers opportunities for better worldwide communications. *NTCIR MOAT* is the first task to provide opportunities for nonnative researchers to develop a sentiment analysis system for low-resource languages and to bridge cultures by clarifying opinion differences across different languages.

[2]http://emm.newsbrief.eu/NewsBrief/clusteredition/en/latest.html.

6.2.3 Subtasks

With the broad range of information sources available on the web and in social media, there has been increased interest by both commercial and governmental parties in trying to analyze and monitor the flow of prevailing attitudes from anonymous users automatically. As a result, the research community has given much attention to automatic identification and processing of the following.

- Sentences in which an opinion is expressed (Wiebe et al. 2004),
- The polarity of the expression (Wilson et al. 2005),
- The opinion holders of the expression (Choi et al. 2005),
- The opinion targets of the experssion (Ruppenhofer et al. 2008), and
- Opinion question and answering (Stoyanov et al. 2005), (Dang 2008).[3]

With these factors in mind, we defined the subtasks in *NTCIR MOAT* as follows.

1. Opinionated sentences
   The judgment of opinionated sentences is a binary decision for all sentences.
2. Relevant sentences
   Each set contains documents that are found to be relevant to an opinion question, such as that shown in Fig. 6.1. For those participating in the relevance subtask evaluation, each opinionated sentence should be judged as either relevant (Y) or non-relevant (N) to the opinion questions. In *NTCIR-8 MOAT*, only opinionated sentences were annotated for relevance.
3. Opinion polarities
   The polarity is determined for each opinion clause. In addition, the polarity is to be determined with respect to the topic description if the sentence is relevant to the topic, and based on the attitude of the opinion if the sentence is not relevant to the topic. The possible polarity values are positive (POS), negative (NEG), or neutral (NEU).
4. Opinion holders
   The opinion holders are annotated in terms of opinion clauses that express an opinion. However, the opinion holder for an opinion clause can occur anywhere in the document. The assessors performed a kind of co-reference resolution by marking the opinion holder with the opinion clause if the opinion holder makes an anaphoric reference noting the antecedent of the anaphora. Each opinion clause must have at least one opinion holder.
5. Opinion targets
   The opinion targets were annotated in a similar manner to the opinion holders. Each opinion clause must have at least one opinion target.
6. Cross-lingual opinion Q&A
   The cross-lingual subtask is defined as the opinion Q&A task. Together with the questions in English, the answer opinions should be extracted in different languages. To keep it simple, the extraction unit is defined as a sentence. The

[3] https://tac.nist.gov/2008/qa/index.html.

```
<TOPIC>
<NUM>N03</NUM>
<TITLE>Bali Island Terrorist Bombing</TITLE>
<QUESTION>What reasons behind the 2002 Bali bombings were discussed?</QUESTION>
<POLARITY>Neutral</POLARITY>
<OPTYPE>Reason controversy</OPTYPE>
<CONC>Bali Island Terrorist Bombing</CONC>
<PERIOD>2002–1010</PERIOD>
</TOPIC>
```

Fig. 6.1 Example: opinion question fields at *NTCIR-8 MOAT*

answer set is defined as the combination of the annotations for the conventional subtasks, with opinionatedness, polarity, and answeredness being matched with the definition in the question description.

6.2.4 Opinion Corpus Annotation Requirements

Opinion corpus annotation for multiple domains (as in news topics) usually requires expert linguistic knowledge because crowdsourcing annotation (such as the Amazon Mechanical Turk) does not fit the *NTCIR MOAT* annotation framework. We conducted our evaluation using agreed (intersection) annotations from multiple expert assessors. To check the stability of this evaluation strategy, we compared the evaluation results for agreed (intersection) annotation and selective (union) annotation to arrive at a gold standard for using *NTCIR-8 MOAT* submission data.

For the English cases in Table 6.2 (the κ coefficient between assessor annotations was 0.73) and the Traditional Chinese cases in Table 6.3 (κ coefficient 0.46), the rank of the participants' systems is different. Although the rank differences for the English cases were within statistical significance, among the Traditional Chinese cases, the precision-oriented systems (*CTL* and *WIA*) tended to be ranked higher for cases of agreed (intersection) annotation, and recall-oriented systems (*KLELAB-1* and *NTU*) tended to be ranked lower. For the Simplified Chinese cases in Table 6.4 (κ coefficient 0.97) and the Japanese cases in Table 6.5 (κ coefficient 0.72), there was no rank difference for the participants' systems despite the different strategies because of either high κ agreement (Simplified Chinese) or a low number of participants (Japanese). From these observations, we concluded that the κ coefficient between assessor annotations should exceed 0.7 for stable evaluation. We also found that strong opinion definition and online annotation tools were helpful, but using expert linguistic annotators remained necessary to achieve high κ agreement.

Table 6.2 Evaluation strategy analysis using *NTCIR-8 MOAT* English raw submission data

English (F1-score)/$\kappa = 0.73$

Rank on agreed	Significance								Rank on non-agreed
UNINE-1	A								UNINE-1
NECLC-bsf	A	B							NECLC-bs1
NECLC-bs0	A	B	C						NECLC-bsf
NECLC-bs1	A	B	C	D					NECLC-bs0
UNINE-2		B	C	D					UNINE-2
KLELAB-3		B	C	D	E				NTU-2
KAISTIRNLP-2			C	D	E				KLELAB-2
KLELAB-2			C	D	E				KLELAB-3
KAISTIRNLP-1			C	D	E				NTU-1
NTU-2				D	E	F			KLELAB-1
KLELAB-1				D	E	F			KAISTIRNLP-2
NTU-1				D	E	F			KAISTIRNLP-1
OPAL-2					E	F	G		OPAL-1
OPAL-3						F	G		OPAL-2
OPAL-1						F	G		OPAL-3
PolyU-1							G		SICS-1
SICS-1							G	H	PolyU-1
PolyU-2								H	PolyU-2

6.2.5 Cross-Lingual Topic Analysis

We ranked topics by averaging their F1-scores, the harmonic mean of precision and recall, obtained from all *NTCIR-8 MOAT* raw submissions in the opinionated judgment subtask. The best three (easy) topics and worst three (difficult) topics and the opinion percentage in the source documents are shown in Table 6.6.

From these results, we found that the topic difficulty is strongly related to each language. We also found that, with many opinions in the source, the topics tended to be easier. Exceptions to this rule included the opinion question for topic *N16*: "What reasons have been given for the anti-Japanese demonstrations that took place in April, 2005 in Peking and Shanghai in China?" We surmise that this was caused by the systems' difficulty in judging quite sensitive opinions expressed in newspaper articles in each language.

Table 6.3 Evaluation strategy analysis using *NTCIR-8 MOAT* traditional Chinese raw submission data

Traditional Chinese (F1-score)/$\kappa = 0.46$					
Rank on agreed	Significance				Rank on non-agreed
CityUHK-2	A				CityUHK-1
CTL-1	A				CityUHK-3
CityUHK-1	A				KLELAB-1
CityUHK-3	A				NTU-1
WIA-1	A				NTU-2
WIA-2	A				CityUHK-2
KLELAB-3		B			cyut-1
KLELAB-1		B			KLELAB-3
NTU-2		B			WIA-1
NTU-1		B			WIA-2
cyut-1		B			cyut-2
cyut-2		B	C		CTL-1
UNINE-1			C	D	UNINE-1
cyut-3				D	cyut-3

Table 6.4 Evaluation strategy analysis using *NTCIR-8 MOAT* simplified Chinese raw submission data

Simplified Chinese (F1-score)/$\kappa = 0.97$						
Rank on agreed	Significance					Rank on non-agreed
PKUTM-2	A					PKUTM-2
PKUTM-1	A	B				PKUTM-1
BUPT-2	A	B				BUPT-2
CTL-1		B				CTL-1
PKUTM-3		B	C			PKUTM-3
BUPT-1		B	C			BUPT-1
WIA-1			C	D		WIA-1
WIA-2			C	D		WIA-2
NECLC-bsf				D		NECLC-bsf
NECLC-bs0				D		NECLC-bs0
NECLC-bs1				D		NECLC-bs1
PolyU-1					E	PolyU-1

Table 6.5 Evaluation strategy analysis using *NTCIR-8 MOAT* Japanese raw submission data

Japanese (F1-score)/$\kappa = 0.72$					
Rank on agreed	Significance				Rank on non-agreed
TUT-1	A				TUT-1
TUT-3	A	B			TUT-3
IISR-3		B	C		IISR-3
TUT-2		B	C		TUT-2
IISR-1		B	C		IISR-1
IISR-2			C		IISR-2
UNINE-1				D	UNINE-1

Table 6.6 Cross-lingual topic analysis using *NTCIR-8 MOAT* raw submission data

	English		Traditional Chinese		Simplified Chinese		Japanese	
	Topic	Opinion % in doc set	Topic	Opinion % in doc set	Topic	Opinion % in doc set	Topic	Opinion % in doc set
Easy topics	N27	25.4	N14	56.5	N18	24.5	N41	34.6
	N39	21.2	N05	55.6	N20	20.7	N11	35.3
	N14	21.3	N27	57.6	N06	22.7	N13	28.1
Difficult topics	N18	7.6	N16	19.4	N07	9.5	N24	35.3
	N13	8.9	N13	15.0	N41	14.9	N18	37.7
	N06	10.0	N20	18.8	N16	20.6	N32	27.0
Average	Avg.	16.7	Avg.	32.1	Avg.	18.6	Avg.	33.9

6.3 Opinion Analysis Research Since *MOAT*

6.3.1 Research Using the NTCIR MOAT Test Collection

Some researchers have used the *NTCIR MOAT* test collection and presented their work at top-rated conferences, particularly those focused on cross-lingual sentiment analysis. Two representative examples are as follows.

1. Joint Bilingual Sentiment Classification
   Lu et al. (2011) hypothesized that aligned sentences between languages should be similar in opinion polarity and strongness. They proposed a method for improving the polarity classification performance that used the *MPQA* opinion corpus and the *NTCIR MOAT* corpus as labeled corpora, and aligned news corpora in Chinese and English as unlabeled corpora. They extended their work by using a cross-

lingual mixture model (Meng et al. 2012) to improve performance when learning polarity clues from unlabeled corpora.

2. Cross-lingual Sentiment Lexicon Learning
   Gao et al. (2015) proposed a method for generating low-resource language sentiment lexicons using available English sentiment lexicons. They created Chinese sentiment lexicons using a bilingual word graph label propagation approach. They evaluated Chinese sentiment classification at the sentence level by using the *NTCIR MOAT* corpus and found increased effectiveness of sentiment classification when using their generated sentiment lexicon to generate features.

6.3.2 Opinion Corpus in News

Several opinion corpora involving news have been developed after *NTCIR MOAT* was published. In this subsection, we introduce the *SemEval-2007 Task 14: Affective Corpus* (Strapparava and Mihalcea 2007) and the *sentiment-annotated quotation set* (Balahur and Steinberger 2009; Balahur et al. 2010).

In the *SemEval-2007 Affective Corpus*, six emotion labels and two polarity labels have been annotated to headlines collected from 1,250 news websites and newspaper articles. The *sentiment-annotated quotation set* contains a set of 1,590 English language quotations (reported speech), manually annotated by two independent sets of annotators for sentiment (positive, negative, or objective/neutral) expressed toward the entities mentioned inside the quotation. Web crawling for news articles employed the *EMM* (Steinberger et al. 2009)[4] developed by the *European Commission Joint Research Centre*.

The *NTCIR MOAT* corpus, however, remains in use as a large cross-lingual news opinion corpus targeted at Chinese, Japanese, and English.

6.3.3 Current Opinion Analysis Research: The Social Media Corpus and Deep NLP

After *NTCIR MOAT* was published, Twitter[5] and other microblog media came into widespread use by many users. The NLP/IR researchers also focused on tweet sentiment analysis (Martinez-Camara et al. 2013). To improve sentiment classification in Twitter, specific clues were found to be useful because a tweet is much shorter than a news article, including tweet context (Jiang et al. 2011), emoticons and hashtags (Purver and Battersby 2012), lengthened words (Brody and Diakopoulos 2011), and emoji (Felbo et al. 2017).

[4] http://emm.newsbrief.eu/overview.html.

[5] http://twitter.com.

On the other hand, deep NLP research such as *Stanford Sentiment Treebank* (Socher et al. 2013)[6] has become mainstream from a technological point of view. In this research, the learning model builds up a representation of whole sentences based on the sentence structure. An opinion corpus called the *Stanford Sentiment Treebank* has been developed to estimate compositionality in the sentiment detection task. It includes the fine-grained sentiment labels "very negative", "negative", "neutral", "positive", and "very positive" for 215,154 phrases in trees parsed with the Stanford Parser from 11,855 sentences extracted from movie reviews (Pang and Lee 2005).

In *SemEval 2018* (Mohammad et al. 2018), an opinion corpus has been created from 10,983 English, 4,381 Arabic, and 7.094 Spanish tweets, and used to evaluate the systems. Several tasks are defined that provide annotations for the mental state of the tweeter, including (1) the intensities of the four basic emotions (anger, fear, joy, and sadness), (2) the intensity of sentiment/valence (very negative, moderately negative, slightly negative, neutral or mixed, slightly positive, moderately positive, and very positive), and (3) multi-label emotion classification across 12 emotions (anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, trust, and neutral). The corpus used *best–worst scaling* (Louviere et al. 2015), a comparative annotation method in which assessors were asked what was the best (highest in terms of the property) and worst (lowest in terms of the property), given n items (typically $n = 4$). Real-valued scores for the association between the items and the property were determined based on the number of times an item was chosen as the best and the worst. The median number of assessors for each tweet was seven. The inter-annotator agreements (*Fleiss's* κ) for the multi-label emotion classification were 0.21, 0.29, and 0.28 for the 12 classes, and 0.40, 0.48, and 0.45 for the four basic emotions in English, Arabic, and Spanish. Most of the participants employed SVM/SVR, LSTMs, and Bi-LSTMS as machine learning algorithms, and also took word embedding, affect lexicon features, and word n-grams as features.

Although the document genres being focused on and the annotation properties have changed over time, cross-lingual opinion corpora remain important in current research.

6.4 Conclusion

In this paper, we have discussed the contributions made by our development of *NTCIR MOAT*. We created a cross-lingual opinion corpus using the news document genre, following which, several researchers have conducted cross-lingual opinion research using our test collections. Although sentiment classification accuracy is improved by using a cross-lingual corpus, research investigating linguistic opinion properties characterized by languages rooted in different cultures and opinion retrieval strategies preferable for different language characteristics remain to be undertaken.

[6] https://nlp.stanford.edu/sentiment/.

In recent research, high-quality contextual representations based on neural architectures such as ELMo (Peters et al. 2018a) and BERT (Devlin et al. 2019) are proving to be effective in NLP research. In addition, linguistic properties such as morphological, local-syntax, and longer-range semantics tend to be treated at different layers, such as the word-embedding layer, lower contextual layers, or upper layers in each of these cases (Peters et al. 2018b; Jawahar et al. 2019). As an extension of bilingual sentiment word-embedding frameworks (Zhou et al. 2015), cross-lingual sentiment retrieval research that considers syntax and semantics in different languages will be an interesting direction for future work.

Acknowledgements This work was partially supported by JSPS Grants-in-Aid for Scientific Research (B) (#19H04420).

References

Balahur A, Steinberger R (2009) Rethinking sentiment analysis in the news: from theory to practice and back. In: Proceedings of workshop on opinion mining and sentiment analysis' (WOMSA), Sevilla, Spain, pp 1–12

Balahur A, Steinberger R, Kabadjov M, Zavarella V, van der Goot E, Halkia M, Pouliquen B, Belyaeva J (2010) Sentiment analysis in the news. In: Proceedings of the 7th international conference on language resources and evaluation (LREC 2010), Valletta, Malta, pp 2216–2220

Blitzer J, Dredze M, Pereira F (2007) Biographies, bollywood, boom-boxes and blenders: domain adaptation for sentiment classification. In: Proceedings of the 45th annual meeting of the association of computational linguistics (ACL), pp 440–447

Brody S, Diakopoulos N (2011) Cooooooooooooooollllllllllll!!!!!!!!!!!!!! using word lengthening to detect sentiment in microblogs. In: Proceedings of the 2011 conference on empirical methods in natural language processing (EMNLP 2011), Edinburgh, Scotland, UK, pp 562–570

Choi Y, Cardie C, Riloff E, Patwardhan S (2005) Identifying sources of opinions with conditional random fields and extraction patterns. In: Proceedings of the 2005 human language technology conference and conference on empirical methods in natural language processing (HLT/EMNLP 2005), Vancouver, B. C

Dang HT (2008) Overview of the TAC 2008 opinion question answering and summarization tasks. In: Text analysis conference (TAC 2008) workshop notebook papers. Accessed 1 Apr 2010. Available from: http://www.nist.gov/tac/tracks/2008/index.html

Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), Minneapolis, Minnesota, pp 4171–4186. https://doi.org/10.18653/v1/N19-1423

Elming J, Plank B, Hovy D (2014) Robust cross-domain sentiment analysis for low-resource languages. In: Proceedings of the 5th workshop on computational approaches to subjectivity, sentiment and social media analysis, association for computational linguistics, Baltimore, Maryland, pp 2–7. https://doi.org/10.3115/v1/W14-2602, https://www.aclweb.org/anthology/W14-2602

Felbo B, Mislove A, Søgaard A, Rahwan I, Lehmann S (2017) Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In: Proceedings of the 2017 conference on empirical methods in natural language processing, Copenhagen, Denmark, pp 1615–1625. https://doi.org/10.18653/v1/D17-1169

Gao D, Wei F, Li W, Liu X, Zhou M (2015) Cross-lingual sentiment lexicon learning with bilingual word graph label propagation. Comput Linguist 41:21–40

Jawahar G, Sagot B, Seddah D (2019) What does BERT learn about the structure of language? In: Proceedings of the 57th annual meeting of the association for computational linguistics, Florence, Italy, pp 3651–3657. https://doi.org/10.18653/v1/P19-1356

Jiang L, Yu M, Zhou M, Liu X, Zhao T (2011) Target-dependent twitter sentiment classification. In: Proceedings of the 49th annual meeting of the association for computational linguistics (ACL 2011), Portland, Oregon, pp 151–160

Le TA, Moeljadi D, Miura Y, Ohkuma T (2016) Sentiment analysis for low resource languages: a study on informal Indonesian tweets. In: Proceedings of the 12th workshop on Asian language resources (ALR), The COLING 2016 Organizing Committee, Osaka, Japan, pp 123–131. https://www.aclweb.org/anthology/W16-5415

Louviere JJ, Flynn TN, Marley AAJ (2015) Best-worst scaling: theory, methods and applications. Cambridge University Press, Cambridge

Lu B, Tan C, Cardie C, Tsou BK (2011) Joint bilingual sentiment classification with unlabeled parallel corpora. In: Proceedings of the 49th annual meeting of the association for computational linguistics (ACL 2011), Portland, Oregon, pp 320–330

Macdonald C, Ounis I (2006) The TREC Blogs06 collection: creating and analysing a blog test collection. Technical report TR-2006-224, Department of Computing Science, University of Glasgow

Martinez-Camara E, Martin-Valdivia MT, Urena-Lopez LA, Montejo-Raez AR (2013) Sentiment analysis in twitter. Nat Lang Eng 11(2):1–28

Meng X, Wei F, Liu X, Zhou M, Xu G, Wang H (2012) Cross-lingual mixture model for sentiment classification. In: Proceedings of the 50th annual meeting of the association for computational linguistics (ACL 2012), Jeju, Republic of Korea, pp 572–581

Mohammad SM, Bravo-Marquez F, Salameh M, Kiritchenko S (2018) SemEval-2018 task 1: affect in tweets. In: Proceedings of the 12th international workshop on semantic evaluation (SemEval-2018), New Orleans, Louisiana, pp 1–17. https://doi.org/10.18653/v1/S18-1001

Pang B, Lee L (2005) Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In: Proceedings of the 43rd annual meeting of the association for computational linguistics (ACL 2005), Ann Arbor, Michigan, pp 115–124

Pang B, Lee L (2008) Opinion mining and sentiment analysis. Now Publishers Inc, Boston

Pang B, Lee L, Vaithyanathan S (2002) Thumbs up? sentiment classification using machine learning techniques. In: Proceedings of the 2002 conference on empirical methods in natural language processing (EMNLP 2011), pp 79–86

Peters M, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L (2018a) Deep contextualized word representations. In: Proceedings of the 2018 conference of the North American Chapter of the association for computational linguistics: human language technologies, volume 1 (long papers), New Orleans, Louisiana, pp 2227–2237. https://doi.org/10.18653/v1/N18-1202

Peters M, Neumann M, Zettlemoyer L, tau Yih W (2018b) Dissecting contextual word embeddings: architecture and representation. In: Proceedings of the 2018 conference on empirical methods in natural language processing (EMNLP 2018), Brussels, Belgium, pp 1499–1509

Purver M, Battersby S (2012) Experimenting with distant supervision for emotion classification. In: Proceedings of the 13th Eurpoean chapter of the association for computational linguistics (EACL 2012), Avignon, France, pp 482–491

Ruppenhofer J, Somasundaran S, Wiebe J (2008) Finding the sources and targets of subjective expressions. In: Proceedings of the 6th international language resources and evaluation (LREC'08), Marrakech, Morocco

Seki Y, Evans DK, Ku LW, Chen HH, Kando N, Lin CY (2007) Overview of opinion analysis pilot task at NTCIR-6. In: Proceedings of the 6th NTCIR workshop meeting, NII, Japan, pp 265–278

Seki Y, Evans DK, Ku LW, Sun L, Chen HH, Kando N (2008) Overview of multilingual opinion analysis task at NTCIR-7. In: Proceedings of the 7th NTCIR workshop meeting, NII, Japan, pp 185–203

Seki Y, Ku LW, Sun L, Chen HH, Kando N (2010) Overview of multilingual opinion analysis task at NTCIR-8 - a step toward cross lingual opinion analysis. In: Proceedings of the 8th NTCIR workshop meeting, NII, Japan, pp 209–220

Shanahan JG, Qu Y, Wiebe JM (2006) Computing attitude and affect in text: theory and applications. Springer, Berlin

Socher R, Perelygin A, Wu J, Manning JCCD, Ng A, Potts C (2013) Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of the (2013) conference on empirical methods in natural language processing (EMNLP 2013). Association for Computational Linguistics, Seattle, Washington, USA, pp 1631–1642

Steinberger R, Pouliquen B, van der Goo E, (2009) An introduction to the Europe media monitor. In: Proceedings of ACM SIGIR 2009 workshop: information access in a multilingual world, Boston, MA, USA

Stoyanov V, Cardie C, Wiebe J (2005) Multi-perspective question answering using the OpQA corpus. In: Proceedings of the 2005 human language technology conference and conference on empirical methods in natural language processing (HLT/EMNLP 2005), Vancouver, B. C

Strapparava C, Mihalcea R (2007) SemEval-2007 task 14: affective text. In: Proceedings of the 4th international workshop on semantic evaluations (SemEval-2007), Prague, pp 70–74

Turney P (2002) Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th annual meeting on association for computational linguistics, Philadelphia, USA, pp 417–424

Wan X (2009) Co-training for cross-lingual sentiment classification. In: Proceedings of the 47th annual meeting of the association of computational linguistics (ACL), Suntec, Singapore, pp 235–243

Wiebe J, Breck E, Buckley C, Cardie C, Davis P, Fraser B, Litman D, Pierce D, Riloff E, Wilson T (2002) NRRC summer workshop on multiple-perspective question answering final report

Wiebe J, Wilson T, Cardie C (2005) Annotating expressions of opinions and emotions in language. Lang Resour Eval 39(2–3):165–210

Wiebe JM, Wilson T, Bruce RF, Bell M, Martin M (2004) Learning subjective language. Comput Linguist 30(3):277–308

Wilson T, Wiebe J, Hoffmann P (2005) Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of the 2005 human language technology conference and conference on empirical methods in natural language processing (HLT/EMNLP 2005), Vancouver, B. C

Zhou H, Chen L, Shi F, Huang D (2015) Learning bilingual sentiment word embeddings for cross-language sentiment classification. In: Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: long papers), Beijing, China, pp 430–440. https://doi.org/10.3115/v1/P15-1042

Chapter 7
Patent Translation

Isao Goto

Abstract The NTCIR patent translation task was the first task for the machine translation of patents that used large-scale patent parallel sentence pairs. In this chapter, we first present the history of machine translation; the contribution of evaluation workshops to machine translation research, and previous evaluation workshops; and the challenge of patent translation at the time of the first patent translation task at NTCIR. We then describe the innovations at NTCIR, including the sharing of research infrastructure, the progress of corpus-based machine translation technologies, and evaluation methods for patent translation. Finally, we outline the developments in machine translation technologies, including patent translation and remark on the future of patent translation.

7.1 Introduction

Research on machine translation began in the 1950s immediately after the birth of computers. The first machine translation technology was *Rule-Based Machine Translation (RBMT)*, which used manually built translation rules. RBMT was actively developed from the 1970s to the 1980s. In the late 1980s, research began on *Statistical Machine Translation (SMT)*, which is a learning-based machine translation technology based on corpus statistics, (Brown et al. 1993). However, there was little research on SMT for about 10 years. Then the situation changed. From the late 1990s to around 2000, that is, since high-performance computers began to be in widespread use, large parallel corpora became available, automatic evaluation methods, such as BLEU (Papineni et al. 2002), were developed, and research on SMT began to progress rapidly.

The progress of the research was facilitated by evaluation workshops. Evaluation workshops played a dual role in providing large datasets and making evaluations comparable using shared tasks. This made it possible to conduct experiments by

I. Goto (✉)
NHK, 1-10-11 Kinuta, Setagaya-ku Tokyo 157-8510, Japan
e-mail: goto.i-es@nhk.or.jp

T. Sakai et al. (eds.), *Evaluating Information Retrieval and Access Tasks*,
The Information Retrieval Series 43,
https://doi.org/10.1007/978-981-15-5554-1_7

sharing research infrastructure and to verify the effectiveness of methods by performing comparisons using the same data. Evaluation workshops made research more active, and research on machine translation progressed. The following is a list of major evaluation workshops on machine translation that were in existence by the mid-2000s:

- Defense Advanced Research Projects Agency (DARPA) Translingual Information Detection, Extraction, and Summarization (TIDES) project (2001 to 2005): The translation languages were Chinese to English and Arabic to English, and the target domain was news. This project was succeeded by the DARPA Global Autonomous Language Exploitation (GALE) project.
- International Workshop on Spoken Language Translation (IWSLT) (2004 to present (2019)[1]): As of 2004 to 2007, speech translation of travel conversations was targeted. Several languages were included, including Japanese and English. The size of the training parallel corpus was 20,000 to 40,000 sentence pairs.
- Workshop on Statistical Machine Translation (WMT) (2006 to present (2019) see footnote 1): Machine translation between European languages is the target. As of 2006 to 2007, the proceedings of the European Parliament and news were the target domains.

As of 2007, research on SMT was in progress for several language pairs and fields. For the Japanese–English language pair, the domain covered in the evaluation workshops was travel conversations only. Because the sentence lengths were short and the topic was narrow, the shared task for travel conversation translation was technically easy. By contrast, there was no shared task for long sentence translation between Japanese and English, which is useful for advancing translation technology for long sentences between languages that differ significantly in word order.

As a domain that includes long sentence translation between Japanese and English, patent translation has substantial demand, such as translation for foreign applications and translation of patents in foreign languages to understand the content of existing patents. The machine translation of patents has been required by sectors that produce and use intellectual property in countries and many companies. Therefore, if machine translation performs well for patent translation, there will be a substantial impact on society.

In 2007, RBMT systems were on the market for the machine translation of patents between Japanese and English. Through years of research and development, RBMT systems have achieved translation quality at a level that is useful as a rough translation for manual post-editing.[2] However, there was a barrier to further improving the translation quality of RBMT. Simply increasing the number of translation rules did not improve translation quality. Manually adding translation rules so that the appropriate translation rules can be selected in accordance with the context from many

[1]This chapter was written in 2019. Thus, this year does not indicate the final year.

[2]In fact, when the organizers of NTCIR-7 asked Japanese–English translators to produce multi-reference translations of the test sentences, the organizers found that an RBMT system was used for rough translation, and the reference translations had to be retranslated to avoid the bias of a specific machine translation (MT) system (Fujii et al. 2008).

candidates has been a serious challenge that requires craftsmanship. It was also a serious challenge to make sentences generated by combining translation rules into natural sentences as written by a person. Moreover, both the accumulated amount of bilingual patent data and computational power could be expected to increase over time. Thus, to overcome the barriers to RBMT and aim for translation quality at the level of human translation, corpus-based machine translation technology, which automatically acquires translation knowledge and sentence generation knowledge from patent data, was required. However, before 2007, there were few studies on corpus-based machine translation for the patent field.

7.2 Innovations at NTCIR

As explained in the previous section, in 2007, to advance long sentence translation technology between languages differing greatly in word order, it was appropriate timing for shared tasks of patent translation between Japanese and English. At that time, the NTCIR-7 organizers extracted over one million Japanese–English parallel sentence pairs from parallel patent applications and launched the shared task of patent translation. This led to research on corpus-based machine translation for long patent sentences between Japanese and English. Patent translation tasks were conducted four times, from NTCIR-7 to NTCIR-10, over six years (Fujii et al. 2008, 2010; Goto et al. 2011, 2013). In NTCIR-9, the Chinese–English patent translation task was added.

In the following, we present a summary of the comparison between SMT and RBMT for patent translation.

- From the evaluation results of NTCIR-7 in 2008, the translation quality of RBMT was higher than that of SMT for Japanese–English and English–Japanese translation.
- From the evaluation results of NTCIR-9 in 2011, the translation quality of SMT for English–Japanese caught up with that of RBMT. The translation quality of SMT for Chinese–English was higher than that of RBMT because the translation quality of RBMT was low.
- From the evaluation results of NTCIR-10 in 2013, SMT outperformed RBMT for English–Japanese translation. Although SMT could not catch up with RBMT for Japanese–English translation, the top SMT system for Japanese–English translation at NTCIR-10 improved compared with the top SMT system at NTCIR-9.

Thus, through four rounds of shared tasks over 6 years, the performance of SMT substantially improved for patent translation including long sentences for Japanese–English and English–Japanese, and Chinese–English. As a result, corpus-based machine translation could make it possible to overcome the challenges encountered by RBMT. This was the biggest innovation in the patent translation tasks.

In the following, the purpose of each patent translation task is described, and an overview of each of the four tasks, major findings, and innovations is provided. The goals of the patent machine translation tasks were as follows:

- to develop challenging and significant practical research into patent machine translation;
- to investigate the performance of state-of-the-art machine translation in terms of patent translations involving Japanese, English, and Chinese;
- to compare the effects of different methods of patent translation by applying them to the same test data;
- to explore practical MT performance in real scenarios for patent machine translation;
- to create publicly available parallel corpora of patent documents and human evaluation of MT results for patent information processing research;
- to drive machine translation research, which is an important technology for the cross-lingual access to information written in unfamiliar languages; and
- ultimately, to foster scientific cooperation.

7.2.1 Patent Translation Task at NTCIR-7 (2007–2008)

As described in Sect. 7.1, 2007 was a time when SMT technology was progressing. Because there was an open-source SMT tool called Moses (Koehn et al. 2007) at that time, it was easy to conduct experiments on SMT if a bilingual parallel corpus was available. SMT could translate short sentences, such as travel conversations, to some extent. By contrast, the translation quality of SMT was low for long sentences between language pairs with a largely different word order. Therefore, translating a patent document that included long sentences between Japanese and English, which largely differ in word order, was a serious challenge for SMT.

In 2007, the organizers constructed a Japanese–English parallel patent dataset that consisted of approximately 1.8 million parallel sentence pairs and launched the shared tasks of Japanese–English and English–Japanese patent translation. This was the first time that more than one million parallel sentence pairs in Japanese and English became widely available for research. The task organizers extracted the Japanese–English parallel patent sentence pairs from Japanese–English bilingual patent families. A patent family is a set of patents taken in more than one country to protect a single invention. The extraction of parallel sentence pairs was conducted by applying an automatic sentence alignment method (Utiyama and Isahara 2007) to approximately 85,000 patent families from 10 years of Japanese patents published by the Japan Patent Office (JPO) and 10 years of English patents published by the United States Patent and Trademark Office.

In the NTCIR-7 patent translation task, human evaluation was performed. For Japanese–English translation, human evaluation was performed for a total of 15 system outputs that consisted of the 14 system outputs submitted by the participating teams and a system output of the SMT tool Moses used by the organizers. The results showed that the automatic evaluation BLEU-4 score of SMT was higher than that of RBMT; however, in the human evaluation, the results indicated that the actual translation quality of RBMT was better than that of SMT. For English–Japanese

translation, human evaluation was performed for some representative systems, and the results showed that the trend of the comparison between SMT and RBMT was similar to that of Japanese–English translation.

Additionally, the organizers compared the effect when English–Japanese machine translation was used for cross-lingual patent retrieval (CLPR) as an extrinsic evaluation. They used a standard retrieval method for CLPR. Because the standard retrieval method did not use the order of words in queries and documents, the order of words did not affect the retrieval results. The CLPR results were highly correlated with the BLEU score, and SMT was better than RBMT; that is, the results showed that SMT was more effective than RBMT in terms of translation word selection.

7.2.2 Patent Translation Task at NTCIR-8 (2009–2010)

The Japanese–English and English–Japanese patent translation tasks continued. The organizers expanded the size of the bilingual corpus by extracting parallel sentence pairs from 15 years of patent families, and provided the task participants with a Japanese–English parallel corpus that consisted of approximately 3.2 million sentence pairs. In the tasks, no purely RBMT system was included in the evaluation and no human evaluation was performed. Therefore, SMT and RBMT could not be compared.

The system with the highest BLEU score for Japanese–English translation first translated Japanese sentences into English using RBMT, and then post-edited the translation results using SMT (Ehara 2010). The results showed that the word reordering performance of SMT had not caught up with that of RBMT. Additionally, the shared task of the automatic evaluation of machine translation was also conducted using the human evaluation results of NTCIR-7. The task evaluated automatic evaluation methods based on the human evaluation results.

7.2.3 Patent Translation Task at NTCIR-9 (2010–2011)

The organizers[3] added a Chinese–English patent translation task in addition to the Japanese–English and English–Japanese patent translation tasks. Chinese–English translation is a globally required language pair and is popular in the machine translation research community. For the Japanese–English and English–Japanese translation tasks, the training dataset was the same as that of NTCIR-8, that is, approximately 3.2 million sentence pairs, and the test dataset was newly produced. For the Chinese–English translation task, the organizers provided the task participants with a training dataset that consisted of one million parallel sentence pairs of Chinese–English bilingual patents. The organizers produced translation results using com-

[3]The organizers of the patent translation task at NTCIR-9 changed from the organizers at NTCIR-8.

mercial RBMT systems to compare SMT and RBMT. They also performed human evaluation. Twenty-one teams around the world participated in the patent translation tasks. The introduction of the Chinese–English translation task led to the participation of top international teams, such as BBN (Ma and Matsoukas 2011), IBM Watson Research (Lee et al. 2011), and RWTH Aachen University (Feng et al. 2011).

The findings obtained from the evaluation results were as follows: For English–Japanese translation, the top SMT system achieved a translation quality equal to or better than that of the top RBMT system. For the first time in patent translation from English to Japanese, the top SMT system had caught up with the top RBMT system. The top SMT system improved substantially in translation quality by improving word reordering performance using a pre-ordering method (Sudoh et al. 2011). It became clear that separating word reordering from the decoding process could obtain a large effect in a simple manner. For Chinese–English translation, the translation quality of SMT was higher than that of RBMT because the performance of the Chinese–English RBMT systems was low.

The organizers created and applied a new human evaluation criterion, that is, "Acceptability," in addition to "Adequacy," which is a conventional human evaluation criterion. The criteria for each grade of Adequacy were ambiguous, and the actual ratings were compared mainly on a relative basis to distinguish between the systems to be evaluated. Therefore, the translation quality was not necessarily the same for the same grade. For example, grade 3 when only low-level systems were evaluated and grade 3 when only high-level systems were evaluated would be different translation qualities. Thus, it was not possible to know the actual quality using such relatively scored grades. By contrast, Acceptability was defined as an objective and clearer standard, with the aim of making the quality of the same grade constant. The Acceptability results showed that the percentage of translated sentences that could convey all the meanings of the source sentences was 60% for the top systems for both Japanese–English and English–Japanese translation, and the percentage was 80% for the top system for Chinese–English translation.

7.2.4 Patent Translation Task at NTCIR-10 (2012–2013)

The Japanese–English, English–Japanese, and Chinese–English patent translation tasks were continued at NTCIR-10. The training dataset was the same as that at NTCIR-9 and the test dataset was newly produced. Twenty-one teams participated in the tasks.

The findings obtained from the evaluation results were as follows: For English–Japanese translation, the top SMT system (Sudoh et al. 2013) outperformed the RBMT systems in terms of translation quality. For Japanese–English translation, RBMT was still better than SMT; however, the translation quality of the top SMT system had improved from NTCIR-9 (Sudoh et al. 2013). For Chinese–English trans-

lation, the top system used neural networks in a language model to improve performance (Huang et al. 2013), and the effectiveness of neural networks for machine translation was thus demonstrated.

If the test data was simply selected from the automatically extracted parallel corpus, biases, such as lengths or included expressions, may result. To reduce biases, the organizers selected test sentences using two methods. For one method, the organizers first calculated the distribution of sentence lengths in monolingual patent documents in the source language, and divided the cumulative length distribution into quartiles (25% each). Each quartile was called a sentence length class. Next, they classified the automatically aligned sentences in the source language into four classes according to their sentence lengths and extracted the same number of sentences from each class as test sentences. For the other method, the organizers randomly selected test sentences from all the description sentences in the source language patents for bilingual patents. Translators translated the test sentences to produce their reference translations. The data produced by the second method was used for the human evaluation.

At NTCIR-9, the top systems performed well for sentence-level evaluations. Therefore, the NTCIR-10 organizers wanted to see how useful the top systems were for practical scenarios. Patent examination was one of the practical scenarios. The organizers performed Patent Examination Evaluation (PEE), which measures the usefulness of MT systems for patent examinations. PEE is described as follows: PEE assumes that the patent is examined in English. When a patent application in English is filed, an examiner examines existing patents and rejects the patent application if almost identical technology is described in an existing patent. If a patent application is rejected by referencing an existing patent, the examiner writes the final decision document (*Shinketsu*), which describes the facts about the existing patent on which the rejection is based. Assuming that the referenced patents were written in a foreign language, the organizers extracted the part that described the facts from the referenced patents and used the extracted sentences as test data. The test data in foreign languages (Japanese/Chinese) were translated into English using machine translation, and the translation results were evaluated according to whether the facts that were used to reject patent applications could be recognized from the translation result. PEE was performed by two experienced patent examiners. For Japanese–English translation, for the best system, all facts were recognized in 66% of referenced patents, and at least half of the facts were recognized in 100% of referenced patents. For Chinese–English translation, for the best system, all facts were recognized in 20% of referenced patents, and at least half of the facts were recognized in 88% of referenced patents. PEE achieved the evaluation of usefulness in one representative practical scenario of patent machine translation. The PEE results and translations can be used as standards of usefulness in patent examination. Specifically, by comparing new translation results for the PEE test data with the PEE evaluated translations at NTCIR-10, their usefulness in patent examination for other systems can be assessed roughly.

7.3 Developments After NTCIR-10

The evaluation workshop on Asian translation (WAT) for machine translation was launched in 2014. WAT targets machine translation between language pairs that include Asian languages. The activities of WAT have promoted the construction and sharing of research infrastructure for machine translation involving Asian languages. WAT features an open innovation platform. The test data and reference translations have been published with the training data, and the use of the same test data every year facilitates comparisons. In the following, we describe the activities and findings of WAT.

In the first workshop (WAT 2014) (Nakazawa et al. 2014), the organizers set the shared tasks of scientific paper translation between Japanese and English, and between Japanese and Chinese. An SMT system using syntactic structures achieved the highest performance.

In the second workshop (WAT 2015) (Nakazawa et al. 2015), in addition to the scientific paper translation tasks, Chinese–Japanese and Korean–Japanese patent translation tasks were included. The size of the training dataset for each patent translation task was one million sentence pairs. The results showed that the translation quality of the top SMT system was higher than that of the RBMT systems for patent translation for Chinese–Japanese and Korean–Japanese. For scientific paper translation, a reranking method using *Neural Machine Translation (NMT)* achieved the highest translation quality. The effectiveness of the scoring by NMT was thus demonstrated.

In the third workshop (WAT 2016) (Nakazawa et al. 2016), Japanese–English and English–Japanese patent translation tasks were added. The size of the training dataset for each patent translation task was one million sentence pairs. For Japanese–English patent translation, the results confirmed that the translation quality of NMT and SMT outperformed the translation quality of RBMT. This was the first time that a corpus-based machine translation system yielded Japanese–English patent translation results comparable with those of RBMT systems. The translation quality of NMT evaluated by humans was higher than that of SMT for Japanese–English patent translation. For Japanese–English and English–Japanese scientific paper translation, pure NMT systems, not SMT reranking, achieved the best performance. In the field of machine translation, where large-scale parallel data was available, the mainstream technology for machine translation was changed from SMT to NMT. For English–Japanese patent translation, NMT achieved a translation quality close to that of the top SMT.

In the fourth workshop (WAT 2017) (Nakazawa et al. 2017), news translation tasks between Japanese and English and recipe translation tasks between Japanese and English were added. In Japanese–English patent translation, the results showed that 86% of translated sentences conveyed all the meanings of the source sentences for the top NMT system, which was trained using ten million parallel sentence pairs in addition to the shared task data of one million parallel sentence pairs. By contrast, for Japanese–English news translation, 5% of translated sentences conveyed all the meanings for the top NMT system. This percentage is substantially lower than that of the top system for Japanese–English patent translation. The small size of the

training data was one of the reasons. An essential reason was that the quality of the parallel translation of news was lower than that of patents. The reason for the low quality of parallel translation of news compared with that of patents is as follows: In patent applications, because the content in Japanese is translated literally to make an English version of the patent to file as a patent family, the translation quality at the sentence level is high. By contrast, news translation is not only translation but news writing. In news writing, writers select the content in consideration of the difference between readers of news in the source language and readers of news in the target language, and writers edit articles to change the structure to that of an English news structure. Thus, even if the sentences are aligned in same-topic bilingual news articles in Japanese and English, the parallel translation quality at the sentence level is lower than that of patents. It was shown that the translation of news with low-quality parallel data was a challenge for machine translation. Additionally, in the Chinese–Japanese patent translation task, 62% of translated sentences conveyed all the meanings of the source sentences. The performance improved from 29% in the previous year. Chinese–Japanese patent translation is in high demand in Japan.

In the fifth workshop (WAT 2018) (Nakazawa et al. 2018), the translation tasks between Myanmar and English, and between seven Indic languages and English were added. For Japanese–English scientific paper translation, the percentage of translated sentences that conveyed all the meanings of the source sentences improved from 34% in WAT 2017 to 61% in WAT 2018.

We have outlined research trends in machine translation, including patent translation from the activities of WAT. In the following, we describe other events. Google Translate changed from SMT to NMT in 2016. The change to NMT improved the translation quality, and people recognized the effectiveness of NMT. As a global trend, artificial intelligence (AI) technologies using deep learning have attracted attention since 2012. NMT is an AI technology. NMT's translation quality first caught up with SMT's translation quality in 2014, and NMT's translation quality has improved each year. There were very rapid advances in translation quality in the four years from 2015 to 2018.

Finally, we discuss the future of patent translation. Patent translation is an area in which large-scale high-quality parallel corpora are available. For example, a parallel corpus exists that contains over 100 million sentences.[4] Although machine translation is not perfect, the translation quality of NMT will become close to translators for sentences without low-frequency words or new words as a result of training using a parallel corpus with the scale of 100 million sentence pairs. Because patent claims in Japanese have special styles, special pre-processing is necessary. The translation of sentences in claim sections is expected to be of high quality in the future. However, the translation of low-frequency words and new words is a problem that is difficult to solve using a corpus-based mechanism alone, and another approach will be necessary. Methods that use subword units, such as byte pair encoding (Sennrich et al. 2016), alleviate this problem. However, the translation of low-frequency words whose elements are not compositional and low-frequency subwords is still a problem.

[4] ALAGIN JPO corpus https://alaginrc.nict.go.jp/resources/jpo-info/jpo-outline.html.

There have been some studies on using automatically discovered bilingual words, and such techniques might be applied to NMT. Although machine translation may make errors, machine translation can do many things. Machine translation can be used for new translation needs that take advantage of its low cost and high speed. The patent offices of several countries, such as JPO, have already incorporated machine translation into their work. Machine translation has also been used in commercial services that provide foreign language patents in their customers' preferred language. Machine translation of patents will be used in society as an indispensable tool to overcome the language barrier in intellectual property.

Acknowledgements The author would like to thank Prof. Mikio Yamamoto for his support in writing this manuscript, particularly the chapters on NTCIR-7 and NTCIR-8. The author also thanks Prof. Doug Oard for his helpful comments. The author thanks Maxine Garcia, PhD, from Edanz Group for editing a draft of this manuscript.

References

Brown PE, Della Pietra SA, Della Pietra VJ, Mercer RL (1993) The mathematics of statistical machine translation: parameter estimation. Comput Linguist 19(2):263–311

Ehara T (2010) Machine translation for patent documents combining rule-based translation and statistical post-editing. In: Proceedings of the NTCIR-8 workshop, pp 384–386

Feng M, Schmidt C, Wuebker J, Peitz S, Freitag M, Ney H (2011) The RWTH Aachen system for NTCIR-9 PatentMT. In: Proceedings of the NTCIR-9 workshop, pp 600–605

Fujii A, Utiyama M, Yamamoto M, Utsuro T (2008) Overview of the patent translation task at the NTCIR-7 workshop. In: Proceedings of the NTCIR-7 workshop, pp 389–400

Fujii A, Utiyama M, Yamamoto M, Utsuro T, Ehara T, Echizen-ya H, Shimohata S (2010) Overview of the patent translation task at the NTCIR-8 workshop. In: Proceedings of the NTCIR-8 workshop, pp 371–376

Goto I, Lu B, Chow KP, Sumita E, Tsou BK (2011) Overview of the patent machine translation task at the NTCIR-9 workshop. In: Proceedings of the NTCIR-9 workshop, pp 559–578

Goto I, Chow KP, Lu B, Sumita E, Tsou BK (2013) Overview of the patent machine translation task at the NTCIR-10 workshop. In: Proceedings of the NTCIR-10 workshop, pp 260–286

Huang Z, Devlin J, Matsoukas S (2013) BBN's systems for the Chinese-English sub-task of the NTCIR-10 PatentMT. In: Proceedings of the NTCIR-10 workshop, pp 287–293

Koehn P, Hoang H, Birch A, Callison-Burch C, Federico M, Bertoldi N, Cowan B, Shen W, Moran C, Zens R, Dyer C, Bojar O, Constantin A, Herbst E (2007) Moses: Open source toolkit for statistical machine translation. In: Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions, association for computational linguistics, Prague, Czech Republic, pp 177–180

Lee YS, Xiang B, Zhao B, Franz M, Roukos S, Al-Onaizan Y (2011) IBM Chinese-to-English PatentMT system for NTCIR-9. In: Proceedings of the NTCIR-9 workshop, pp 606–613

Ma J, Matsoukas S (2011) BBN's systems for the Chinese-English sub-task of the NTCIR-9 PatentMT evaluation. In: Proceedings of the NTCIR-9 workshop, pp 579–584

Nakazawa T, Mino H, Goto I, Kurohashi S, Sumita E (2014) Overview of the 1st workshop on Asian translation. In: Proceedings of the 1st workshop on Asian translation (WAT2014), workshop on Asian translation, Tokyo, Japan, pp 1–19. https://www.aclweb.org/anthology/W14-7001

Nakazawa T, Mino H, Goto I, Neubig G, Kurohashi S, Sumita E (2015) Overview of the 2nd workshop on Asian translation. In: Proceedings of the 2nd workshop on Asian translation (WAT2015), workshop on Asian translation, Kyoto, Japan, pp 1–28. https://www.aclweb.org/anthology/W15-5001

Nakazawa T, Ding C, Mino H, Goto I, Neubig G, Kurohashi S (2016) Overview of the 3rd workshop on Asian translation. In: Proceedings of the 3rd workshop on Asian translation (WAT2016), the COLING 2016 organizing committee, Osaka, Japan, pp 1–46. https://www.aclweb.org/anthology/W16-4601

Nakazawa T, Higashiyama S, Ding C, Mino H, Goto I, Kazawa H, Oda Y, Neubig G, Kurohashi S (2017) Overview of the 4th workshop on Asian translation. In: Proceedings of the 4th workshop on Asian translation (WAT2017), Asian federation of natural language processing, Taipei, Taiwan, pp 1–54. https://www.aclweb.org/anthology/W17-5701

Nakazawa T, Sudoh K, Higashiyama S, Ding C, Dabre R, Mino H, Goto I, Pa WP, Kunchukuttan A, Kurohashi S (2018) Overview of the 5th workshop on Asian translation. In: Proceedings of the 32nd Pacific Asia conference on language, information and computation: 5th workshop on Asian translation: 5th workshop on Asian translation, association for computational linguistics, Hong Kong. https://www.aclweb.org/anthology/Y18-3001

Papineni K, Roukos S, Ward T, Zhu WJ (2002) Bleu: a method for automatic evaluation of machine translation. In: Proceedings of 40th annual meeting of the association for computational linguistics, pp 311–318

Sennrich R, Haddow B, Birch A (2016) Neural machine translation of rare words with subword units. In: Proceedings of the 54th annual meeting of the association for computational linguistics (Vol 1: Long Papers), association for computational linguistics, Berlin, Germany, pp 1715–1725. https://doi.org/10.18653/v1/P16-1162, https://www.aclweb.org/anthology/P16-1162

Sudoh K, Duh K, Tsukada H, Nagata M, Wu X, Matsuzaki T, Tsujii J (2011) NTT-UT statistical machine translation in NTCIR-9 PatentMT. In: Proceedings of the NTCIR-9 workshop, pp 585–592

Sudoh K, Suzuki J, Tsukada H, Nagata M, Hoshino S, Miyao Y (2013) NTT-NII statistical machine translation for NTCIR-10 PatentMT. In: Proceedings of the NTCIR-10 workshop, pp 294–300

Utiyama M, Isahara H (2007) A Japanese-English patent parallel corpus. In: Proceedings of the machine translation summit XI, pp 475–482

Chapter 8
Component-Based Evaluation for Question Answering

Teruko Mitamura and Eric Nyberg

Abstract This chapter describes the component-based evaluation of automatic question answering (QA) systems, which was pioneered in the NTCIR-7 ACLIA challenge and has became a fundamental part of QA system development, especially for difficult real-world datasets which require a multi-strategy, multi-component approach. We summarize the history of component evaluation for QA and describe more recent work at Carnegie Mellon (on TREC Genomics, BioASQ, and LiveQA datasets) which has descended directly from our experiences in NTCIR.

8.1 Introduction

In this chapter, we first describe the component-based evaluations for question answering that were developed as part of past NTCIR challenges. We introduce the CMU JAVELIN Cross-lingual Question Answering (CLQA) system and show how the JAVELIN architecture supports component-level evaluation, which can accelerate overall system development. This component-based evaluation concept was used in the NTCIR-7 ACLIA tasks, not only to evaluate each component but also to evaluate different combinations of Information Retrieval (IR) and Question Answering (QA) modules.

In later sections, we describe more recent developments in component-based evaluation within the Open Advancement of Question Answering (OAQA) and Configuration Space Exploration (CSE) projects. We also describe automatic component evaluation for biomedical QA systems. All of these later developments were influenced by the original vision of component-based evaluation embodied in the NTCIR QA tasks. To conclude, we discuss remaining challenges and future directions for component-based evaluation in QA.

T. Mitamura (✉) · E. Nyberg
Carnegie Mellon University, Pittsburgh, PA, USA
e-mail: teruko@andrew.cmu.edu

E. Nyberg
e-mail: ehn@cs.cmu.edu

T. Sakai et al. (eds.), *Evaluating Information Retrieval and Access Tasks*,
The Information Retrieval Series 43,
https://doi.org/10.1007/978-981-15-5554-1_8

8.1.1 History of Component-Based Evaluation in QA

The JAVELIN Cross Language Question Answering (CLQA) system, developed by the Language Technologies Instutute (LTI) at Carnegie Mellon University (CMU) had five main components: question analysis, keyword translation, document retrieval, information extraction, and answer generation (Mitamura et al. 2007). This system contains an English-to-Japanese QA system and an English-to-Chinese QA system with the same overall architecture, which supported direct comparison of the two systems on a per-module basis. After analyzing the observed performance of each module on the evaluation data, we created gold-standard data (perfect input) for each module in order to determine upper bounds on module performance. The overall architecture is shown in Fig. 8.1.

The Question Analysis (QA) module is responsible for parsing the input question, choosing the appropriate answer type, and producing a set of keywords. The Translation Module (TM) translates the keywords into task-specific languages. The Retrieval Strategist (RS) module is responsible for finding relevant documents which might contain answers to the question, using translated keywords produced by the Translation Module. The Information Extractor (IX) module extracts answers from the relevant documents. The Answer Generation (AG) module normalizes the answers and ranks them in order of correctness.

Although traditional QA systems consist of several modules with a cascaded approach, as far as we know the JAVELIN CLQA system was the first one to incorporate component-based evaluation for QA. We participated in the NTCIR-5 CLQA1 task and demonstrated our results (Lin et al. 2005). A more detailed analysis of our component-based evaluation was presented at LREC 2006 (Shima et al. 2006).

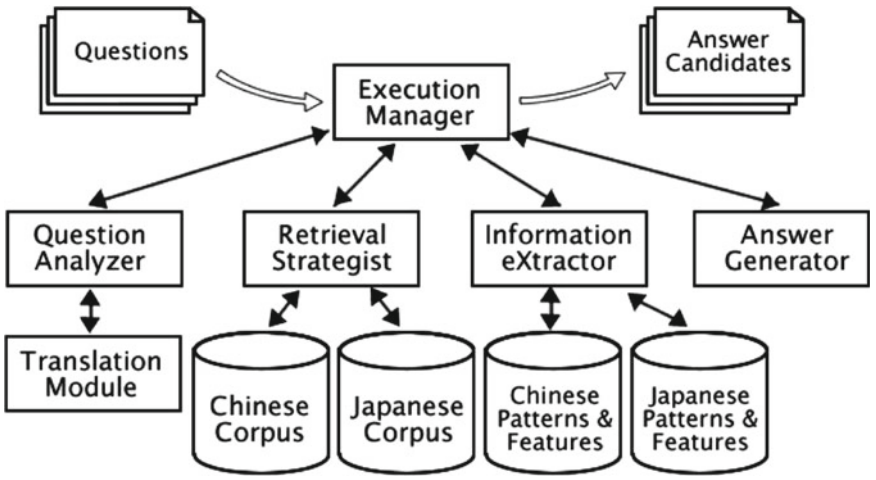


Fig. 8.1 JAVELIN architecture

8.1.2 Contributions of NTCIR

NTCIR first included a question answering challenge (QAC) evaluation for Japanese in 2002 (NTCIR-3). The NTCIR-4 and the NTCIR-5 challenges continued to include QAC tasks in 2004 and 2005 respectively. The NTCIR-5 challenge also added the first cross-lingual QA task, which contained five subtasks for three languages: English, Japanese, and Chinese. The JAVELIN system was evaluated on the CLQA tasks for all three languages. When developing cross-lingual capabilities with three languages, system and component development became more complicated, and error analysis became very challenging. Therefore, we developed a component-based evaluation approach for error analysis and improvement of the JAVELIN CLQA system (Lin et al. 2005; Shima et al. 2006).

Input questions in English are processed by these modules in the order listed above. The answer candidates are returned in one of the two target languages (Japanese and Chinese) as final outputs. The QA module is responsible for parsing the input question, choosing the expected answer type, and producing a set of keywords. The QA module calls the Translation Module, which translates the keywords into the language(s) required by the task.

In order to gain different perspectives on the tasks and our system's performance, a module-by-module analysis was performed. We used the formal run dataset from NTCIR task CLQA1, which includes English–Chinese (EC) and English–Japanese (EJ) subtasks. 200 input questions were provided for each of the subtasks. This analysis was based on gold-standard answer data, which also provides information about the documents that contain the correct answer for each question. We judged the QA module by the accuracy of its answer type classification, and the Translation Module by the accuracy of its keyword translation. For the RS and IX modules, if a correct document or answer is returned, regardless of its ranking, we consider the module to be successful. To separate the effects of errors introduced by earlier modules, we created gold-standard data by manually correcting answer type and keyword translation errors. We also create "perfect" IX input using the gold-standard document set. In Table 8.1, the overall performance (top 1 average accuracy) is shown in the last two columns of the top rows for EC and EJ. The symbol "R" indicates recall versus the standard gold answer set; the symbol "R+U" indicates recall versus the standard gold answer set plus other (unofficial) correct answers ("Unsupported"). If we examine only such global measures, we will not be able to understand the performance of individual modules in a complex system.

Our analysis of per-module performance from gold-standard input shows that the QA module and the RS module are already performing fairly well, but there is still room in the IX module and the AG module for future improvement.

Table 8.1 Modular performance analysis (Shima et al. 2006)

	Gold standard input	AType accuracy (%)	TM accuracy (%)	RS top 15 (%)	IX top 100 (%)	MRR	Overall top 1 R (%)	Top 1 R+U (%)
EC	None	86.5	69.3	30.5	30.0	0.130	7.5	9.5
EC	TM	86.5	–	57.5	50.0	0.254	9.5	20.0
EC	TM+AType	–	–	57.5	50.5	0.260	9.5	20.5
EC	TM+AType+RS	–	–	–	63.0	0.489	41.0	43.0
EJ	None	93.5	72.6	44.5	31.5	0.116	10.0	12.5
EJ	TM	93.5	–	67.0	41.5	0.154	9.5	15.0
EJ	TM+AType	–	–	68.0	45.0	0.164	10.0	15.5
EJ	TM+AType+RS	–	–	–	51.5	0.381	32.0	32.5

8.2 Component-Based Evaluation in NTCIR

In 2007, LTI/CMU became an organizer of Advanced Cross-lingual Information Access (ACLIA) task for NTCIR-7. In this task, we started the formal component-based evaluation for Japanese (JA), Simplified Chinese (CS), Traditional Chinese (CT), and English for the first time (Mitamura et al. 2008). There were two major tasks: (1) Information Retrieval for Question Answering (IR4QA) and (2) Complex Cross-Lingual Question Answering (CCLQA) tasks. Within the CCLQA task, we had three subtasks: Question Analysis track, CCLQA Main Track, and IR4QA+CCLQA collaboration tracks (obligatory track and optional track). The ACLIA task data flow is illustrated in Fig. 8.2.

As a central problem in question answering evaluation, the lack of standardization made it difficult to compare systems under a shared condition. In NLP research at that time, system design was moving away from monolithic, black-box architectures and more toward modular, architectural approaches that include an algorithm-independent formulation of the system's data structures and data flows, so that multiple algorithms implementing a particular function can be evaluated on the same task. Therefore, the ACLIA data flow includes a pre-defined schema for representing the inputs and outputs of the document retrieval step, as illustrated in Fig. 8.2. This novel standardization effort made it possible to evaluate IR4QA (Information Retrieval for Question Answering) in the context of a closely related QA task. During the evaluation, the question text and QA system question analysis results were provided as input to the IR4QA task, which produced retrieval results that were subsequently fed back into the end-to-end QA systems. The modular design and XML interchange format supported by the ACLIA architecture made it possible to perform such embedded evaluations in a straightforward manner.

The modular design of this evaluation data flow is motivated by the following goals: (a) to make it possible for participants to contribute component algorithms to an evaluation, even if they cannot field an end-to-end system; (b) to make it possible to conduct evaluations on a per-module basis, in order to target metrics and error

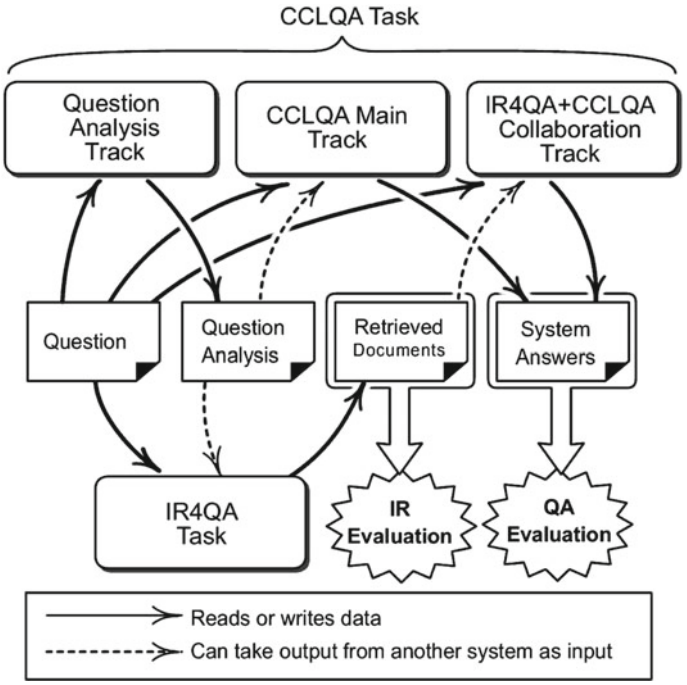


Fig. 8.2 Data flow in ACLIA task cluster showing how interchangeable data model made inter-system and inter-task collaboration possible (Mitamura et al. 2008)

analysis on important bottlenecks in the end-to-end system; and (c) to determine which combination of algorithms works best by combining the results from various modules built by different participants.

8.2.1 Shared Data Schema and Tracks

In order to combine a Cross-Lingual Information Retrieval (CLIR) module with a cross-lingual Question Answering (CLQA) system for module-based evaluation, we defined five types of XML schema to support exchange of results among participants and submission of results to be evaluated:

- **Topic format**: The organizer distributes topics in this format for formal run input to IR4QA and CCLQA systems.
- **Question Analysis format**: CCLQA participants who chose to share Question Analysis results submit their data in this format. IR4QA participants can accept task input in this format.
- **IR4QA submission format**: IR4QA participants submit results in this format.

- **CCLQA submission format:** CCLQA participants submit results in this format.
- **Gold-Standard Format**: Organizer distributes CCLQA gold-standard data in this format.

Participants in the ACLIA CCLQA task submitted results for the following four tracks:

- **Question Analysis Track**: Question Analysis results contain key terms and answer types extracted from the input question. These data are submitted by CCLQA participants and released to IR4QA participants.
- **CCLQA Main Track**: For each topic, a system returned a list of system responses (i.e., answers to the question), and human assessors evaluated them. Participants submitted a maximum of three runs for each language pair.
- **IR4QA+CCLQA Collaboration Track (obligatory)**: Using possibly relevant documents retrieved by the IR4QA participants, a CCLQA system-generated QA results in the same format used in the main track. Since we encouraged participants to compare multiple IR4QA results, we did not restrict the maximum number of collaboration runs submitted and used automatic measures to evaluate the results. In the obligatory collaboration track, only the top 50 documents returned by each IR4QA system for each question were utilized.
- **IR4QA+CCLQA Collaboration Track (optional)**: This collaboration track was identical to the obligatory collaboration track, except that participants were able to use the full list of IR4QA results available for each question (up to 1000 documents per-topic).

8.2.2 Shared Evaluation Metrics and Process

In order to build an answer key for evaluation, third party assessors created a set of weighted nuggets for each topic. A "nugget" is defined as the minimum unit of correct information that satisfies the information need.

In this section, we present the evaluation framework used in ACLIA, which is based on weighted nuggets. Both human-in-the-loop evaluation and automatic evaluation were conducted using the same topics and metrics. The primary difference is in the step where nuggets in system responses are matched with gold-standard nuggets. During human assessment, this step is performed manually by human assessors, who judge whether each system response nugget matches a gold-standard nugget. In automatic evaluation, this decision is made automatically. The subsections that follow, we detail the differences between these two types of evaluation.

8.2.2.1 Human-in-the-loop Evaluation Metrics

In CCLQA, we evaluate how well a QA system can return answers that satisfy information needs on average, given a set of natural language questions. We adopted the

nugget pyramid evaluation method (Lin and Demner-Fushman 2006) for evaluating CCLQA results, which requires only that human assessors make a binary decision whether a system response matches a gold-standard "vital" nugget (necessary for the answer to be correct) or "ok" nugget (not necessary, but not incorrect). This method was used in the TREC 2005 QA track for evaluating definition questions, and in the TREC 2006–2007 QA tracks for evaluating "other" questions. We evaluated each submitted run by calculating the macroaverage F-score over all questions in the formal run dataset.

In the TREC evaluations, a character allowance parameter C is set to 100 non-whitespace characters for English (Voorhees 2003). Based on the micro-average character length of the nuggets in the formal run dataset, we derived settings of C = 18 for CS, C = 27 for CT and C = 24 for JA.

Note that precision is an approximation, imposing a simple length penalty on the System Response (SR). This is due to Voorhees' observation that "nugget precision is much more difficult to compute since there is no effective way of enumerating all the concepts in a response" (Voorhees 2004). The precision is a length-based approximation with a value of 1 as long as the total system response length per question is less than the allowance, i.e., C times the number of nuggets defined for a topic. If the total length exceeds the allowance, the score is penalized. Therefore, although there is no limit on the number of SRs submitted for a question, a long list of SRs harms the final F-score.

The F ($\beta = 3$) or simply F3 score has emphasizes recall over precision, with the β value of 3 indicating that recall is weighted three times as much as precision. Historically, a β of 5 was suggested by a pilot study on definitional QA evaluation (Voorhees 2003). In the later TREC QA tasks, the value has been to 3.

8.2.2.2 Automatic Evaluation Metrics

ACLIA also utilized automatic evaluation metrics for evaluating the large number of IR4QA+CCLQA Collaboration track runs. Automatic evaluation is also useful during developing, where it provides rapid feedback on algorithmic variations under test. The main goal of research in automatic evaluation is to devise an automatic metric for scoring that correlates well with human judgment. The key technical requirement for automatic evaluation of complex QA is a real-valued matching function that provides a high score to system responses that match a gold-standard answer nugget, with a high degree of correlation with human judgments on the same task.

The simplest nugget matching procedure is exact match of the nugget text within the text of the system response. Although exact string match (or matching with simple regular expressions) works well for automatic evaluation of factoid QA, this model does not work well for complex QA, since nuggets are not exact texts extracted from the corpus text; the matching between nuggets and system responses requires a degree of understanding that cannot be approximated by a string or regular expression match for all acceptable system responses, even for a single corpus.

$$a_{BINARIZED} = \sum_{n \in Nuggets} \max_{s \in SRs} \mathrm{I}_{\theta}(n, s)$$

$$\mathrm{I}_{\theta}(n, s) = \begin{cases} 1 : \mathrm{NuggetRecall}_{\mathrm{token}}(n, s) > \theta \\ 0 : \mathrm{otherwise} \end{cases}$$

Fig. 8.3 Formulas of the binarized metric used for official ACLIA automatic evaluation (Mitamura et al. 2008)

For the evaluation of complex questions in the TREC QA track, Lin and Demner-Fushman (2006) devised an automatic evaluation metric called POURPRE. Since the TREC target language was English, the evaluation procedure simply tokenized answer texts into individual words as the smallest units of meaning for token matching. In contrast, the ACLIA evaluation metric tokenized Japanese and Chinese texts into character unigrams. We did not extract word-based unigrams since automatic segmentation of CS, CT, and JA texts is non-trivial; these languages lack white space and there are no general rules for comprehensive word segmentation. Since a single character in these languages can bear a distinct unit of meaning, we chose to segment texts into character unigrams, a strategy that has been followed for other NLP tasks in Asian languages (e.g., Named Entity Recognition Asahara and Matsumoto 2003). One of the disadvantages of POURPRE is that it gives a partial score to a system response if it has at least one common token with any one of the nuggets. To avoid over-estimating the score via aggregation of many such partial scores, we devised a novel metric by mapping the POURPRE soft match score values into binary values (see Fig. 8.3). We set the threshold θ to be somewhere in between no match and an exact match, i.e., 0.5, and we used this BINARIZED metric as our official automatic evaluation metric for ACLIA.

Reliability of Automatic Evaluation: We compared per-run (# of data points = # of human evaluated runs for all languages) and per-topic (# of data points = # of human evaluated runs for all languages times # of topics) correlation between scores from human-in-the-loop evaluation and automatic evaluation. The following Table 8.2 from the ACLIA Overview (Mitamura et al. 2008) shows that the correlation between the automatic and human evaluation metrics.

The Pearson measure indicates the correlation between individual scores, while the Kendall measure indicates the rank correlation between sets of data points. The results show that our novel nugget matching algorithm BINARIZED outperformed SOFTMATCH for both correlation measures, and we chose BINARIZED as the official automatic evaluation metric for the CCLQA task.

Table 8.2 Per-run and per-topic correlation between automatic nugget matching and human judgment (Mitamura et al. 2008)

Algorithm	Token	Per-run (N = 40) Pearson	Per-run (N = 40) Kendall	Per-topic (N = 40 × 100) Pearson	Per-topic (N = 40 × 100) Kendall
Exactmatch	Char	0.4490	0.2364	0.5272	0.4054
Softmatch	Char	0.6300	0.3479	0.6383	0.4230
Binarized	Char	0.7382	0.4506	0.6758	0.5228

8.3 Recent Developments in Component Evaluation

The introduction of modular QA design and component-based QA evaluation by NTCIR had a strong influence on subsequent research in applied QA systems. In this section, we summarize key developments in QA research that followed directly from our experiences with NTCIR.

8.3.1 Open Advancement of Question Answering

Shared modular APIs and common data exchange formats have become fundamental requirements for general language processing frameworks like UIMA (Ferrucci et al. 2009a) and specific language applications (like the Jeopardy! Challenge) (Ferrucci et al. 2010). In 2009, a group of academic and industry researchers published a technical report on the fundamental requirements for the Open Advancement of Question Answering (OAQA) (Ferrucci et al. 2009b); chief among these requirements are the shared modular design, common data formats, and automatic evaluation metrics first introduced by NTCIR:

> To support this vision of shared modules, dataflows, and evaluation measures, an open collaboration will include a shared logical architecture—a formal API definition for the processing modules in the QA system, and the data objects passed between them. For any given configuration of components, standardized metrics can be applied to the outputs of each module and the end-to-end system to automatically capture system performance at the micro and macro level for each test or evaluation. (Ferrucci et al. 2009b)

> By designing and building a shared infrastructure for system integration and evaluation, we can reduce the cost of interoperation and accelerate the pace of innovation. A shared logical architecture also reduces the overall cost to deploy distributed parallel computing models to reduce research cycle time and improve run-time response. (Ferrucci et al. 2009b)

A group of eight universities followed these principles in collaborating with IBM Research to develop the Watson system for the Jeopardy! challenge (Andrews 2011). The Watson system utilized a shared, modular architecture which allowed the exploration of many different implementations of question-answering components. In

particular, hundreds of components were evaluated, as part of an answer-scoring ensemble that was used to select Watson's final answer for each clue (Ferrucci et al. 2010).

Following the success of the Watson system in the Jeopardy! Challenge (where the system won a tournament against two human champions, Ken Jennings and Brad Rutter), Carnegie Mellon continued to refine the OAQA approach and engaged with other industrial sponsors (most notably, Hoffman-Laroche) to develop open-source architectures and solutions for question answering (discussed below).

8.3.2 Configuration Space Exploration (CSE)

In January of 2012, Carnegie Mellon launched a new project on biomedical question answering, with support from Hoffman-Laroche. Given the goal of building a state-of-the-art QA system for a current dataset (at that time, the TREC Genomics dataset), the CMU team chose to survey and evaluate published approaches (at the level of architecture and modules) to determine the best baseline solution. This triggered a new emphasis on defining and exploring a space of possible end-to-end pipelines and module combinations, rather than selecting and optimizing a single architecture based on preference, convenience, etc. The Configuration Space Exploration project (Garduño et al. 2013) explored the following research questions (taken from Yang et al. 2013):

- How can we formally define a configuration space to capture the various ways of configuring resources, components, and parameter values to produce a working solution? Can we give a formal characterization of the problem of finding an optimal configuration from a given configuration space?
- Is it possible to develop task-independent open-source software that can easily create a standard task framework and incorporate existing tools and efficiently explore a configuration space using distributed computing?
- Given a real-world information processing task, e.g., biomedical question answering, and a set of available resources, algorithms, and toolkits, is it possible to write a descriptor for the configuration space, and then find an optimal configuration in that space using the CSE framework?

The CSE concept of operations is shown in Fig. 8.4. Given a labeled set of input–output pairs (the *information processing task*), the system searches a space of possible solutions (algorithms, toolkits, knowledge bases, etc.) using a set of standard benchmarks (metrics) to determine which solution(s) have the best performance over all the inputs in the task. The goal of CSE is to find an optimal or near-optimal solution while exploring (formally evaluating) only a smart part of the total configuration space.

Based on a shared component architecture and implemented in UIMA, the Configuration Space Exploration (CSE) project was the first to automatically choose an optimal configuration from a set of QA modules and associated parameter values,

given a set of labeled training instances (Garduño et al. 2013). As part of his Ph.D. thesis at Carnegie Mellon, Zi Yang applied the CSE framework to several biomedical information processing problems (Yang 2017). In the following subsection, we discuss the main results of component evaluation for biomedical QA systems.

8.3.3 Component Evaluation for Biomedical QA

Using the Configuration Space Exploration techniques described in the previous subsection (Garduño et al. 2013), a group of researchers at CMU were able to automatically identify a system configuration which signficantly outperformed published baselines for the TREC Genomics task (Yang et al. 2013). Subsequent work showed that it was possible to build high-performance QA systems by applying this optimization approach to an ensemble of subsystems, for the related set of tasks in the BioASQ challenge (Yang et al. 2015).

Table 8.3 shows a summary of the different components that were evaluated for the TREC genomics task: various tokenizers, part-of-speech taggers, named entity recognizers, biomedical knowledge bases, retrieval tools, and reranking algorithms. As shown in Fig. 8.4, the team evaluated about 2,700 different end-to-end configurations, executing over 190 K test examples in order to select the best-performing configuration (Table 8.4). After 24 hours of clock time, the system (running on 30 compute nodes) was able to find a configuration that significantly outperformed the published state of the art on the 2006 TREC Genomics task, achieving a document MAP of 0.56 (versus a published best of 0.54) and a passage MAP of 0.18 (versus a published best of 0.15). Table 8.5 shows the analogous results for the 2007 TREC

Table 8.3 Summary of components integrated for TREC Genomics. (Yang et al. 2013)

Category	Components
NLP tools	LingPipe HMM-based tokenizer
	LingPipe HMM-based POS tagger
	LingPipe HMM-based named entity recognizer
	Rule-based lexical variant generator
KBs	UMLS for syn/acronym expansion
	EntrezGene for syn/acronym expansion
	MeSH for syn/acronym expansion
Retrieval tools	Indri system
Reranking algorithms	Important sentence identification
	Term proximity-based ranking
	Score combination of different retrieval units
	Overlapping passage resolution

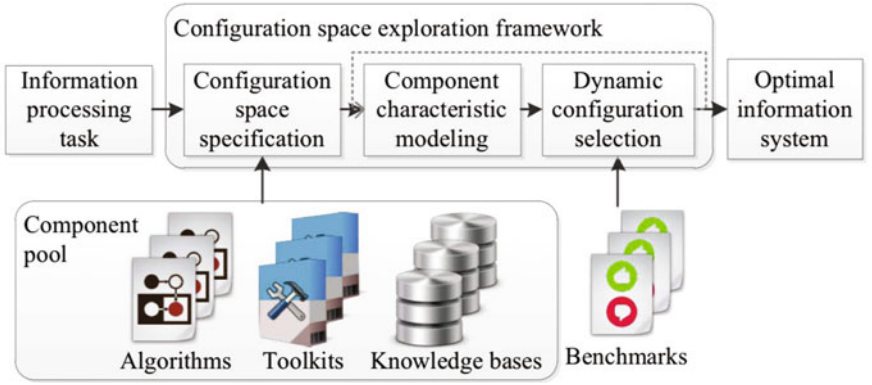


Fig. 8.4 Overview of configuration space exploration framework architecture (Yang et al. 2013)

Table 8.4 Performance of automatically configured components (CSE) versus TREC Genomics 2006 participants (Yang et al. 2013)

	TREC 2006	CSE
No. components	1,000	12
No. configurations	1,000	32
No. traces	92	2,700
No. executions	1,000	190,680
Capacity (hours)	N/A	24
DocMAP max	0.5439	0.5648
DocMAP median	0.3083	0.4770
DocMAP min	0.0198	0.1087
PsgMAP max	0.1486	0.1773
PsgMAP median	0.0345	0.1603
PsgMAP min	0.0007	0.0311

Table 8.5 Performance of automatically configured components versus TREC Genomics 2007 participants (Yang et al. 2013)

	TREC 2007	CSE
DocMAP max	0.3286	0.3144
DocMAP median	0.1897	0.2480
DocMAP min	0.0329	0.2067
PsgMAP max	0.0976	0.0984
PsgMAP median	0.0565	0.0763
PsgMAP min	0.0029	0.0412

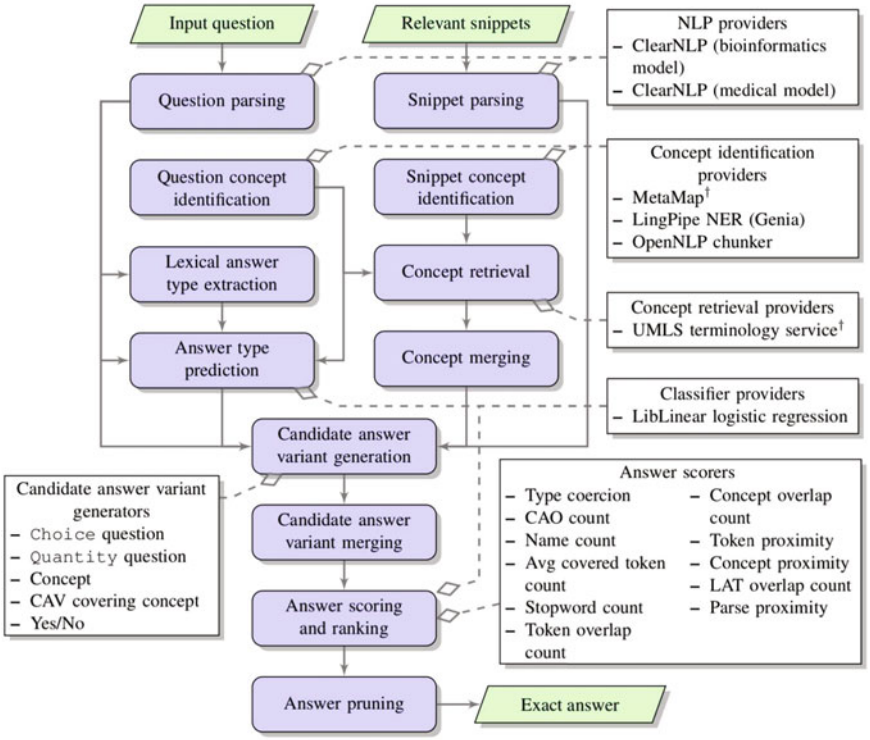


Fig. 8.5 Modular architecture and components for BioASQ phase B (Yang et al. 2015)

Genomics Task, where CSE was also able to find a significantly better combination of components.

The positive results from applying CSE to the TREC Genomics tasks were extended by applying CSE to a much larger, more complex task with many subtasks: The BioASQ Challenge (Chandu et al. 2017; Yang et al. 2015, 2016). Using a shared corpus of biomedical documents (PubMed articles), the BioASQ organizers created a set of interrelated tasks for question answering: retrieval of relevant medical concepts, articles, snippets and RDF triples, plus generation of both exact and "ideal" (summary) answers for each question. Figure 8.5 illustrates the modular architecture used to generate exact answers for 2015 BioASQ Phase B (Yang et al. 2015). Across the five batch tests in Phase B, the CMU system achieved top scores in concept retrieval, snippet retrieval, and exact answer generation. As shown in Fig. 8.5, this involved evaluating and optimizing ensembles of language models, named entity extractors, concept retrievers, classifiers, candidate answer generators, and answer scorers.

8.4 Remaining Challenges and Future Directions

Much recent work in question-answering has focused on neural models which are trained on large numbers of question-answer pairs created by human curators (e.g., SQUAD (Rajpurkar et al. 2016), SQUAD 2 (Rajpurkar et al. 2018). While neural QA approaches are effective when large numbers of labeled training examples are available (e.g., more than 100,000 examples), in practice neural approaches are very sensitive to the distribution of answer texts and corresponding questions that are created by the human curators. For example, a recent study showed that an advanced question curation strategy, using the original answer texts from SQUAD produced a dataset (ParallelQA) that was much tougher for neural models; models evaluated on SQUAD and ParallelQA did approximately 20% worse on ParallelQA (Wadhwa et al. 2018c). In the future, we believe that QA research must focus more energy on defining effective curation strategies, so that the best components and models may be chosen and built into an effective solution using the least amount of labeled data and human resources. In preliminary work, we have adopted a comparative evaluation framework (Wadhwa et al. 2018a) that allows us to compare the performance of different neural QA approaches across datasets, in order to identify the approach with the most general capability.

It is also the case that neural approaches to QA often assume that a single neural model or an ensemble of neural models will produce an effective solution. In reality, it is difficult for any one model to learn all of the varied ways in which answers correspond to questions presented by the user. Due to the high cost of training and evaluating neural models, researchers often don't consider more sophisticated combinations of models, or ensembles with non-neural components. This movement away from the multi-strategy, multi-component approach that reached its zenith in IBM Watson is unfortunate, because it has focused the QA field on just a few, artificially created datasets that are comparatively easy for neural QA approaches.

It is ironic that the best-performing automatic QA system in the LiveQA evaluations (Wang and Nyberg 2015b, 2016, 2017) combined sophisticated neural models with an optimized version of the classic BM25 algorithm; neither the neural model nor BM25 was competitive by itself, but the combination of these two algorithms provided the most effective solution for the Yahoo! Answers data set. While it is true that curating datasets which can be solved by neural methods has stimulated the development of more capable, sophisticated neural models, neural approaches still rely on hundreds of thousands of labeled examples, and do not perform well when (a) there is limited training data, (b) there is a large variance in the lengths of the question versus answer texts, and (c) there is little lexical overlap between question and answer texts (Wadhwa et al. 2018b, c).

8.5 Conclusion

As we have discussed in this chapter, the development of common interchange formats for language processing modules in the JAVELIN project (Lin et al. 2005; Mitamura et al. 2007; Shima et al. 2006) led to the use of common schemas in the NTCIR IR4QA embedded task (Mitamura et al. 2008), which we believe is the first example of a common QA evaluation using a shared data schema and automatic combination runs. Although it is expensive to use human evaluators to judge all possible combinations of systems, automatic metrics (such as ROUGE) can be used to find novel combinations that seem to perform well or better than the state of the art; this subset of novel systems can then be evaluated by humans. In the OAQA project (which followed JAVELIN at CMU), development participants began to create gold-standard datasets that include expected outputs for all stages in the QA pipeline, not just the final answer (Garduño et al. 2013). This allowed precise automatic evaluation and more effective error analysis, leading to the development of high-performance QA incorporating hundreds of different strategies in real time (IBM Watson) (Ferrucci et al. 2010). The OAQA approach was also used to evaluate and optimize several multi-strategy QA systems, some of which achieved state-of-the-art performance on the TREC Genomics datasets (2006 and 2007) (Yang et al. 2013) and BioASQ tasks (2015–2018) (Chandu et al. 2017; Yang et al. 2015, 2016).

Although academic datasets in the QA field have recently focused on specific parts of the QA task (such as answer sentence and answer span selection) (Rajpurkar et al. 2016, 2018) which can be solved by a single deep learning or neural architecture, systems which achieve state-of-the-art performance on messy, real-world datasets (such as Jeopardy! or Yahoo! Answers) must employ a multi-strategy approach. For example, neural QA components were combined with classic information-theoretic algorithms (e.g., BM25) to achieve the best automatic QA system performance on the TREC LiveQA task (2015–2017) (Wang and Nyberg 2015a, b, 2016, 2017), which was based on a Yahoo! Answers community QA dataset. It is our expectation that a path to more general QA performance will be found by upholding the tradition of multi-strategy, multi-component evaluations pioneered by NTCIR. In our most recent work, we have tried to extend the state of the art in neural QA by performing comparative evaluations of different neural QA architectures across QA datasets (Wadhwa et al. 2018a), and we expect that future work will also focus on how to curate the most challenging (and realistic) datasets for real-world QA tasks (Wadhwa et al. 2018c).

Acknowledgements We would like to thank to the editors and to the past organizers and participants of the NTCIR ACLIA QA tasks. Special thanks go to Hideki Shima, who worked on CMU's Javelin QA system to develop the component-based evaluation and helped to organize the ACLIA tasks. We also thank the other students and staff who contributed to the JAVELIN, ACLIA, OAQA, and LiveQA projects.

References

Andrews C (2011) Ibm announces eight universities contributing to the watson computing system's development. PR Newswire. https://tinyurl.com/yxsmx8q5

Asahara M, Matsumoto Y (2003) Japanese named entity extraction with redundant morphological analysis. In: Proceedings of the 2003 human language technology conference of the North American chapter of the association for computational linguistics. pp 8–15. https://www.aclweb.org/anthology/N03-1002

Chandu KR, Naik A, Chandrasekar A, Yang Z, Gupta N, Nyberg E (2017) Tackling biomedical text summarization: OAQA at bioasq 5b. In: Cohen KB, Demner-Fushman D, Ananiadou S, Tsujii J (eds) BioNLP 2017, Association for Computational Linguistics, Vancouver, Canada. pp 58–66. 10.18653/v1/W17-2307. https://doi.org/10.18653/v1/W17-2307

Ferrucci D, Lally A, Verspoor K, Nyberg E (2009a) Unstructured information management architecture (uima) version 1.0. OASIS Standard. OASIS

Ferrucci D, Nyberg E, Allan J, Barker K, Brown E, Chu-Carroll J, Ciccolo A, Duboue P, Fan J, Gondek D et al (2009b) Towards the open advancement of question answering systems. IBM, IBM Res Rep, Armonk, NY

Ferrucci D, Brown E, Chu-Carroll J, Fan J, Gondek D, Kalyanpur AA, Lally A, Murdock JW, Nyberg E, Prager J et al (2010) Building watson: an overview of the deepqa project. AI Magazine 31(3):59–79

Garduño E, Yang Z, Maiberg A, McCormack C, Fang Y, Nyberg E (2013) CSE framework: a uima-based distributed system for configuration space exploration. In: Klügl P, Eckart de Castilho R, Tomanek K (eds) Proceedings of the 3rd workshop on unstructured information management architecture, vol 1038. CEUR-WS.org, CEUR Workshop, Darmstadt, Germany, pp 14–17. http://ceur-ws.org/Vol-1038/paper_10.pdf

Lin F, Shima H, Wang M, Mitamura T (2005) CMU JAVELIN system for NTCIR5 CLQA1. In: Kando N (ed) Proceedings of the Fifth NTCIR workshop meeting on evaluation of information access technologies: information retrieval, question answering and cross-lingual information access, NTCIR-5, National Center of Sciences, National Institute of Informatics (NII), Tokyo, Japan. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings5/data/CLQA/NTCIR5-CLQA1-LinF.pdf

Lin J, Demner-Fushman D (2006) Will pyramids built of nuggets topple over? In: Proceedings of the human language technology conference of the NAACL, main conference. Association for Computational Linguistics, New York, USA, pp 383–390. https://www.aclweb.org/anthology/N06-1049

Mitamura T, Lin F, Shima H, Wang M, Ko J, Betteridge J, Bilotti MW, Schlaikjer AH, Nyberg E (2007) JAVELIN III: cross-lingual question answering from japanese and chinese documents. In: Kando N (ed) Proceedings of the 6th NTCIR workshop meeting on evaluation of information access technologies: information retrieval, question answering and cross-lingual information access, NTCIR-6. National Center of Sciences, Tokyo, Japan, National Institute of Informatics (NII). http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings6/NTCIR/51.pdf

Mitamura T, Nyberg E, Shima H, Kato T, Mori T, Lin C, Song R, Lin C, Sakai T, Ji D, Kando N (2008) Overview of the NTCIR-7 ACLIA tasks: advanced cross-lingual information access. In: Kando N (ed) Proceedings of the 7th NTCIR workshop meeting on evaluation of information access technologies: information retrieval, question answering and cross-lingual information access, NTCIR-7. National Center of Sciences, National Institute of Informatics (NII), Tokyo, Japan. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings7/pdf/NTCIR7/C1/CCLQA/01-NTCIR7-OV-CCLQA-MitamuraT.pdf

Rajpurkar P, Zhang J, Lopyrev K, Liang P (2016) Squad: 100,000+ questions for machine comprehension of text. arXiv:160605250

Rajpurkar P, Jia R, Liang P (2018) Know what you don't know: Unanswerable questions for squad. arXiv:180603822

Shima H, Wang M, Lin F, Mitamura T (2006) Modular approach to error analysis and evaluation for multilingual question answering. In: Calzolari N, Choukri K, Gangemi A, Maegaard B, Mariani J, Odijk J, Tapias D (eds) Proceedings of the fifth international conference on language resources and evaluation, LREC 2006. European Language Resources Association (ELRA), Genoa, Italy, pp 1143–1146. http://www.lrec-conf.org/proceedings/lrec2006/pdf/782_pdf.pdf

Voorhees EM (2003) Overview of the TREC 2003 question answering track. In: Voorhees EM, Buckland LP (eds) Proceedings of The twelfth text Retrieval conference, TREC 2003, vol Special Publication 500-255. National Institute of Standards and Technology (NIST), Gaithersburg, Maryland, USA, pp 54–68. http://trec.nist.gov/pubs/trec12/papers/QA.OVERVIEW.pdf

Voorhees EM (2004) Overview of the TREC 2004 question answering track. In: Voorhees EM, Buckland LP (eds) Proceedings of the thirteenth text retrieval conference, TREC 2004, vol Special Publication 500-261. National Institute of Standards and Technology (NIST), Gaithersburg, Maryland, USA. http://trec.nist.gov/pubs/trec13/papers/QA.OVERVIEW.pdf

Wadhwa S, Chandu KR, Nyberg E (2018a) Comparative analysis of neural QA models on squad. http://arxiv.org/abs/1806.06972

Wadhwa S, Chandu KR, Nyberg E (2018b) Comparative analysis of neural QA models on squad. In: Choi E, Seo M, Chen D, Jia R, Berant J (eds) Proceedings of the workshop on machine reading for question answering@ACL 2018. Association for Computational Linguistics, Melbourne, Australia, pp 89–97. 10.18653/v1/W18-2610. https://www.aclweb.org/anthology/W18-2610/

Wadhwa S, Embar V, Grabmair M, Nyberg E (2018c) Towards inference-oriented reading comprehension: parallelQA. http://arxiv.org/abs/1805.03830

Wang D, Nyberg E (2015a) CMU OAQA at TREC 2015 liveqa: discovering the right answer with clues. In: Voorhees EM, Ellis A (eds) Proceedings of the twenty-fourth text retrieval conference, TREC 2015, vol Special Publication 500-319. National Institute of Standards and Technology (NIST), Gaithersburg, Maryland, USA. http://trec.nist.gov/pubs/trec24/papers/oaqa-QA.pdf

Wang D, Nyberg E (2015b) A long short-term memory model for answer sentence selection in question answering. In: Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing of the Asian federation of natural language processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 2: Short Papers. The Association for Computer Linguistics, pp 707–712. http://aclweb.org/anthology/P/P15/P15-2116.pdf

Wang D, Nyberg E (2016) CMU OAQA at TREC 2016 liveqa: An attentional neural encoder-decoder approach for answer ranking. In: Voorhees EM, Ellis A (eds) Proceedings of the twenty-fifth text retrieval conference, TREC 2016, Gaithersburg, Maryland, USA, November 15–18, 2016, National Institute of Standards and Technology (NIST), vol Special Publication 500-321. http://trec.nist.gov/pubs/trec25/papers/CMU-OAQA-QA.pdf

Wang D, Nyberg E (2017) CMU OAQA at TREC 2017 liveqa: a neural dual entailment approach for question paraphrase identification. In: Voorhees EM, Ellis A (eds) Proceedings of the twenty-sixth text retrieval conference, TREC 2017, Gaithersburg, Maryland, USA, November 15–17, 2017, National Institute of Standards and Technology (NIST), vol Special Publication 500-324. https://trec.nist.gov/pubs/trec26/papers/CMU-OAQA-QA.pdf

Yang Z (2017) Analytics meta learning. PhD thesis, Carnegie Mellon University, 5000 Forbes Avenue

Yang Z, Garduño E, Fang Y, Maiberg A, McCormack C, Nyberg E (2013) Building optimal information systems automatically: configuration space exploration for biomedical information systems. In: He Q, Iyengar A, Nejdl W, Pei J, Rastogi R (eds) 22nd ACM international conference on information and knowledge management, CIKM'13, San Francisco, CA, USA, October 27 November 1, 2013, ACM, pp 1421–1430. https://doi.org/10.1145/2505515.2505692

Yang Z, Gupta N, Sun X, Xu D, Zhang C, Nyberg E (2015) Learning to answer biomedical factoid & list questions: OAQA at bioasq 3b. In: Cappellato L, Ferro N, Jones GJF, SanJuan E (eds) Working notes of CLEF 2015—conference and labs of the evaluation forum, Toulouse, France, September 8–11, 2015., CEUR-WS.org, CEUR Workshop Proceedings, vol 1391. http://ceur-ws.org/Vol-1391/114-CR.pdf

Yang Z, Zhou Y, Nyberg E (2016) Learning to answer biomedical questions: OAQA at BioASQ 4B. In: Proceedings of the fourth BioASQ workshop, association for computational linguistics, Berlin, Germany, pp 23–37. https://doi.org/10.18653/v1/W16-3104, https://www.aclweb.org/anthology/W16-3104

Chapter 9
Temporal Information Access

Masaharu Yoshioka and Hideo Joho

Abstract This chapter introduces the research background and details of temporal information access tasks in the NTCIR. The GeoTime task was the first attempt to evaluate temporal information retrieval as an extension of an information-retrieval-for-question-answering task. Temporalia was a task to investigate the role of temporal factors in a search.

9.1 Introduction

Temporal information is important to understand the document and to represent users' information needs. In the early age of Named Entity Recognition (NER), tasks such as MUC-6 (Sundheim 1995) and IREX (Sekine and Isahara 2000), date and time were selected as categories for NER. In information access technology research, there had been several studies on using such temporal information (e.g., Mani et al. 2004), but there have not been many studies on temporal information retrieval (Alonso et al. 2007).

Compared to the usage of temporal information, Geographical Information Retrieval (Geographic Information Retrieval (GIR)) had attracted more researchers, and a series of workshops on Geographic Information Retrieval (GIR) was started in 2004 (Purves and Jones 2004). In this series of workshops, temporal information was only discussed as a related topic of the task.

At NTCIR-8, *GeoTime* (geographic and temporal information retrieval) tasks (Gey et al. 2010) were launched as first attempts to construct a test collection for

M. Yoshioka (✉)
Faculty of Information Science and Technology, and Global Station for Big Data and Cybersecurity, Global Institution for Collaborative Research and Education, Hokkaido University, Sapporo, Japan
e-mail: yoshioka@ist.hokudai.ac.jp

H. Joho
Faculty of Library, Information and Media Science, University of Tsukuba, 1-2 Kasuga, Tsukuba, Ibaraki 305-8550, Japan
e-mail: hideo@slis.tsukuba.ac.jp

T. Sakai et al. (eds.), *Evaluating Information Retrieval and Access Tasks*,
The Information Retrieval Series 43,
https://doi.org/10.1007/978-981-15-5554-1_9

temporal information retrieval. This task was designed as an extension of IR4QA tasks (Mitamura et al. 2008). There were two types of temporal-related queries. One query type asked for temporal information ("when" question), while the other query type used temporal information as constraints (winning team of Superbowl in 2002). Details of the information related to the task are discussed in Sect. 9.2.

Following the success of GeoTime Tasks in NTCIR-8 and 9, a new task was proposed to further investigate the role of temporal factors in the search. The task was called *Temporalia* (Temporal Information Access) and was run twice in NTCIR-11 and 12. One of the important innovations in Temporalia was to provide a test collection that allowed researchers to examine the performance of time-aware search applications using categories such as past, recent, future, and atemporal rather than focusing on recency queries. Details of the information related to the task are discussed in Sect. 9.3.

9.2 Temporal Information Retrieval

There are several IR applications that utilize temporal information; e.g., ad hoc retrieval, hit-list clustering based on the temporal aspect, exploratory search, and visualization of results based on the temporal relationships (Alonso et al. 2007). However, there was no IR evaluation campaign for temporal information retrieval except for some discussions related to Geographic Information Retrieval (GIR) (Purves and Jones 2004).

To utilize and incorporate the discussion related to *Geographic Information Retrieval (GIR), GeoTime* (geographic and temporal information retrieval) tasks (Gey et al. 2010) were launched at NTCIR-8 as an extension of IR4QA for handling spatial and temporal-related queries (Mitamura et al. 2008).

9.2.1 NTCIR-8 GeoTime Task

The NTCIR-8 GeoTime Task was designed as an IR4QA task for the geographical and temporal-related queries.

Parts of queries were constructed using the information of notable events listed in Wikipedia,[1] and several queries were derived from the *ACLIA* collection (Sakai et al. 2010). This task used the New York Times collection for the English document database and the Mainichi Japanese newspaper collection for the Japanese document database.

For the evaluation, because most of the queries have both temporal and spatial aspects, the articles that can be used for answering questions for temporal and spatial aspects were categorized as "fully relevant" and ones that can answer only one aspect

[1] For example, notable events in 2002 are listed at https://en.wikipedia.org/wiki/2002.

(temporal or spatial) are categorized as "partially relevant". The submitted results were evaluated by the same schemes used for the ACLIA IR4QA collection (Sakai et al. 2010).

The following are examples of the queries.

- How old was Max Schmeling when he died, and where did he die?
- When and where did a massive earthquake occur in December 2003?

The former question asks for temporal information using a "when" question. The latter question also has the "when" question style, but it also uses temporal information to represent constraints ("in December 2003").

There were 14 teams that participated in NTCIR-8 GeoTime (8 and 7 teams submitted runs for Japanese and English runs, respectively) using various approaches (Gey et al. 2010). The baseline system utilized ordinary ad hoc IR systems such as probabilistic IR with blind relevance feedback. This baseline system worked well for the English run but underperformed in the Japanese run. Another approach utilized a NER system and/or geographic resources to extract named entity information including geographic and temporal information from the queries and documents. The best performing NTCIR-8 Japanese run was a hybrid approach that combined the probabilistic approach and weighted Boolean query formulation based on the NER results (Yoshioka 2010). There were approaches that focused on geographic information including the hierarchical relationship among location names (e.g., Tokyo is a part of Japan) and the distance between the extracted location of the query and document, and there were several discussions about the temporal information.

Another approach emphasized the style of the query in GeoTime. Because the query was provided as a question in IR4QA style, the relevant documents should contain the information for its answer. Based on this understanding, one team counted the number of temporal or geographic mentions that can be candidates for the answer for re-ranking (Kishida 2010). Another approach decomposes the question into one for geographic information and another for temporal information. After decomposing the question, they used a factoid question answering system to determine the answer and utilize its information for constructing new queries (Mori 2010). However, those approaches did not perform well for the task.

From the analysis of the difficult queries based on the evaluation of the submitted results, two types of difficult queries were identified. One type is that the system tends to misinterpret the constraint of the query. An example of the query is "When and where were the 2010 Winter Olympics host city location announced?". In this question, "2010" is used as a part of an event name and not as a constraint-specifying articles should be selected from those published in 2010 or after. Another type of difficult query requires a list of events to determine relevant articles. An example of this type of query is "When and where were the last three Winter Olympics held?". It is difficult to retrieve relevant articles without generating an event list that satisfies the query constraint. Details of the discussion about the difficulties of the problem are addressed in Sect. 9.2.3.

9.2.2 NTCIR-9 GeoTime Task Round 2

By comparing the English runs and Japanese runs, there were queries that have large performance variability for the same topics. Therefore, the news article data for English runs were expanded to include newspapers from different countries. In addition to the news articles of the New York Times collection, English versions of Korean Times (Korea), Mainichi (Japan), and Xinhua (China) were used to construct a document database.

There were 12 teams that participated for NTCIR-9 GeoTime (5 and 9 teams submitted runs for the Japanese and English runs, respectively) using various approaches (Gey et al. 2011). One large difference from the previous GeoTime was the usage of external resources such as Yahoo PlaceMaker, Wikipedia, DBpedia, Geonames, Google Maps, and the Alexandria Digital Library gazetteer. Most of the teams utilized such information for improving the retrieval results related to the geographic queries. However, the query that required reverse geocoding (finding place names from a latitude/longitude information) was not appropriately handled except that the team manually extracted the related event name using Wikipedia.

The best performing team for both Japanese and English runs used manual query expansion with a related event name and/or name of the location using Wikipedia and Google Maps (Sato 2011). Because this approach was not automatic, it was difficult to compare this result with others. However, this result suggested that the extraction of such related event names or locations is crucial for improving the recall of the related articles.

9.2.3 Issues Discussed Related to Temporal IR

One of the difficult queries in NTCIR-8 GeoTime was "When and where were the 2010 Winter Olympics host city location announced?". To discuss the difficulties of this query, it was necessary to discuss the types of temporal expression. Alonso et al. (2007) proposed the following types of temporal expression.

1. Explicit. Temporal expression directly describes its information (e.g., September 11, 2001).
2. Implicit. There is imprecise temporal information, such as names of holidays or events. It is possible to extract temporal information using knowledge about such holidays or events (e.g., Labor Day, 2001, can be mapped to September 1, 2001, and Vancouver Winter Olympics can be mapped to February 2010).
3. Relative. Temporal expressions represent temporal entities that refer to other temporal entities. Temporal information resolution is necessary to extract its temporal information (e.g., "yesterday" of the news article published on September 12, 2001, can be mapped to September 11, 2001).

In the query discussed above, "the 2010 Winter Olympics" is the name of the event and can be treated as an implicit temporal expression. However, it is not a

constraint for selecting relevant articles. It is necessary to have a mechanism to select which kinds of temporal expression should be used for constraints to retrieve relevant articles.

Another problem is related to handling the relationship between temporal information and event names that represent imprecise temporal information. An example of this difficult query is "When and where were the last three Winter Olympics held?"; "the last three" uses relative and imprecise temporal information to select relevant event names (three Winter Olympic event names). Because most of the relevant documents contain such event names but do not have such relative expression, it is difficult to retrieve such articles without event names. As we confirmed in the case of NTCIR-9 GeoTime, query expansion using such event names significantly improves the performance.

9.3 Temporal Query Analysis and Search Result Diversification

To facilitate research on temporal information access, Temporalia-1 in NTCIR-11 (Joho et al. 2014) focused on each of the four categories in a structured way, while Temporalia-2 in NTCIR-12 (Joho et al. 2016) was designed to encourage researchers to explore ways to combine the four categories in a meaningful way. Both were designed to address the temporal ambiguity and diversity of the search space.

9.3.1 NTCIR-11 Temporal Information Access Task

Temporalia-1 at NTCIR-11 consisted of two subtasks: Temporal query intent classification and temporal information retrieval.

9.3.1.1 Temporal Query Intent Classification

The Temporal Query Intent Classification (TQIC) subtask was used to classify a given query into one of the following classes: past, recency, future, and atemporal. Example queries are ground truth temporal classes are shown in Table 9.1. The classes were defined as follows.

Past: class characterizing queries about past entities/events whose search results are not expected to change much with the passage of time.

Recency: class characterizing queries about recent entities/events, whose search results are expected to be timely and up to date. The information contained in the search results usually changes quickly with the passage of time. Note that this type of query usually refers to events that happened in the near past or at the present time.

Table 9.1 Example queries and ground truth temporal classes for the TQIC subtask (dry run)

Query class	Query example
Past	Price hike in Bangladesh 2008
Past	Who was Martin Luther
Past	When did the Titanic sink
Past	Yuri Gagarin cause of death
Past	History of Coca-Cola
Recency	Apple stock price
Recency	Number of millionaires in USA
Recency	Time in London
Recency	Trendy plus size clothing
Recency	Did the Pirates win today
Future	2013 MLB playoff schedule
Future	Release date for iOS7
Future	College baseball regional projections
Future	Disney prices 2014
Future	Long-term weather forecast
Atemporal	Blood pressure monitor
Atemporal	Distance from earth to sun
Atemporal	How to start a conversation
Atemporal	New York Times
Atemporal	Lose weight quickly

In contrast, the "past" query category tends to refer to events in a relatively distant past.

Future: class characterizing queries about predicted or scheduled events, and the search results of which should contain future-related information.

Atemporal: class characterizing queries without any clear temporal intent (i.e., their returned search results are not expected to be related to time and should not change much over time). Navigational queries are considered to be atemporal.

Participants were handed a set of query strings and query submission dates and were asked to develop a system to classify each of the query strings to one of the four above-mentioned temporal classes. As this problem rather requires different kinds of knowledge (e.g., historical information or information on planned events), the participants were allowed to use any external resources to complete the TQIC subtask as long as the details of external resource usage were described in their reports. Each participating team was asked to submit a temporal class (past, recency, future, or atemporal) for each one of the queries. The performance of submitted runs was measured by the number of queries with correct temporal classes divided by the total number of queries.

Table 9.2 Example topics for the TIR subtask (dry run)

	Girl with the Dragon Tattoo
Description	I have recently watched a film called Girl with the Dragon Tattoo and I really liked it. Therefore, I would like to gather information about the movie
Past question	How did the casting of the film develop?
Recency question	What did the recent reviews say about the film?
Future question	Is there any plan about a sequel?
Atemporal question	What are the names of the main actors and actresses of the film?
Search date	28 Feb 2013 GMT+0:00

9.3.1.2 Temporal Information Retrieval

The Temporal Information Retrieval (TIR) subtask was used to retrieve a set of documents in response to a search topic that incorporates a time factor. In addition to a typical search topic description (i.e., title, description, and subtopics), the TIR search topic description also contains a query submission date (see Table 9.2). This subtask required indexing of the document collection with any standard information retrieval toolkit. Participants were asked to submit the top 100 documents for each temporal question per topic (e.g., top 100 documents for a past question and another 100 for a recency question). The retrieval effectiveness was evaluated by the precision at 20 for each of the temporal questions. Similar to the TQIC subtask, the results section presents an analysis of the performance across temporal questions.

9.3.2 NTCIR-12 Temporal Information Access Task Round 2

Temporalia-2 at NTCIR-12 also consisted of two subtasks: temporal intent disambiguation and temporally diversified retrieval.

9.3.2.1 Temporal Intent Disambiguation

The Temporal Intent Disambiguation (TID) subtask determined a probability distribution of a query over four classes denoting the types of temporal intent: past, recency, future, and atemporal. The definitions of the four classes were based on TQIC in Temporalia-1. An example of the probability distribution of temporal intents is shown in Tables 9.3.

Table 9.3 Example queries for the TID subtask (dry run) with query submission date of May 1, 2013. Ground truth probability of temporal intents was determined by votes from crowd workers

Query	Past	Recency	Future	Atem.
Australian open	0.091	0.0	0.455	0.455
Motorcycle accident June	0.7	0.0	0.3	0.0
NBA finals	0.1	0.0	0.4	0.5
NBA playoff schedule	0.0	0.2	0.6	0.2
Price of oil	0.0	0.9	0.0	0.1
How to lose weight	0.0	0.1	0.0	0.9
Time in India	0.0	1.0	0.0	0.0
History of volleyball	1.0	0.0	0.0	0.0

Table 9.4 Example topics for the TDR subtask

	Junk food health effect
Description	I am concerned about the health effects of junk food in general. I need to know more about their ingredients, impact on health, history, current scientific discoveries, and any prognoses
Past question	When did junk foods become popular?
Recency question	What are the latest studies on the effect of junk foods on our health?
Future question	Will junk food continue to be popular in the future?
Atemporal question	How junk foods are defined?
Search date	29 May 2013 GMT+0:00

9.3.2.2 Temporally Diversified Retrieval

The Temporally Diversified Retrieval (TDR) subtask required participants to retrieve a set of documents relevant to each of four temporal intent classes for a given topic description (see Table 9.4). Participants were also asked to return a set of documents that is temporally diversified for the same topic. They received a set of topic descriptions, query issuing times, and indicative search questions for each temporal class (past, recency, future, and atemporal). The objective of the indicative search questions was to show one possible subtopic under a particular temporal class. Participants were asked to develop systems that can produce a total of five search results per topic (past, recency, future, atemporal, and diversified).

9.3.3 Implications from Temporalia

This section discusses the implications of Temporalia tasks on system development and test collection, respectively.

9.3.3.1 Implications on System Development

From the meta-analysis of 17 runs submitted to the TQIC subtask, the classification of recency queries was found to be the most difficult with 56% accuracy, and past queries were the easiest with 73%. Another overall trend was that no single approach was effective across the four temporal classes. A confusion matrix showed that: (1) atemporal queries are likely to be confused as either recency or past queries (16.7% and 9.6%, respectively), (2) past queries are likely to be confused as atemporal queries (13.1%), (3) recency queries are likely to be confused as future or atemporal queries (28.2% and 13.5%, respectively), and (4) future queries tend to be confused as recency queries (25.9%). Correlation analysis suggested that it was difficult to apply the same technique to predict recency queries and atemporal queries with high accuracy.

The TIR subtask showed a similar pattern with varied performance across the four classes. No single system was able to perform the best for all classes. The learning-to-rank approach was effective for atemporal and past queries, while BM25 performed well for recency and future topics.

The meta-analysis of 37 runs submitted to the TID subtask suggested that when a query was temporally ambiguous and multiple temporal classes can be inferred, detecting atemporal features was the most difficult. Also, some techniques were good at modeling temporally less diverse queries (i.e., a fewer number of nonzero probability classes), while other methods were good at modeling temporally more diverse queries.

The results of the TDR subtask suggested that a learning-to-rank approach was effective in retrieving relevant documents for all classes compared to BM25. However, the best performance on temporal search result diversification was obtained by a round-robin of BM25 rankings of four temporal classes, suggesting that there is still room for improvement in this area.

9.3.3.2 Implications for Test Collection

Document Collections

One of the challenges in building a test collection for temporal-aware technologies was to obtain access to document collections that have rich temporal features. Temporalia was fortunate to have support to use the "LivingKnowledge news and blogs annotated sub-collection" constructed by the LivingKnowledge project and distributed by the Internet Memory Foundation. The collection was approximately

20 GB large when uncompressed and over 5 GB large when zipped. The collection spanned from May 2011 to March 2013 and contains around 3.8 million documents collected from about 1,500 different blogs and news sources. The data were split into 970 files based on the date and sources (there might be more than one file per day). Texts in the collection were annotated by entities and by temporal expressions that were resolved to a specific day, month, or year (Matthews et al. 2010). The relative expressions such as "next month" was resolved based on the publication date of the articles.

In Temporalia-2, we also made efforts to diversify the target language of document collections to Chinese using SogouCA-2012[2] and SogouT-2012.[3] Similar to the English collection, SogouCA-2012 was based on news articles from major publishers in China. For annotating temporal expressions, a variant of the standard format TIMEX3 used in TempEval task was applied.[4]

Relevance Assessments

Another challenge we faced during the construction of Temporalia test collections was relevance assessment. The temporality of topics and relevance can be subjective and not always deterministic. Therefore, we used a mixture of methods to ensure that both queries and documents were temporally annotated for evaluation.

We had a combination of workshops and crowdsourcing in formal runs. In another series of workshops, participants (not necessarily the same people as topic creators) were asked to read the formal run topic descriptions carefully and assess the relevance of the retrieved documents.

The documents were then evaluated using crowdsourcing as for their relevance to each of the temporal subclasses. For each assigned subtopic, CrowdFlower workers were asked to identify at least one highly relevant and one irrelevant document. They were also asked to note the relevant text from original documents in the case of highly relevant documents. The relevance of these documents was verified by a third person during the workshop to improve their reliability.

The documents initially identified by the workshop participants were then used as "test questions" of crowdsourcing jobs. Test questions were questions that crowdsourcing workers had to pass to participate in our relevance assessment jobs. We used CrowdFlower to run relevance judgments. Our configuration of crowdsourcing is based on common settings used by various IR evaluations (e.g., Kazai et al. 2013).

- Each task had five documents to judge
- Ten cents were paid for one task
- Each task had 120 s of minimum work time
- Each document had at least three judgments

[2] http://www.sogou.com/labs/dl/ca.html.

[3] http://www.sogou.com/labs/dl/t-e.html.

[4] http://www.timeml.org/tempeval2/tempeval2-trial/guidelines/timex3guidelines-072009.pdf.

We had several iterations of revising job instructions and relevance criteria before running all formal run subtopics. We tested both detailed instructions and simple instructions, but we received mixed responses from workers. Also, detailed instructions caused the time required for relevance assessment to increase too much. After several iterations, we decided to use the following three levels of relevance criteria.

Not Relevant The web page does not contain any information to answer the search question.

Highly Relevant The web page discusses the answer to the search question exhaustively. In the case of a multifaceted search question, all or most subthemes or viewpoints are covered. Typical extent: several text paragraphs, at least four sentences or facts.

Relevant The web page contains some information to answer the question, but the presentation is not exhaustive. In the case of a multifaceted search question, only some of the subthemes or viewpoints are covered. Typical extent: one text paragraph, or one to three sentences or facts.

9.4 Related Work and Broad Impacts

After introducing temporal information retrieval task as a part of GeoTime task at NTCIR 8, there were several lines of research emerged as a variation of temporal information retrieval. Kanhabua et al. (2015) is a comprehensive textbook that introduces such research results. Moulahi et al. (2016) also summarizes past efforts in temporal information retrieval evaluation and discuss future directions. From these results, we would like to introduce some research that is highly related to the tasks discussed above.

Strötgen and Gertz (2013), Daoud and Huang (2013) both proposed proximity methods for the Geotemporal Information Retrieval task. In this method, proximity of the geographic and temporal information are considered for ranking documents in addition to the standard information retrieval ranking such as BM25. Another interesting example is event-centric search and exploration (Strötgen and Gertz 2012). This framework was proposed for analyzing historic documents using geographic and temporal constraints constructed from event information. In the discussion of GeoTime, there was a consideration of using the name of an event for time constraints. This event-centric approach utilizes these characteristics to find documents relevant to the event for exploration.

There have been related efforts to construct test collections for Information Access technologies with temporal awareness, such as the TREC Temporal Summarization Track (2012–2015) (Aslam et al. 2015; Guo et al. 2013) and TREC Knowledge Base Acceleration Track (2012–2014) (Frank et al. 2014). The TREC Temporal Summarization Track had two subtasks: Sequential Update Summarization and Value Tracking. Sequential Update Summarization sought to find timely, sentence-level, reliable, relevant, and nonredundant updates about developing event, while Value Tracking

aimed at tracking values of event-related attributes that were of high importance to the event. TREC Knowledge Base Acceleration Track was a challenge for filtering a large stream of text to find documents that can help update knowledge bases like Wikipedia, Facebook, or Crunchbase. Both efforts either explicitly or implicitly had a focus on recency information about entities. NTCIR Temporalia was, on the other hand, designed to facilitate research on diverse temporal attributes in a systematic manner.

There have been several extensions of the original work. For example, Hasanuzzaman et al. (2016) applied temporal query intent classification techniques to stock market analysis. Rizzo and Montesi (2017) used the LivingKnowledge collection to conduct a temporal analysis of a digital library collection. Finally, Joho et al. (2013, 2015) used the Temporalia test collection to study temporal information-seeking behavior in a controlled user study and a questionnaire-based study. The studies identified the difference in resource selection and relevant content types across temporal attributes of information needs. These are some of the ways in which the test collection for temporal information access can have broader impacts than the original objectives of the resources.

See the citation of the overview papers (Gey et al. 2010, 2011; Joho et al. 2014, 2016) for more details of broader impacts from GeoTime and Temporalia.

9.5 Conclusion

We have introduced two tasks related to temporal information access in the NTCIR workshop. GeoTime was the first attempt to place more emphasis on temporal search, and Temporalia provided a framework to examine the performance of time-aware search application using a test collection. The review of the literature suggests that these resources have been useful for researchers to advance temporal information access technologies and to better understanding temporal information-seeking behavior.

Acknowledgements The authors would like to thank Prof. Fred Gey, Prof. Ray Larson, Dr. Jorge Machado, Dr. Roi Blanco, Dr. Adam Jatowt, Dr. Haitao Yu, Dr. Shuhei Yamamoto, and Mr. Hajime Naka for their efforts to organize the NTCIR GeoTime and NTCIR Temporalia Tasks. The authors also thank all participants in the tasks for their interest and contributions to the technological advance in temporal information access. LivingKnowledge project, Internet Memory Foundation, and Sogou have provided valuable document collections that enabled us to build the test collections. Finally, the authors would like to thank Doug Oard, Tetsuya Sakai, and Noriko Kando for their editorial work on this manuscript.

References

Alonso O, Gertz M, Baeza-Yates R (2007) On the value of temporal information in information retrieval. SIGIR Forum 41(2):35–41. https://doi.org/10.1145/1328964.1328968

Aslam JA, Diaz F, Ekstrand-Abueg M, McCreadie R, Pavlu V, Sakai T (2015) TREC 2015 temporal summarization track overview. In: Voorhees EM, Ellis A (eds) Proceedings of the 24th text retrieval conference, TREC 2015, Gaithersburg, Maryland, USA, 17–20 November 2015, National Institute of Standards and Technology (NIST), vol Special Publication, pp 500–319. http://trec.nist.gov/pubs/trec24/papers/Overview-TS.pdf

Daoud M, Huang JX (2013) Modeling geographic, temporal, and proximity contexts for improving geotemporal search. J Am Soc Inf Sci Technol 64(1):190–212. https://doi.org/10.1002/asi.22648

Frank JR, Kleiman-Weiner M, Roberts DA, Voorhees EM, Soboroff I (2014) Evaluating stream filtering for entity profile updates in TREC 2012, 2013, and 2014. In: Voorhees EM, Ellis A (eds) Proceedings of the 23rd text retrieval conference, TREC 2014, Gaithersburg, Maryland, USA, 19–21 November 2014, National Institute of Standards and Technology (NIST), vol Special Publication, pp 500–308. http://trec.nist.gov/pubs/trec23/papers/overview-kba.pdf

Gey FC, Larson RR, Kando N, Machado J, Sakai T (2010) NTCIR-GeoTime overview: evaluating geographic and temporal search. In: Kando et al 2010, pp 147–153. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings8/NTCIR/01-NTCIR8-OV-GeoTime-GeyF.pdf

Gey FC, Larson RR, Machado J, Yoshioka M (2011) NTCIR9-GeoTime overview - evaluating geographic and temporal search: round 2. In: Kando et al 2011. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings9/NTCIR/01-NTCIR9-OV-GEOTIME-GeyF.pdf

Guo Q, Diaz F, Yom-Tov E (2013) Updating users about time critical events. In: Serdyukov P, Braslavski P, Kuznetsov S, Kamps J, Rüger S, Agichtein E, Segalovich I, Yilmaz E (eds) Advances in information retrieval. Lecture notes in computer science, vol 7814. Springer, Berlin, pp 483–494. https://doi.org/10.1007/978-3-642-36973-541

Hasanuzzaman M, Sze WL, Salim MP, Dias G (2016) Collective future orientation and stock markets. In: Proceedings of the 22nd European conference on artificial intelligence, ECAI'16, pp 1616–1617. IOS Press, Amsterdam. https://doi.org/10.3233/978-1-61499-672-9-1616

Joho H, Jatowt A, Blanco R (2013) A survey of temporal web search experience. In: Carr L, Laender AHF, Lóscio BF, King I, Fontoura M, Vrandecic D, Aroyo L, de Oliveira JPM, Lima F, Wilde E (eds) 22nd international world wide web conference, WWW'13, Rio de Janeiro, Brazil, 13–17 May 2013, Companion volume. International World Wide Web Conferences Steering Committee/ACM, pp 1101–1108. https://doi.org/10.1145/2487788.2488126

Joho H, Jatowt A, Blanco R, Naka H, Yamamoto S (2014) Overview of NTCIR-11 temporal information access (temporalia) task. In: Kando N, Joho H, Kishida K (eds) Proceedings of the 11th NTCIR conference on evaluation of information access technologies, NTCIR-11, National Center of Sciences, Tokyo, Japan, 9–12 December 2014, National Institute of Informatics (NII). http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings11/pdf/NTCIR/OVERVIEW/01-NTCIR11-OV-TEMPORALIA-JohoH.pdf

Joho H, Jatowt A, Blanco R (2015) Temporal information searching behaviour and strategies. Inf Process Manag 51(6):834–850. https://doi.org/10.1016/j.ipm.2015.03.006, http://www.sciencedirect.com/science/article/pii/S0306457315000448

Joho H, Jatowt A, Blanco R, Yu H, Yamamoto S (2016) Overview of NTCIR-12 temporal information access (temporalia-2) task. In: Kando N, Sakai T, Sanderson M (eds) Proceedings of the 12th NTCIR conference on evaluation of information access technologies, National Center of Sciences, Tokyo, Japan, 7–10 June 2016, National Institute of Informatics (NII). http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings12/pdf/ntcir/OVERVIEW/01-NTCIR12-OV-TEMPORALIA-JohoH.pdf

Kando N, Kishida K, Sugimoto M (eds) (2010) Proceedings of the 8th NTCIR workshop meeting on evaluation of information access technologies: information retrieval, question answering and cross-lingual information access, NTCIR-8, National Center of Sciences, Tokyo, Japan, 15–

18 June 2010, National Institute of Informatics (NII). http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings8/NTCIR/

Kando N, Ishikawa D, Sugimoto M (eds) (2011) Proceedings of the 9th NTCIR workshop meeting on evaluation of information access technologies: information retrieval, question answering and cross-lingual information access, NTCIR-9, National Center of Sciences, Tokyo, Japan, 6–9 December 2011, National Institute of Informatics (NII). http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings9/NTCIR/toc_ntcir.html

Kanhabua N, Blanco R, NØrvåg K, (2015) Temporal information retrieval. Found Trends Inf Retr 9(2):91–208. https://doi.org/10.1561/1500000043

Kazai G, Kamps J, Milic-Frayling N (2013) An analysis of human factors and label accuracy in crowdsourcing relevance judgments. Inf Retr 16(2):138–178. https://doi.org/10.1007/s10791-012-9205-0

Kishida K (2010) Vocabulary-based re-ranking for geographic and temporal searching at NTCIR GeoTime task. In: Kando et al 2010, pp 181–184. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings8/NTCIR/06-NTCIR8-GeoTime-KishidaK.pdf

Mani I, Pustejovsky J, Sundheim B (2004) Introduction to the special issue on temporal information processing. ACM Trans Asian Lang Inf Process (TALIP) 3(1):1–10. https://doi.org/10.1145/1017068.1017069

Matthews M, Tolchinsky P, Blanco R, Atserias J, Mika P, Zaragoza H (2010) Searching through time in the New York Times. In: HCIR 2010, pp 41–44. http://www.hugo-zaragoza.net/academic/pdf/matthews_HCIR2010.pdf

Mitamura T, Nyberg E, Shima H, Kato T, Mori T, Lin C, Song R, Lin C, Sakai T, Ji D, Kando N (2008) Overview of the NTCIR-7 ACLIA tasks: advanced cross-lingual information access. In: Kando N (ed) Proceedings of the 7th NTCIR workshop meeting on evaluation of information access technologies: information retrieval, question answering and cross-lingual information access, NTCIR-7, National Center of Sciences, Tokyo, Japan, 16–19 December 2008, National Institute of Informatics (NII). http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings7/pdf/NTCIR7/C1/CCLQA/01-NTCIR7-OV-CCLQA-MitamuraT.pdf

Mori T (2010) A method for GeoTime information retrieval based on question decomposition and question answering. In: Kando et al 2010, pp 167–172. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings8/NTCIR/04-NTCIR8-GeoTime-MoriT.pdf

Moulahi B, Tamine L, Yahia SB (2016) When time meets information retrieval: past proposals, current plans and future trends. J Inf Sci 42(6):725–747. https://doi.org/10.1177/0165551515607277

Purves R, Jones C (2004) Workshop on geographic information retrieval, SIGIR 2004. SIGIR Forum 38(2):53–56. https://doi.org/10.1145/1041394.1041406

Rizzo SG, Montesi D (2017) Quantification of time in digital libraries: temporal Zipf's law. In: Proceedings of the 21st international database engineering and applications symposium, IDEAS 2017, pp 143–152. ACM, New York. https://doi.org/10.1145/3105831.3105866

Sakai T, Shima H, Kando N, Song R, Lin C, Mitamura T, Sugimito M, Lee C (2010) Overview of NTCIR-8 ACLIA IR4QA. In: Kando et al 2010, pp 63–93. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings8/NTCIR/01-NTCIR8-OV-IR4QA-SakaiT.pdf

Sato T (2011) NTCIR-9 GeoTime at Osaka Kyoiku University - toward automatic extraction of place/time terms. In: Kando et al 2011. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings9/NTCIR/09-NTCIR9-GEOTIME-SatoT.pdf

Sekine S, Isahara H (2000) IREX: IR & IE evaluation project in Japanese. In: Proceedings of the 2nd international conference on language resources and evaluation (LREC'00), European Language Resources Association (ELRA), Athens, Greece. http://www.lrec-conf.org/proceedings/lrec2000/pdf/27.pdf

Strötgen J, Gertz M (2012) Event-centric search and exploration in document collections. In: Proceedings of the 12th ACM/IEEE-CS joint conference on digital libraries, JCDL'12, pp 223–232. ACM, New York. https://doi.org/10.1145/2232817.2232859

Strötgen J, Gertz M (2013) Proximity2-aware ranking for textual, temporal, and geographic queries. In: Proceedings of the 22nd ACM international conference on information and knowledge management, CIKM'13, pp 739–744. ACM, New York. https://doi.org/10.1145/2505515.2505640

Sundheim BM (1995) Overview of results of the MUC-6 evaluation. In: Proceedings of the 6th conference on message understanding, association for computational linguistics, MUC6'95, Stroudsburg, PA, USA, pp 13–31. https://doi.org/10.3115/1072399.1072402

Yoshioka M (2010) On a combination of probabilistic and boolean IR models for GeoTime task. In: Kando et al 2010, pp 154–158. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings8/NTCIR/02-NTCIR8-GeoTime-YoshiokaM.pdf

Chapter 10
SogouQ: The First Large-Scale Test Collection with Click Streams Used in a Shared-Task Evaluation

Ruihua Song, Min Zhang, Cheng Luo, Tetsuya Sakai, Yiqun Liu, and Zhicheng Dou

Abstract Search logs are very precious for information retrieval studies. In this chapter, we will introduce a real Chinese query log dataset, SogouQ, which was released by SogouQ corporation in 2010 for the NTCIR-9 Intent task. SogouQ contains more than 30 million clicks collected in 2008. It is the first large-scale query logs used in a shared-task evaluation (i.e., the NTCIR tasks). SogouQ has been adopted in a number of follow-up evaluation tasks, NTCIR-10 Intent-2, NTCIR-11 IMine, NTCIR-12 IMine-2, as well as in several Chinese domestic tasks. Moreover, SogouQ has a broader impact on other research areas, such as natural language processing and social science. It has been acquired by more than 200 institutions.

R. Song (✉)
Microsoft XiaoIce, Beijing 100080, China
e-mail: rsong@microsoft.com

M. Zhang · Y. Liu
Tsinghua University, Beijing 100084, China
e-mail: z-m@tsinghua.edu.cn

Y. Liu
e-mail: yiqunliu@tsinghua.edu.cn

C. Luo
MegaTech.AI, Beijing 100080, China
e-mail: luochengleo@gmail.com

T. Sakai
Waseda University, Shinjuku-ku Okubo 3-4-1 63-05-04, Tokyo 169-8555, Japan
e-mail: tetsuyasakai@acm.org

Z. Dou
Renmin University of China, Beijing 100872, China
e-mail: dou@ruc.edu.cn

T. Sakai et al. (eds.), *Evaluating Information Retrieval and Access Tasks*,
The Information Retrieval Series 43,
https://doi.org/10.1007/978-981-15-5554-1_10

10.1 Introduction

When we were preparing the NTCIR-9 Intent task that aims to investigate query intents and search result diversification (Song et al. 2011) in 2010, Sogou corporation was so generous to provide a real Chinese query log to NTCIR participants and further research communities. The data is called SogouQ and contains 30 million clicks collected in 2008. It is the first large-scale query logs used in a shared-task evaluation, such as NTCIR tasks.

The NTCIR-9 Intent task attracted 16 teams for Subtopic Mining subtask and 8 teams for Document Ranking subtask. It became the largest track in NTCIR-9 partially because participants are interested in SogouQ and how to use query logs for mining intents and diversifying document ranking. Since then SogouQ is used for NTCIR-10 Intent-2 task (Sakai et al. 2013), NTCIR-11 IMine task (Liu et al. 2014), and NTCIR-12 IMine-2 task (Yamamoto et al. 2016). The total number of participants groups is more than 80. They are from Australia, Canada, China, Germany, France, Japan, Korea, Spain, UK, and United States.

Later SogouQ had an even bigger impact on research. The usage of SogouQ data collection goes beyond the research on query intent. SogouQ is also used for improving fundamental natural language processing modules, such as name entity identification and new word discovery, user behavior studies, and Sociological topics. More than 200 institutes have acquired SogouQ related datasets from Tsinghua-Sohu Joint Laboratory on Search Technology. We believe that a more practical impact has happened but not been reported.

The remainder of this chapter is organized as follows: Sect. 10.2 describes the details of SogouQ and its related data collections. Section 10.3 briefly describes how organizers and participants use SogouQ in the NTCIR tasks. Section 10.4 reports more research impact beyond the works published in NTCIR proceedings. Section 10.5 concludes this chapter.

10.2 SogouQ and Related Data Collections

SogouQ was constructed by the Tsinghua-Sohu Joint Lab on Search Technology. It is a web query log of Sogou search engine for about one month (June 2008). There are about 30 million clicks included. The size of compressed SogouQ is about 1.9 gigabytes and is available for download.[1]

It should be noted that several similar click datasets were also released by several organizations for research purpose:

- **AOL Query logs** (2006/36M queries/English) includes user ids and click data. This dataset was intentional and intended for research purposes. However, the queries were not filtered and further lead to much controversy about privacy issues.

[1] http://www.sogou.com/labs/resource/q.php.

- **MSN Query logs** (2006/100M queries/English) includes session ids and click-through information, but not user ids (Craswell et al. 2009).
- **Yandex Query logs** (unknown time/210M queries/Russian) includes user sessions extracted from Yandex logs, with user ids, queries, query terms, URLs, their domains, URL rankings, and clicks. However, the user data is fully anonymized.[2]

The data format of SogouQ is as follows:

[Access time]\t[User ID]\t[Query]\t[Rank of the URL in the returned result]\t[The sequence number of user click]\t[URL that user clicked]\n

Here User ID is automatically assigned according to the cookie information when a user accesses the search engine by using the browser. Different queries that are input by the same browser correspond to the same user ID.

Compared to other search log data, SogouQ has several advantages. First, User ID and access time can provide information on sessions, which is important for session-based retrieval or mining-related searches by session. Second, in addition to the clicked URL, SogouQ provides the rank of clicked URL when it was shown to the user and which sequence the user clicked URLs for a query. Such information is valuable for research on user click modeling. Third, if we have only URLs, the content of URLs is difficult to obtain because the web keeps evolving. URLs may expire or the content of some URLs may change. Fortunately, Sogou released a document collection called SogouT[3] in 2010, which were crawled in June 2008. Therefore, researchers can get the corresponding page content at the same time.

We appreciate Sogou corporation and Tsinghua-Sohu Joint Lab of Search Technology. Due to their deep understanding of search and courage, research communities can have such valuable data collections.

10.3 SogouQ and NTCIR Tasks

The NTCIR-9 Intent task comprises the Subtopic Mining subtask (given a query, output a ranked list of possible subtopic strings) and the Document Ranking subtask (given a query, output a ranked list of URLs that are selectively diversified). In the Subtopic Mining subtask, a subtopic could be a specific interpretation of an ambiguous query (e.g., "microsoft windows" or "house windows" in response to "windows") or an aspect of a faceted query (e.g., "windows 7 update" in response to "windows 7"). The subtopics collected from participants were pooled, manually clustered, and thereby used as a basis for identifying the search intents of the query. The probability of each intent given the query was estimated through assessor voting. In the Document Ranking subtask, in contrast to traditional relevance assessments where the assessors determine the relevance of each pooled document with respect to a topic, we required the assessor to provide graded relevance assessments with

[2] https://www.kaggle.com/c/yandex-personalized-web-search-challenge/data.

[3] http://www.sogou.com/labs/resource/t.php.

respect to each intent of a given query. Finally, the relevance and diversity of the ranked subtopics or documents were evaluated using diversified information retrieval metrics (Sakai and Song 2014).

SogouQ was used by every participant for mining subtopics for given queries or estimating the importance of subtopics according to the number of clicks (Han et al. 2011; Wang et al. 2013; Xue et al. 2011; Yu and Ren 2014). The subtopics and their importance will influence document ranking then. Thus when user queries and clicks are introduced to the subtopic pool via SogouQ, our manually labeled intents or documents model the information needs of real users more accurately. Such an evaluation benchmark helps research on information retrieval in universities or labs without commercial search engines as experimental platforms.

In NTCIR-10 Intent-2 task, organizers provide the following instruction on subtopic:

> A subtopic string of a given query is a query that specialises and/or disambiguates the search intent of the original query. If a string returned in response to the query does neither, it is considered incorrect.
>
> e.g. original query: "harry potter" (underspecified) subtopic string: "harry potter philosophers stone movie" incorrect: "harry potter hp" (doe not specialise)
>
> It is encouraged that participants submit subtopics of the form "<originalquery> <additionalstring>"

Assessors were asked to provide a label for each intent cluster in the form "<originalquery><additionalstring>". Such a change provides valuable data to better understand a query in the perspective of two intent roles, i.e., kernel-object and modifier (Ren and Yu 2016; Yu and Ren 2012; Zheng et al. 2018). In contrast to the NTCIR-9 Intent task where we had up to 24 intents for a single topic, organizers of Intent-2 decided to select up to 9 intents per topic based on votes because search result diversification is mainly about diversifying the first search result page, which can only accommodate around ten URLs.

NTCIR-11 IMine task continued Subtopic Mining subtask and Document Ranking subtask and started a new subtask called TaskMine, which aims to explore the methods of automatically finding subtasks of a given task (e.g., for a given task "lose weight", the possible outputs can be "do physical exercise", "take calories intake", "take diet pills", etc.). In the Subtopic Mining subtask, participants are expected to generate a two-level hierarchy of underlying subtopics by analysis into the provided document collection, user behavior data including SogouQ, or other kinds of external data sources. For example, given the ambiguous query "windows", the first-level subtopic may be "microsoft windows", "software on windows platform", or "house windows". In the category of "microsoft windows", users may be interested in different aspects (second-level subtopics), such as "windows 8" and "windows update". The hierarchical structure of subtopics is closely related with the knowledge graph. However, the hierarchical subtopics here are used to describe users' possible information needs instead of the manually created knowledge structure of entity names. Organizers encouraged participants not to use the graph directly even when a knowledge graph exists for a given query. Therefore, user behavior data, such as SogouQ,

play important roles in creating the hierarchy of subtopics as real user queries reflect users' possible information needs.

NTCIR-12 IMine-2 task focuses on vertical intents behind a query as well as its topical intents because many commercial Web search engines merge several types of search results and generate a SERP (search engine results page) in response to a user's query. For example, the results of query "flower" now may contain image results and encyclopedia results as well as usual Web search results. We refer to such "types" of search results as verticals. Accordingly, the IMine-2 task comprises two subtasks: the Query Understanding subtask and the Vertical Incorporating subtask. The Query Understanding subtask is a successive task of the Subtopic Mining subtask but the difference is that participants are asked to identify the relevant verticals for each subtopic. For example, for the query "iPhone 6", a possible result list of the Query Understanding subtask is:

[tid] [subtopic] [vertical] [score]
IMINE2-E-000 iPhone 6 apple.com Web 0.9
IMINE2-E-000 iPhone 6 sales News 0.90
IMINE2-E-000 iPHone 6 photo Image 0.88
IMINE2-E-000 iPhone 6 review Web 0.78

The Vertical Incorporating subtask is also a successive task of the Document Ranking subtask. The difference is that the participants should decide whether the result list should contain vertical result or not. SogouQ is still a useful resource of user behaviors for Chinese subtasks. Similarly, Yahoo! Japan provides the participants of Japanese subtasks a Web search related query data, which is generated from the query log of Yahoo! Japan Search from July 2009 to June 2013.[4]

10.4 Impact of SogouQ

As by April 30, 2019, we can find 82 papers when we search the keyword "SogouQ" in Google Scholar.[5] Most of them are not published in NTCIR proceedings.

Some works such as Gu et al. (2016), Han et al. (2011), Ren et al. (2015), Xue et al. (2011), Kim and Lee (2015), and Zheng et al. (2015) use SogouQ to mine subtopics (Song et al. 2018; Wang et al. 2013; Yu and Ren 2014), or suggestions (Li and Wang 2014; Liu et al. 2017; Shu et al. 2013). Some works like Zheng et al. (2018) use SogouQ for better understanding a query in the perspective of two intent roles, i.e., kernel-object and modifier (Ren and Yu 2016; Yu and Ren 2012). Some other works investigate intent shifting (Wang and Chen 2011), query specification (Xiangbin et al. 2015), and search task identification (Du et al. 2018). Some works use SogouQ for improving some fundamental modules of natural language processing, such as unsupervised dependency parsing (Qiao et al. 2016), new word identification

[4] http://research.nii.ac.jp/ntcir/news-20150717-ja.html.

[5] http://scholar.google.com.

(Xuewei 2014), and person name recognition (Lv et al. 2013; Wen et al. 2013). Moreover, the rich information of SogouQ provides evidence to get statistics, e.g., query per second (Fang et al. 2017), sample queries (Liu and Li 2014); or mine a particular type of queries, e.g., time-sensitive search queries (Pei et al. 2016) and health search queries; or predict authoritative of website (Yu and Ren 2018).

Some usage of SogouQ is on broader research topics. Rao et al. (2014) constructs query co-occurrence network from SogouQ and compares the network with Named Entity Person co-occurrence network and the network based on the co-occurrence of words in sentences of news articles; Wang and Pleimling (2017) use it to investigate foraging patterns in online searches. Authors analyze three different click-through logs and discover an increased efficiency of the search engines. In the language of foraging, the newer logs indicate that online searches overwhelmingly yield local searches (i.e., on one page of links provided by the search engines), whereas for the older logs, the foraging processes are a combination of local searches and relocation phases that are power law distributed. It follows that good search engines enable the users to find the information they are looking for through a local exploration of a single page with search results, whereas for poor search engine, users are often forced to do a broader exploration of different pages.

According to the statistics from Tsinghua-Sohu Joint Lab on Search Technology, more than 200 institutions have acquired SogouQ related datasets. We believe that a more practical impact has happened but not been reported.

10.5 Conclusion

The problems that are explored in NTCIR Intent and IMine tasks require a data collection of query logs. With the great support of Sogou corporation, SogouQ becomes the first query logs that are used in a shared evaluation. Compared to other query logs, SogouQ has richer information on session, ranking, and orders of clicks, and corresponding documents if being combined with SogouT. Therefore, SogouQ does not only support research on query understanding of intent and vertical, but also enable many works on broader research topics on web search user behaviors. More than 200 institutes have acquired SogouQ data and they are using the query logs for various research and applications.

As query logs are too sensitive, it is difficult to obtain more shared query logs. Some efforts were done to simulate click-through data, such as Sogou-QCL (Zheng et al. 2018), to enable the neural-based works that need a larger amount of data.

References

Craswell N, Jones R, Dupret G, Viegas E (eds) (2009) Proceedings of the 2009 workshop on web search click data (WSCD'09). ACM, New York, NY, USA

Du C, Shu P, Li Y (2018) CA-LSTM: search task identification with context attention based LSTM. In: The 41st international ACM SIGIR conference on research & development in information retrieval. ACM, pp 1101–1104

Fang Z, Yu T, Mengshoel OJ, Gupta RK (2017) Qos-aware scheduling of heterogeneous servers for inference in deep neural networks. In: Proceedings of the 2017 ACM on conference on information and knowledge management. ACM, pp 2067–2070

Gu J, Feng C, Gao X, Wang Y, Huang H (2016) Query intent detection based on clustering of phrase embedding. In: Chinese national conference on social media processing. Springer, pp 110–122

Han J, Wang Q, Orii N, Dou Z, Sakai T, Song R (2011) Microsoft research Asia at the NTCIR-9 INTENT task. In: Proceedings of NTCIR-9

Kim SJ, Lee JH (2015) Subtopic mining using simple patterns and hierarchical structure of subtopic candidates from web documents. Inf Process Manag 51(6):773–785

Li L, Wang H (2014) Multi-strategy query expansion method based on semantics. J Digit Inf Manag 12(3)

Liu C, Li Y, (2014) Non-iteration parallel algorithm for frequent pattern discovery. In: 2014 13th international symposium on distributed computing and applications to business, engineering and science. IEEE, pp 127–132

Liu J, Li Q, Lin Y, Li Y (2017) A query suggestion method based on random walk and topic concepts. In: 2017 IEEE/ACIS 16th international conference on computer and information science (ICIS). IEEE, pp 251–256

Liu Y, Song R, Zhang M, Dou Z, Yamamoto T, Kato MP, Ohshima H, Zhou K (2014) Overview of the NTCIR-11 IMine task. In: Proceedings of NTCIR-11

Lv X, Wu R, Wen B (2013) Chinese personal name recognition in web queries via bootstrapping. In: 2013 9th international conference on computational intelligence and security. IEEE, pp 415–419

Pei J, Huang D, Ma J, Song D, Sang L (2016) Dut-nlp-ch@ NTCIR-12 temporalia tid subtask. In: Proceedings of NTCIR-12

Qiao X, Cao H, Zhao T (2016) Improving unsupervised dependency parsing with knowledge from query logs. ACM Trans Asian Low-Resour Lang Inf Process (TALLIP) 16(1):3

Rao L, Luo Z, Tang J, Wang T (2014) Research on the query co-occurrence networks. Management innovation and information technology, vol 61, pp. 275

Ren F, Yu H (2016) Role-explicit query extraction and utilization for quantifying user intents. Inf Sci 329:568–580

Ren P, Chen Z, Ma J, Wang S, Zhang Z, Ren Z (2015) Mining and ranking users' intents behind queries. Inf Retr J 18(6):504–529

Sakai T, Song R (2014) Evaluating diversified search results using per-intent graded relevance. In: ACM SIGIR 2011

Sakai T, Dou Z, Yamamoto T, Liu Y, Zhang M, Song R, Kato M, Iwata M (2013) Overview of the NTCIR-10 INTENT-2 task. In: Proceedings of NTCIR-10

Shu B, Niu Z, Jiang X, Mustafa G (2013) A novel query suggestion method based on sequence similarity and transition probability. In: Proceedings of the international conference on data mining (DMIN), The Steering Committee of The World Congress in Computer Science, p 1

Song R, Zhang M, Sakai T, Kato MP, Liu Y, Sugimoto M, Wang Q, Orii N (2011) Overview of the NTCIR-9 INTENT task. In: Proceedings of NTCIR-9

Song W, Liu Y, Liu Lz, Wang Hs (2018) Semantic composition of distributed representations for query subtopic mining. Front Inf Technol Electron Eng 19(11):1409–1419

Wang CJ, Chen HH (2011) Intent shift detection using search query logs. Int J Comput Linguist Chin Lang Process 16(3–4)

Wang Q, Qian Y, Song R, Dou Z, Zhang F, Sakai T, Zheng Q (2013) Mining subtopics from text fragments for a web query. Inf Retr 16(4):484–503

Wang X, Pleimling M (2017) Foraging patterns in online searches. Phys Rev E 95(3):032145
Wen B, Xiao S, Luo Y, LV X, (2013) Unsupervised Chinese personal name recognition using search session. J Comput Inf Syst 9(6):2201–2208
Xiangbin T, Wei L, Xiaojuan Z, Shihao H (2015) Feature analysis and automatic identification of query specificity. Data Anal Knowl Discov 31(2):15–23
Xue Y, Chen F, Zhu T, Wang C, Li Z, Liu Y, Zhang M, Jin Y, Ma S (2011) THUIR at NTCIR-9 INTENT task. In: Proceedings of NTCIR-9
Xuewei L, Xueqiang L, Kehui L (2014) Chinese new words identification from query log by extending the context. Data Anal Knowl Discov 30(11):59–65
Yamamoto T, Liu Y, Zhang M, Dou Z, Zhou K, Markov I, Kato MP, Ohshima H, Fujita S (2016) Overview of the NTCIR-12 IMine-2 task. In: Proceedings of NTCIR-12
Yang H, Feng Y (2018) Authoritative prediction of website based on deep learning. In: 2018 IEEE 4th international conference on big data computing service and applications (BigDataService). IEEE, pp 208–212
Yu H, Ren F (2012) Role-explicit query identification and intent role annotation. In: Proceedings of the 21st ACM international conference on information and knowledge management. ACM, pp 1163–1172
Yu HT, Ren F (2014) Subtopic mining via modifier graph clustering. In: Pacific-Asia conference on knowledge discovery and data mining. Springer, pp 337–347
Zhang X, Han S, Lu W (2018) Automatic prediction of news intent for search queries: an exploration of contextual and temporal features. Electron Libr 36(5):938–958
Zhang Z, Sun L, Han X (2015) Learning to mine query subtopics from query log. In: Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 2: short papers), vol 2, pp 341–345
Zheng Y, Fan Z, Liu Y, Luo C, Zhang M, Ma S (2018) Sogou-QCL: a new dataset with click relevance label. In: The 41st international ACM SIGIR conference on research & development in information retrieval. ACM, pp 1117–1120

Chapter 11
Evaluation of Information Access with Smartphones

Makoto P. Kato

Abstract NTCIR 1CLICK and MobileClick are the earliest attempts toward test-collection-based evaluation for information access with smartphones. Those campaigns aimed to develop an IR system that outputs a short text summary for a given query, which is expected to fit a small screen and to satisfy users' information needs without requiring much interaction. The textual output was evaluated on the basis of *iUnits*, pieces of relevant text for a given query, with consideration of users' reading behaviors. This chapter begins with an introduction to NTCIR 1CLICK and MobileClick, explains the evaluation methodology and metrics such as S-measure and M-measure, and finally discusses the potential impacts of those evaluation campaigns.

11.1 Introduction

In 2015, Google announced that more searches took place on mobile devices than on desktop computers in 10 countries including the US and Japan.[1] Among diverse types of mobile devices, the smartphone has become dominant according to a survey in 2015.[2] Thus, there is no doubt that the smartphone is one of the most important search environments for which search engines should be designed, due to its popularity and several differences from traditional devices, e.g., desktop computers.

The search experience difference between desktop computers and smartphones mainly comes from the differences in screen size, internet connection, interaction, and situation. A relatively small screen size limits the amount of content which the users can read at a time. The internet connection is sometimes unstable depending on

[1] https://adwords.googleblog.com/2015/05/building-for-next-moment.html.
[2] https://www.pewglobal.org/2016/02/22/smartphone-ownership-and-internet-usage-continues-to-climb-in-emerging-economies/.

M. P. Kato (✉)
University of Tsukuba, 1-2 Kasuga, Tsukuba, Ibaraki 305-8550, Japan
e-mail: mpkato@acm.org

T. Sakai et al. (eds.), *Evaluating Information Retrieval and Access Tasks*,
The Information Retrieval Series 43,
https://doi.org/10.1007/978-981-15-5554-1_11

where users conduct search. While the keyboard and mouse are typical input devices for desktop computers, touch interaction and speech input are often used for smartphones and may not be suitable for inputting or editing many keywords. Search with smartphones can sometimes be interrupted by the other activities with which the user is engaged simultaneously. To overcome the limitations in search with smartphones, research communities have studied new designs of interface, interaction, and search algorithms suitable for smartphones (Crestani et al. 2017).

NTCIR 1CLICK (Kato et al. 2013a; Sakai et al. 2011b) and MobileClick (Kato et al. 2014, 2016b) are the earliest attempts toward test-collection-based evaluation for information access with smartphones. Those campaigns aimed to develop an IR system that outputs a short text summary for a given query, which is expected to fit a small screen and to satisfy users' information needs without requiring much interaction. The textual output was evaluated on the basis of pieces of relevant text for a given query. The basic task design is similar to *query-biased multi-document summarization* (Carbonell and Goldstein 1998; Tombros and Sanderson 1998), in which a system is expected to generate a summary of a fixed length from multiple documents, satisfying the information need of users who input a certain query. The main difference from the query-biased multi-document summarization task is *position awareness* of presented information. In the NTCIR 1CLICK and MobileClick tasks, more important information is expected to be present at the beginning of the summary so that users can reach such information efficiently. In other words, more relevant information pieces should be ranked at higher positions like an ad hoc retrieval task. Accordingly, evaluation measures used in these tasks were designed to be position-aware, unlike those for text summarization such as recall, precision, and ROUGE (Lin 2004). This task design and evaluation methodology distinguishes NTCIR 1CLICK and MobileClick from the other summarization tasks, and had some impact on mobile information access and related fields.

This chapter first describes the task design of NTCIR 1CLICK and MobileClick, introduces evaluation methodologies used in these campaigns, and finally discusses potential impacts on works published after NTCIR 1CLICK and MobileClick.

11.2 NTCIR Tasks for Information Access with Smartphones

This section provides a brief overview of the task design of the NTCIR 1CLICK and MobileClick tasks. Table 11.1 summarizes four NTCIR tasks to be described in this section.[3]

[3] S♯-measure is a combination of S-measure and T-measure (a precision-like metric) (Sakai and Kato 2012).

Table 11.1 NTCIR tasks for information access with smartphones

Year	NTCIR	Task	Subtasks	Primary metric
2010	9	1CLICK-1		S-measure
2011	10	1CLICK-2	Main & Query classification	S♯-measure[3]
2013	11	MobileClick-1	iUnit retrieval & iUnit summarization	M-measure
2014	12	MobileClick-2	iUnit retrieval & iUnit summarization	M-measure

11.2.1 NTCIR 1CLICK

The history of information access with smartphones in NTCIR began from a subtask of the NTCIR-9 INTENT task, namely, NTCIR-9 1CLICK-1 (formally, *one-click access task*) (Sakai et al. 2011b). While the NTCIR-9 INTENT task targeted search result diversification, the NTCIR-9 1CLICK-1 task focused especially on generating a query-biased summary as a proxy for a search engine result page (or "ten blue links"), for satisfying the user immediately after the user clicks on the search button. Strictly speaking, the NTCIR-9 1CLICK-1 task was designed not for information access with smartphones, but for *Direct and Immediate Information Access*, which was defined in earlier work by the task organizers (Sakai et al. 2011a):

> We define Direct Information Access as a type of information access where there is no user operation such as clicking or scrolling between the user's click on the search button and the user's information acquisition; we define Immediate Information Access as a type of information access where the user can locate the relevant information within the system output very quickly. Hence, a Direct and Immediate Information Access (DIIA) system is expected to satisfy the user's information need very quickly with its very first response.

While the NTCIR-9 1CLICK-1 task was treated as a pilot task and targeted only the Japanese language, the 1CLICK-2 task was organized as an independent task at NTCIR-10 and employed almost the same task design as that of the NTCIR-9 1CLICK-1 task, with the scope extended to Japanese and English.

At both 1CLICK-1 and 1CLICK-2, participants were given a list of queries categorized into four query categories, namely, celebrity, local, definition, and Q&A. The task organizers selected these categories the following work by Li et al. (2009), which investigated Google's desktop and mobile query logs of three countries, and identified frequent query types for *good abandonment*—an abandoned query for which the user's information need was successfully addressed by the search engine result page without clicks or query reformulation.

NTCIR-9 1CLICK-1 and NTCIR-10 1CLICK-2 participants were expected to produce a plain text of X characters for each query ($X = 140$ for Japanese and $X =$

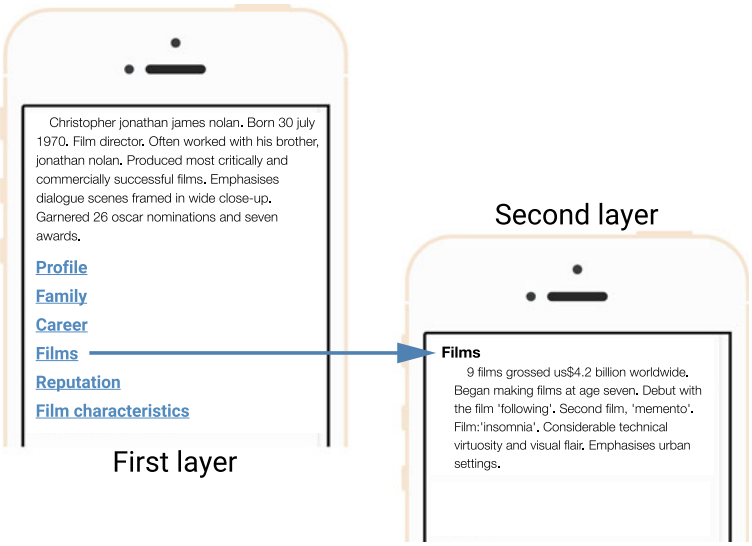


Fig. 11.1 A two-layered summary for query "`christopher nolan`". Users can see the second layer if they click on a link in the first layer

280 for English),[4] based on a given document collection. The output was expected to include important pieces of information first and to minimize the amount of text the user has to read. These requirements are more formally described through the evaluation metrics explained in Sect. 11.3.

11.2.2 NTCIR MobileClick

NTCIR MobileClick, which started from NTCIR-11, took over the spirit of NTCIR 1CLICK, and aimed to directly return a summary of relevant information and immediately satisfy the user *without requiring much interaction*. Unlike the 1CLICK tasks, participants were expected to produce a two-layered summary that consists of a single first layer and multiple second layers, as shown in Fig. 11.1. The first layer is expected to contain information interesting for most of the users, and the links to the second layer; the second layer, which is hidden until its header link is clicked on, is expected to contain information relevant for a particular type of users. In a two-layered summary, users can avoid reading text in which they are not interested, thus saving time spent on non-relevant information, if they can make a binary yes/no decision of each second-layer entry from the head link alone.

[4] Both NTCIR-9 1CLICK-1 and NTCIR-10 1CLICK-2 accepted two types of runs, namely, DESKTOP and MOBILE runs. In this chapter, only MOBILE runs are explained for simplicity.

This unique output was motivated by the discussion at the NTCIR-10 conference in June 2013, and reflected the rapid growth of smartphone users in those years. Although 1CLICK expects no interaction except for clicking on the search button, MobileClick targeted smartphone users and expects users to tap on some links for browsing desired information efficiently.

NTCIR MobileClick assumed different types of users who are interested in different topics. The diversity of users who input a certain query was modeled by *intent probability*, which is the probability over intents for the query. For example, among users who input "apple" as a query, 90% are interested in Apple Inc. and 10% are interested in apple the fruit. A two-layered summary is considered *good* if different types of users are all satisfied with the summary. Thus, the first layer should not contain information in which a particular type of users are interested, while the second layers should not contain information relevant to the majority of users.

The input in the NTCIR-11 MobileClick-1 and NTCIR-12 MobileClick-2 tasks was a list of queries that were basically categorized into four types mentioned earlier. There were two subtasks in these evaluation campaigns: iUnit retrieval and iUnit summarization subtasks. In iUnit retrieval subtask, participants were expected to output a ranked list of information pieces called iUnit in response to a given query. In iUnit summarization subtask, as was explained earlier, the output was a two-layered summary in XML format. While the NTCIR-11 MobileClick-1 required participants to identify information pieces from a document collection, the NTCIR-12 MobileClick-2 only required selecting and ranking or arranging predefined information pieces, mainly for increasing the reusability of the test collection.

11.3 Evaluation Methodology in NTCIR 1CLICK and MobileClick

This section explains and discusses some details of the evaluation methodology used in the NTCIR 1CLICK and MobileClick tasks, which is mainly based on nuggets, or pieces of information we call *iUnits*. We first present the background and explain the differences between summarization and our tasks. We then focus on the notions of nuggets and iUnits, and finally discuss the effectiveness metrics developed and used in the NTCIR tasks.

11.3.1 Textual Output Evaluation

Summarization is one of the most similar tasks to NTCIR 1CLICK and MobileClick. As mentioned earlier, the most notable difference between the summarization and these NTCIR tasks is position awareness of information pieces in the textual output. This subsection details and discusses the difference in terms of the evaluation methodology.

Automatic evaluation of machine-generated summaries has been often conducted by comparison with human-generated summaries (Nenkova and McKeown 2011). ROUGE is a widely used evaluation metric based on word matching between a machine summary and human summaries (Lin 2004). There are several variants of ROUGE such as ROUGE-W (n-gram matching), ROUGE-L (longest common sequence), and ROUGE-S (skip-gram matching). Although these variants are sensitive to the order of words, they are agnostic to the absolute position where each word appears in a machine summary. The Pyramid method identifies Summary Content Units (SCUs), which are word spans expressing the same meaning, from multiple human summaries, and computes a score for each machine summary based on the included SCUs (Nenkova et al. 2007). The weight of an SCU is determined by the number of human summaries including the SCU, and a summary is scored basically by the sum of the weights of SCUs within the summary. The position of SCUs within a machine summary does not affect the score.

The insensitivity for the position of information pieces (*i.e.,* words or SCUs) is reasonable when it is assumed that the whole summary is always read. In such a case, the position of information pieces should not affect the utility of the summary, as all the information pieces are equally consumed by the reader.

On the other hand, the position matters when users may read different parts of a summary. As the textual output in NTCIR 1CLICK is expected to be scanned from top to bottom, like Web search, contents near the end have a smaller probability to be read, and, accordingly, should be discounted when the utility is estimated. The two-layer summary in NTCIR MobileClick can be read in many different ways. A user may read only the first layer, while another user may scan contents in the first layer from top, click on a link interesting for the user, read a second layer shown by the click, and stop reading at the end of the second layer. Therefore, the primary difference from ordinary summarization tasks is how the summary is expected to be read, which naturally required different evaluation methodologies.

11.3.2 From Nuggets to iUnits

The NTCIR-9 1CLICK-1 task evaluated the system output based on nuggets. Nuggets are fragments of text, which were frequently used in summarization and question answering evaluation. TREC Question Answering track defined *an information nugget* as "a fact for which the assessor could make a binary decision as to whether a response contained the nugget" (Voorhees 2003). The possibility of the binary decision is called *atomicity* (Dang et al. 2007). As explained earlier, the Pyramid method (Nenkova et al. 2007) uses SCUs as units of comparison:

> SCUs are semantically motivated, subsentential units; they are variable in length but not bigger than a sentential clause. This variability is intentional since the same information may be conveyed in a single word or a longer phrase. SCUs emerge from annotation of a

> collection of human summaries for the same input. They are identified by noting information that is repeated across summaries, whether the repetition is as small as a modifier of a noun phrase or as large as a clause.

Babko-Malaya described a systematic way to uniform the granularity of nuggets based on several nuggetization rules (Babko-Malaya 2008). Examples of the rules are shown below:

> Nuggets are created out of each core verb and its arguments, where the maximal extent of the argument is always selected.

> Noun phrases are not decomposed into separate nuggets, unless they contain temporal, locative, numerical information, or titles.

Basic elements are another attempt to systematically define nuggets (Hovy et al. 2006), and were defined as follows:

> the head of a major syntactic constituent (noun, verb, adjective or adverbial phrases), expressed as a single item, or a relation between a head-BE and a single dependent, expressed as a triple (head—modifier—relation).

Although several attempts had been made to standardize the nuggetization procedure, the task organizers of NTCIR 1CLICK still found it hard to identify nuggets. The primary difficulty is to uniform the granularity of nuggets. While the notion of atomicity determines the unit of nuggets to some extent, there were some cases in which assessors disagreed. Typical examples are shown below:

1. Tetsuya Sakai was born in 1988.
2. Takehiro Yamamoto received a PhD from Kyoto University in 2011.

The following pieces are candidates for nuggets in sentences 1 and 2.

1-A. Tetsuya Sakai was born in 1988.
1-B. Tetsuya Sakai was born.
2-A. Takehiro Yamamoto received a PhD from Kyoto University in 2011.
2-B. Takehiro Yamamoto received a PhD in 2011.
2-C. Takehiro Yamamoto received a PhD from Kyoto University.
2-D. Takehiro Yamamoto received a PhD.

Although 1-B and 2-D are results of a similar type of decomposition, 1-B does not look appropriate for a nugget, but 2-D does. Whereas, 2-D may not be an appropriate nugget if the query is "When did Takehiro Yamamoto receive his PhD?" since 2-D can be a trivial fact like 1-B. A systematic approach may not be very helpful in this case.

Another difficulty is the way to determine the weight of nuggets. Unlike the Pyramid method and others, the NTCIR-9 1CLICK-1 task extracted nuggets from a document collection from which the textual output is generated, not from those generated by human assessors. This methodology was chosen because there were hundreds of nuggets for some queries, which cannot be included in a few human-generated summaries. The weighting schema used in the Pyramid method cannot

be simply applied to this case, as the number of assessors who found a nugget may simply reflect the frequency of the nugget in the collection, but it might be unrelated to the importance of the nugget. Furthermore, the dependency of nuggets makes the problem more complicated. For example, 2-B entails 2-D. Then, what is the score of a summary including 2-B? Is it the sum of the weights of 2-B and 2-D, or 2-B's alone?

To clarify the definition of nuggets and weighting schema in NTCIR 1CLICK, the task organizers of the NTCIR-10 1CLICK-2 opted to redefine nuggets and call them *information units* or *iUnits*.

iUnits satisfy three properties, *relevant*, *atomic*, and *dependent*, described in detail below. *Relevant* means that an iUnit provides useful factual information to the user on its own. Thus, it does not require other iUnits to be present in order to provide useful information. For example:

1. Tetsuya Sakai was born in 1988.
2. Tetsuya Sakai was born.

If the information need is "Who is Tetsuya Sakai?", (2) alone is probably not useful, and therefore this is not an iUnit. Note that this property emphasizes that the information need determines which pieces of information are iUnits. If the information need is "Where was Tetsuya Sakai born?", both cannot be iUnits.

Atomic means that an iUnit cannot be broken down into multiple iUnits without loss of the original semantics. Thus, if it is broken down into several statements, at least one of them does not pass the relevance test. For example:

1. Takehiro Yamamoto received a PhD from Kyoto University in 2011.
2. Takehiro Yamamoto received a PhD in 2011.
3. Takehiro Yamamoto received a PhD from Kyoto University.
4. Takehiro Yamamoto received a PhD.

(1) can be broken down into (2) and (3), and both (2) and (3) are *relevant* to the information need "Who is Takehiro Yamamoto?". Thus, (1) cannot be an iUnit, but (2) and (3) are iUnits. (2) can be further broken down into (4) and "Takehiro Yamamoto received something in 2011". However, the latter does not convey useful information for the information need. The same goes for (3). Therefore, (2) and (3) are valid iUnits and (4) is also an iUnit.

Dependent means that an iUnit can entail other iUnits. For example:

1. Takehiro Yamamoto received a PhD in 2011.
2. Takehiro Yamamoto received a PhD.

(1) entails (2) and they are both iUnits.

In the NTCIR-10 1CLICK-2, nuggets were first identified from a document collection, and iUnits were extracted from the nuggets.[5] A set of iUnits for query 1C2-J-0001 "倉木麻衣 (Mai Kuraki; a Japanese singer-songwriter)" is shown in Table 11.3,

[5]This approach was taken mainly for increasing the efficiency by dividing the iUnit extraction task into two parts.

Table 11.2 Nuggets for query 1C2-J-0001 "倉木麻衣 (Mai Kuraki; a Japanese singer-songwriter)"

ID	Nugget
S005	99.10月、16歳で"Mai K"名義の『Baby I Like』で全米デビュー。同年12月8日『Love, Day After Tomorrow』で倉木麻衣、日本デビュー。 (She made her American debut with "Baby I Like" as "Mai K" in October 1999, when she was 16 years old. In the same year, on December 8, she made her debut in Japan with "Love, Day After Tomorrow".)
S008	血液型 B型 (Blood type: B)
S012	職業 歌手 (Occupation: Singer)
S022	2005年立命館大学を卒業 (She graduated from Ritsumeikan University in 2005.)
S023	第15回日本ゴールドディスク大賞で『delicious way』がロック・アルバム・オブ・ザ・イヤーを、「Secret of my heart」がソング・オブ・ザ・イヤーを受賞。 ("delicious way" won "Rock album of the Year" and "Secret of my heart" won "Song of the Year" at the 15th annual Japan Gold Disc Awards)

which were extracted from nuggets in Table 11.2. The column "Entails" indicates a list of iUnits that are entailed by the iUnit. For example, iUnit I014 entails I013, and iUnit I085 entails iUnits I023 and I033. A *semantics* is the factual statement that the iUnit conveys. This is used by assessors to determine whether an iUnit is present in a summary.

A *vital string* is a minimally adequate natural language expression and extracted from iUnits. This approximates the minimal string length required so that the user who issued a particular query can read and understand the conveyed information. The vital string of iUnit u that entails iUnits $e(u)$ does not include that of iUnits $e(u)$ to avoid duplication of vital strings, since if iUnit u is present in a summary, iUnits $e(u)$ are also present by definition. For example, the vital string of iUnit I014 does not include that of iUnit I013 as shown in Table 11.3. Even the vital string of I085 is empty as it entails iUnits I023 and I033.

Having extracted iUnits from nuggets, assessors gave the weight to each iUnit on five-point scale (very low (1), low (2), medium (3), high (4), and very high (5)). iUnits were randomly ordered and their entailment relationship was hidden during the voting process. After the voting, we revised iUnit's weight so that iUnit u entailing iUnits $e(u)$ receives the weight of only u excluding that of $e(u)$. This revision is necessary because the presence of iUnit u in a summary entails that of iUnits $e(u)$, resulting in duplicative counting of the weight of $e(u)$ when we take into account the weight of both u and $e(u)$.

For example, suppose that there are only four iUnits:

1. Ichiro was a batting champion (3).
2. Ichiro was a stolen base champion (3).
3. Ichiro was a batting and stolen base champion (7).
4. Ichiro was the first player to be a batting and stolen base champion since Jackie Robinson in 1949 (8).

Table 11.3 iUnits for query 1C2-J-0001 "倉木麻衣 (Mai Kuraki; a Japanese singer-songwriter)"

ID	Entails	Nugget	Semantics	Vital string
I011		S005	1999年日本デビュー (Made her Japanese debut in 1999)	1999年日本デビュー (Made her Japanese debut in 1999)
I012		S008	血液型B型 (Blood type: B)	血液型B型 (Blood type: B)
I013		S022	立命館大卒 (Graduated from Ritsumeikan University)	立命館大卒 (Graduated from Ritsumeikan University)
I014	I013	S022	2005年立命館大卒 (Graduated from Ritsumeikan University in 2005)	2005年 (2005)
I017		S012	職業 歌手 (Occupation: Singer)	歌手 (Singer)
I023		S023	第15回日本ゴールドディスク大賞ソング・オブ・ザ・イヤー受賞 (Won "Song of the Year" at the 15th annual Japan Gold Disc Awards)	第15回日本ゴールドディスク大賞ソング・オブ・ザ・イヤー受賞 (Won "Song of the Year" at the 15th annual Japan Gold Disc Awards)
I033		S023	シングル Secret of my heart (Single "Secret of my heart")	シングル Secret of my heart (Single "Secret of my heart")
I085	I023, I033	S023	第15回日本ゴールドディスク大賞で『Secret of my heart』がソング・オブ・ザ・イヤーを受賞 ("Secret of my heart" won "Song of the Year" at the 15th annual Japan Gold Disc Awards)	

where (4) entails (3), and (3) entails both (1) and (2). A parenthesized value indicates the weight of each iUnit. Suppose that a summary contains (4). In this case, the summary also contains (1), (2), and (3) by definition. If we just sum up the weight of iUnits in the summary, the result is $21(= 3 + 3 + 7 + 8)$, where the weight of (1) and (2) is counted three times and that of (3) is counted twice. Therefore, it is necessary to subtract the weight of entailing iUnits to avoid the duplication; in this example, thus, the weight of iUnits becomes 3, 3, $4(= 7 - 3)$, and $1(= 8 - 7)$, respectively.

More formally, we used the following equation for revising the weight of iUnit u:

$$w(u) - \max_{u' \in e(u)} w(u'), \tag{11.1}$$

where $w(u)$ is the weight of iUnit u. Note that iUnits $e(u)$ in the equation above are ones entailed by iUnit u and the entailment is *transitive*, *i.e.* if i entails j and j entails k, then i entails k.

11.3.3 S-Measure

S-measure (Sakai et al. 2011a) was the primary evaluation metric at NTCIR-9 1CLICK-1 and NTCIR-10 1CLICK-2. Letting M be a set of iUnits identified in a summary, *S*-measure is defined as

$$S\text{−measure} = \frac{1}{\mathcal{N}} \sum_{u \in M} w(u) \max(0, 1 - \text{offset}(u)/L), \tag{11.2}$$

where $\mathcal{N}$ is a normalization factor, $w(u)$ is the weight of an iUnit u, L is a *patience* parameter, and $\text{offset}(u)$ is the offset of an iUnit u in the summary (more precisely, it is the number of characters between the beginning of the summary and the end of the iUnit). This measure basically represents the sum of the weight ($w(u)$) with offset-based decay ($1 - \text{offset}(u)/L$) for iUnits in a summary. Figure 11.2 illustrates S-measure computation with a simple example. As shown in the figure, the decay is assumed to decrease linearly with respect to the offset of an iUnit, and totally cancels the value of an iUnit appearing after L characters (the maximum function simply prevents the decay from being negative). Thus, the patience parameter can be interpreted as how many characters can be read by the user, or, alternatively, how much time the user can spend to read the summary when it is divided by the reading speed. For example, $L = 500$ in Fig. 11.2. If the reading speed is 500 characters per minute for average Japanese users, this patience parameter indicates that the user spends only a minute and leaves right after a minute passes. This corresponds to the fact that the decay factor becomes zero (or no value) after 500 characters.

The normalization factor $\mathcal{N}$ sets the upper bound so that S ranges from 0 to 1, and is defined as

$$\mathcal{N} = \sum_{u \in U} w(u) \max(0, 1 - \text{offset}^*(v(u))/L), \tag{11.3}$$

where U is a set of all iUnits and $\text{offset}^*(v(u))$ is the offset of the vital string of an iUnit u in *Pseudo Minimal Output* (PMO), which is an ideal summary artificially created for estimating the upper bound. The PMO was obtained by sorting all vital strings by $w(u)$ (first key) and $|v(u)|$ (second key) and concatenating them. Note that this procedure of generating an ideal summary may not be optimal, yet it is not a serious problem in practice as discussed in the original paper (Sakai et al. 2011a).

Finally, the original notation of *S*-measure is shown below, though it is obviously equivalent to Eq. 11.2:

$$S\text{-measure} = \frac{\sum_{u \in M} w(u) \max(0, L - \text{offset}(u))}{\sum_{u \in U} w(u) \max(0, L - \text{offset}^*(v(u)))}, \tag{11.4}$$

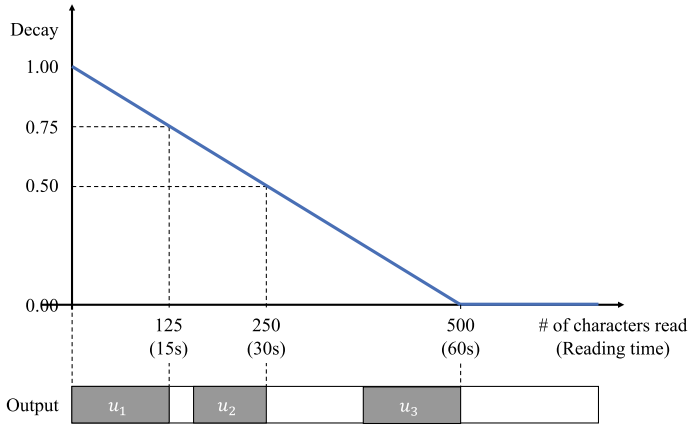


Fig. 11.2 Illustration of S-measure computation. The x-axis represents the number of characters read by the user, and y-axis represents the offset-based decay ($\max(0, 1 - \text{offset}(u))/L$) with $L = 500$. The x-axis can also be interpreted as reading time indicated in the parentheses when the reading speed is 500 characters per minute. The textual output located at the bottom includes three iUnits u_1, u_2, and u_3. The position of iUnits is aligned to the x-axis and their offsets are 125, 250, and 500, respectively. Their weight is 1 for simplicity. S-measure for this textual output can be computed as $S\text{-measure} = \frac{1}{\mathcal{N}}(1 \cdot 0.75 + 1 \cdot 0.50 + 1 \cdot 0.00) = \frac{1}{\mathcal{N}} \cdot 1.25$

11.3.4 M-Measure

M-measure (Kato et al. 2016a) was the primary evaluation metric at NTCIR-11 MobileClick-1 and NTCIR-12 MobileClick-2, which was proposed for two-layered summaries.

Intuitively, a two-layered summary is good if: (1) The summary does not include non-relevant iUnits in the first layer; (2) The first layer includes iUnits relevant for all the intents; and (3) iUnits in the second layer are relevant for the intent that links to them.

To be more specific, the following choices and assumptions were made for evaluating two-layered summaries:

- Users are interested in one of the intents $i \in I_q$ by following the intent probability $P(i|q)$, where I_q is a set of intents for query q.
- Each user reads a summary following these rules:

1. The user starts to read a summary from the beginning of the first layer.
2. When reaching the end of a link l_i which interests a user with intent i, the user clicks on the link and starts to read its second layer $\mathbf{s}_i$.
3. When reaching the end of the second layer $\mathbf{s}_i$, the user goes back to the end of the link l_i and continues reading.
4. The user stops after reading no more than L characters.

- The weight of iUnits is judged per intent. Therefore, an iUnit is important for a user but may not be important for another user.
- The utility of text read by a user is measured by U-measure proposed by Sakai and Dou (2013), which consists of a position-based gain and a position-based decay function.
- The evaluation metric for two-layered summaries, M-measure, is the expected utility of text read by different users.

These choices and assumptions derive all possible *trailtexts* and their probability in a two-layered summary. A trailtext is a concatenation of all the texts read by a user, and can be defined as a list of iUnits and links consumed by the user. According to the user model described above, a trailtext of a user who is interested in intent i can be obtained by inserting a list of iUnits in the second layer $\mathbf{s}_i$ after the link of l_i. More specifically, given the first layer $\mathbf{f} = (u_1, \ldots, u_{j-1}, l_i, u_j, \ldots)$ and second layer $\mathbf{s}_i = (u_{i,1}, \ldots, u_{i,|\mathbf{s}_i|})$, trailtext $\mathbf{t}_i$ of intent i is defined as follows: $\mathbf{t}_i = (u_1, \ldots, u_{j-1}, l_i, u_{i,1}, \ldots, u_{i,|\mathbf{s}_i|}, u_j, \ldots)$. An example of trailtexts in a two-layered summary is shown in Fig. 11.3.

M-measure, an evaluation metric for the two-layered summarization, is the expected utility of text read by users:

$$M = \sum_{\mathbf{t} \in T} P(\mathbf{t}) U(\mathbf{t}), \tag{11.5}$$

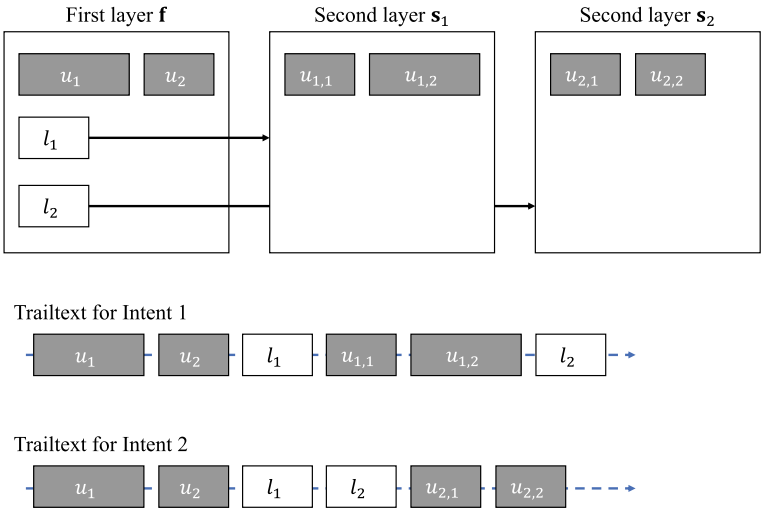


Fig. 11.3 Example of trailtexts in a two-layered summary. Suppose links l_1 and l_2 are interesting for users with intents 1 and 2, respectively. All the users start to read the summary from the beginning of the first layer and read iUnits u_1 and u_2. A user with intent 1 clicks on link l_1, reads the iUnits in the second layer $\mathbf{s}_1$, and goes back to the first layer for reading the rest. A user with intent 2 does not click on link l_1 but clicks on link l_2, reads the iUnits in $\mathbf{s}_2$, and returns to the first layer. These different trails result in different trailtexts shown at the bottom of the figure

where T is a set of all possible trailtexts, $P(\mathbf{t})$ is a probability of going through a trailtext $\mathbf{t}$, and $U(\mathbf{t})$ is the U-measure score of a trailtext $\mathbf{t}$.

For simplicity, a one-to-one relationship between links and intents was assumed in NTCIR-12 MobileClick-2. Therefore, there is only a relevant link and a trailtext for each intent. It follows that the probability of each trailtext being generated is equivalent to the probability of the corresponding intent, *i.e.,* $P(\mathbf{t}_i) = P(i|q)$ where $\mathbf{t}_i$ denotes a trailtext read by users with intent i. Then, M-measure can be rewritten as

$$M = \sum_{i \in I_q} P(i|q) U_i(\mathbf{t}_i). \tag{11.6}$$

where $U_i(\mathbf{t}_i)$ is the U-measure score of a trailtext $\mathbf{t}_i$ for users with intent i.

The computation of U-measure (Sakai and Dou 2013) is the same as that of S-measure except for the normalization factor and definition of the weight. U-measure is defined as follows:

$$U_i(\mathbf{t}) = \frac{1}{\mathcal{N}} \sum_{j=1}^{|\mathbf{t}|} g_i(u_j) d(u_j), \tag{11.7}$$

where $g_i(u_j)$ is the weight of iUnit u_j in terms of intent i, d is a position-based decay function, and $\mathcal{N}$ is a *constant* normalization factor ($\mathcal{N}$=1 in NTCIR MobileClick). Note that a link in the trailtext is regarded as a non-relevant iUnit for the sake of convenience. The position-based decay function is the same as that of S-measure:

$$d(u) = \max(0, 1 - \text{offset}(u)/L). \tag{11.8}$$

11.4 Outcomes of NTCIR 1CLICK and MobileClick

This section highlights the outcomes of NTCIR 1CLICK and MobileClick. We first present the results of each task and then discuss their potential impacts.

11.4.1 Results

Table 11.4 shows the number of participants and submissions at each NTCIR task. While the first round of 1CLICK and MobileClick failed to attract many participants, the second round of each received a sufficient number of submissions from ten or more teams. Due to a small number of participants, we only summarize results from 1CLICK-2 and MobileClick-2.

Table 11.4 The number of participants and submissions at each NTCIR task

Year	NTCIR	Task	# of participants	# of submissions
2010	9	1CLICK-1	3	10
2011	10	1CLICK-2	10	38 (for the Main task)
2013	11	MobileClick-1	4	24 (for retrieval) & 11 (for summarization)
2014	12	MobileClick-2	12	37 (for retrieval) & 29 (for summarization)

The NTCIR-10 1CLICK-2 results showed that simple use of search engine snippets and the first paragraph of Wikipedia articles outperformed more sophisticated approaches for both of the English and Japanese queries. Those simple approaches were particularly effective for celebrity query types, while they were not for the other types such as local queries (Kato et al. 2013b).

The NTCIR-12 MobileClick-2 task results showed that some participants' runs outperformed the baselines. Since the MobileClick task required systems to group iUnits relevant to the same intent, some teams proposed effective methods to measure the similarity between intents and iUnits, and achieved significantly better results than baselines. For example, one of the top performers used word embedding for measuring the intent-iUnit similarity, and another team proposed an extension of topic-sensitive PageRank for the summarization task. Per-query analysis at MobileClick-2 also suggested that celebrity query types were easy, while local and Q&A types of queries are difficult for both baselines and participants' systems (Kato et al. 2016b).

11.4.2 Impacts

An evaluation metric for summaries, ranked lists, and sessions, U-measure, was proposed by Sakai and Dou (2013). As they explained, U-measure was inspired by S-measure and is a generalization of S-measure. U-measure was further extended to the evaluation of customer-helpdesk dialogues by Zeng et al. (2017).

Luo et al. (2017) proposed *height-biased gain* (HBG), an evaluation metric for mobile search engine result pages. HBG is computed by summing up the product of weight and decay that are both modeled in terms of result height in mobile search engine result pages. As the authors mentioned in their paper, U-measure is one of the evaluation metrics that inspired HBG.

Arora and Jones (2017a, b) adapted the definition of iUnits for their study on identifying useful and important information and how people perceive information.

In commercial search engines, *direct answers* or *featured snippets* have become an important part of the search engine result page. This functionality presents a text that answers a question given as a query, just like the textual output of NTCIR 1CLICK. As of May 2019, it seems that they only show a part of a webpage and do

not summarize multiple webpages. The evaluation methodology of NTCIR 1CLICK and MobileClick could be potentially useful when direct answers are composed from multiple webpages and need to be evaluated in detail.

11.5 Summary

This chapter introduced the earliest attempts toward test-collection-based evaluation for information access with smartphones, namely, NTCIR 1CLICK and MobileClick. Those campaigns aimed to develop an IR system that outputs a single, short text summary for a given query, which is expected to fit a small screen and to satisfy users' information needs without requiring much interaction. This chapter mainly discussed the novelty of the evaluation methodology used in those evaluation campaigns by contrasting it with ordinary summarization evaluation. Moreover, the potential impacts of NTCIR 1CLICK and MobileClick were discussed as well.

Acknowledgements I really thank Tetsuya Sakai for his long-term contributions to the NTCIR 1CLICK and MobileClick tasks. A large fraction of the technologies introduced in this chapter was invented by or inspired by him. I also would like to thank the other task organizers of NTCIR 1CLICK and MobileClick, Young-In Song, Virgil Pavlu, Matthew Ekstrand-Abueg, Takehiro Yamamoto, Mayu Iwata, Hajime Morita, and Sumio Fujita. I would like to express deep gratitude to all the participants in the NTCIR-9 1CLICK-1, NTCIR-10 1CLICK-2, NTCIR-11 MobileClick-1, and NTCIR-12 MobileClick-2 tasks. Finally, I would like to express my special thanks to Stefano Mizzaro, who reviewed the initial version of this chapter and provided many useful suggestions.

References

Arora P, Jones GJ (2017a) How do users perceive information: Analyzing user feedback while annotating textual units. In: Proc. of the Second Workshop on Supporting Complex Search Tasks, pp 7–11

Arora P, Jones GJ (2017b) Identifying useful and important information within retrieved documents. In: CHIIR, pp 365–368

Babko-Malaya O (2008) Annotation of nuggets and relevance in gale distillation evaluation. In: LREC

Carbonell JG, Goldstein J (1998) The use of MMR and diversity-based reranking for reordering documents and producing summaries. In: SIGIR, pp 335–336

Crestani F, Mizzaro S, Scagnetto I (2017) Mobile information retrieval. Springer, Berlin

Dang HT, Kelly D, Lin JJ (2007) Overview of the TREC 2007 question answering track. In: TREC

Hovy EH, Lin CY, Zhou L, Fukumoto J (2006) Automated summarization evaluation with basic elements. LREC 6:604–611

Kato MP, Ekstrand-Abueg M, Pavlu V, Sakai T, Yamamoto T, Iwata M (2013a) Overview of the NTCIR-10 1CLICK-2 Task. In: NTCIR-10 conference, pp 243–249

Kato MP, Sakai T, Yamamoto T, Iwata M (2013b) Report from the NTCIR-10 1click-2 Japanese subtask: baselines, upperbounds and evaluation robustness. In: SIGIR, pp 753–756

Kato MP, Ekstrand-Abueg M, Pavlu V, Sakai T, Yamamoto T, Iwata M (2014) Overview of the NTCIR-11 mobileClick task. In: NTCIR-11 Conference

Kato MP, Pavlu V, Sakai T, Yamamoto T, Morita H (2016a) Two-layered summaries for mobile search: does the evaluation measure reflect user preferences? In: Proceeding of the seventh international workshop on evaluating information access (EVIA 2016), pp 29–32
Kato MP, Sakai T, Yamamoto T, Pavlu V, Morita H, Fujita S (2016b) Overview of the NTCIR-12 mobileClick-2 task. In: NTCIR-12 conference
Li J, Huffman S, Tokuda A (2009) Good abandonment in mobile and PC internet search. In: SIGIR, pp 43–50
Lin CY (2004) ROUGE: A package for automatic evaluation of summaries. In: Text summarization branches out: proceedings of the ACL-04 workshop, pp 74–81
Luo C, Liu Y, Sakai T, Zhang F, Zhang M, Ma S (2017) Evaluating mobile search with height-biased gain. In: SIGIR, pp 435–444
Nenkova A, McKeown K (2011) Automatic summarization. Found Trends® Inf Retriev 5(2–3):103–233
Nenkova A, Passonneau R, McKeown K (2007) The pyramid method: incorporating human content selection variation in summarization evaluation. ACM Trans Speech Lang Process (TSLP) 4(2):4
Sakai T, Dou Z (2013) Summaries, ranked retrieval and sessions: a unified framework for information access evaluation. In: SIGIR, pp 473–482
Sakai T, Kato MP (2012) One click one revisited: enhancing evaluation based on information units. In: AIRS, pp 39–51
Sakai T, Kato MP, Song YI (2011a) Click the search button and be happy: evaluating direct and immediate information access. In: CIKM, pp 621–630
Sakai T, Kato MP, Song YI (2011b) Overview of NTCIR-9 1CLICK. In: NTCIR-9, pp 180–201
Tombros A, Sanderson M (1998) Advantages of query biased summaries in information retrieval. In: SIGIR, pp 2–10
Voorhees EM (2003) Overview of the TREC 2003 question answering track. TREC 2003:54–68
Zeng Z, Luo C, Shang L, Li H, Sakai T (2017) Test collections and measures for evaluating customer-helpdesk dialogues. In: Proceedings of the eighth international workshop on evaluating information access (EVIA 2017), pp 1–9

Chapter 12
Mathematical Information Retrieval

Akiko Aizawa and Michael Kohlhase

Abstract We present an overview of the NTCIR Math Tasks organized during NTCIR-10, 11, and 12. These tasks are primarily dedicated to techniques for searching mathematical content with formula expressions. In this chapter, we first summarize the task design and introduce test collections generated in the tasks. We also describe the features and main challenges of mathematical information retrieval systems and discuss future perspectives in the field.

12.1 Introduction

The NTCIR Math Tasks are aimed at developing test collections for mathematical search in STEM (Science/Technology/Engineering/Mathematics) documents to facilitate and encourage research in mathematical information retrieval (MIR) (Liska et al. 2011) and its related fields (Guidi and Sacerdoti Coen 2016; Zanibbi and Blostein 2012).

Mathematical formulae are important for the dissemination and communication of scientific information. They are not only used for numerical calculation but also for clarifying definitions or disambiguating explanations that are written in natural language. Despite the importance of math in technical documents, most contemporary information retrieval systems do not support users' access to mathematical formulae in target documents. One major obstacle to MIR research is the lack of readily available large-scale datasets with structured mathematical formulae, carefully designed tasks, and established evaluation methods.

MIR involves searching for a particular mathematical concept, object, or result, often expressed using mathematical formulae, which—in their machine-readable

A. Aizawa (✉)
National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
e-mail: aizawa@nii.ac.jp

M. Kohlhase
FAU Erlangen-Nürnberg, Martenstr. 3, 91058 Erlangen, Germany
e-mail: michael.kohlhase@fau.de

T. Sakai et al. (eds.), *Evaluating Information Retrieval and Access Tasks*,
The Information Retrieval Series 43,
https://doi.org/10.1007/978-981-15-5554-1_12

forms—are expressed as complex expression trees. To answer MIR queries, a search system should tackle at least two challenges: (1) tree structure search and (2) utilization of textual context information.

To understand the problem, consider an engineer who wants to prevent an electrical system from overheating, thus, needs a tight upper estimate for the energy term

$$\int_a^b |V(t)I(t)|dt$$

for all a, b, where V is voltage and I current. Search engines, such as Google, are restricted to word-based searches of mathematical articles, which barely helps with finding mathematical objects because there are no keywords to search for. Computer algebra systems cannot help either since they do not incorporate the necessary special knowledge. However, the required information is out there, e.g., in the form of

Theorem 17. (Hölder's Inequality)
If f and g are measurable real functions, $l, h \in \mathbb{R}$, and $p, q \in [0, \infty)$, such that $1/p + 1/q = 1$, then

$$\int_l^h |f(x)g(x)|\,dx \leq \left(\int_l^h |f(x)|^p\,dx\right)^{\frac{1}{p}} \left(\int_l^h |g(x)|^q\,dx\right)^{\frac{1}{q}}$$

For mathematical content (here the statement of Hölder's inequality) to be truly searchable, it must be in a form in which an MIR system can find it from a query

$$\int_{\boxed{\text{a}}}^{\boxed{\text{b}}} |V(t)I(t)|dt \leq \boxed{\text{R}}$$

the boxed identifiers are query variables (see Sect. 12.3.2)—and can even extend the calculation to

$$\int_a^b |V(t)I(t)|dt \leq \left(\int_a^b |V(x)|^2\,dx\right)^{\frac{1}{2}} \left(\int_a^b |I(x)|^2\,dx\right)^{\frac{1}{2}}$$

after the engineer chooses $p = q = 2$ (Cauchy–Schwarz inequality). Estimating the individual V and I values is now a much simpler problem.

Admittedly, Google would have found the information by querying for "Cauchy–Schwarz Hölder", but that keyword was the crucial information the engineer was missing in the first place. In fact, it is not unusual for mathematical document collections to be so large that determining the identifier of the sought-after object is harder than recreating the actual object.

In this example we see the effect of both (1) formula structure search and (2) context information as postulated above:

1. The formula structure is mapped by unification (finding a substitution for the boxed query variables to make the query and main formula of Hölder's inequality structurally identical or similar (see Sect. 12.3.2).

2. We have used the context information about the parameters of Hölder's inequality, e.g., that the identifiers f, g, p, and q are universal (thus can be substituted for); the first two are measurable functions and the last two are real numbers.

In the following sections, we summarize our attempts at NTCIR to develop datasets for MIR together with some future perspectives of the field.

12.2 NTCIR Math: Overview

Prior to the NTCIR Math Tasks, MIR had been mainly approached by researchers in digital mathematics libraries, and only a little attention has been paid by the information retrieval community. Unlike other scientific disciplines that require a search for specific types of named entities such as genes, diseases, and chemical compounds, mathematics is based on abstract concepts with many possible interpretations when mapped to a real-world phenomenon. This means that although their mathematical definitions are rigid, mathematical concepts are inherently ambiguous in their applications to the real world. Also, the representation of mathematical formulae can be highly complicated with diverse types of symbols including user-defined functions, constants, and free and bound variables. As such, MIR requires dedicated search techniques such as approximate tree matching or unification. To summarize, in the context of information retrieval, MIR is not only a challenge for novel retrieval targets but also featured as a testbed for (1) retrieval of non-textual objects in documents using their context information and (2) a large-scale complex tree structure search with a realistic application scenario.

The NTCIR Math tasks were the first trial to introduce an evaluation framework of information retrieval to mathematical formula search. NTCIR Math Tasks were organized three times during NTCIR-10, 11, and 12, i.e., the NTCIR-10 Math Pilot Task, NTCIR-11 Math-2 Task, and NTCIR-12 MathIR Task.

12.2.1 NTCIR-10 Math Pilot Task

The NTCIR-10 Math Pilot Task (Aizawa et al. 2013) was the first attempt to develop a common workbench for mathematical formula search. This task was organized as two independent subtasks:

1. The first was the Math Retrieval Subtask in which the objective was to retrieve relevant documents given a math query.
2. The second was the Math Understanding Subtask in which the objective was to identify textual spans that describe math formulae that appear in the document.

The corpus used for this task was based on 100,000 arXiv documents converted from LaTeX to XHTML by the arXMLiv project.[1]

Six teams participated in this task, all six contributing to the Math Retrieval Subtask and only one to the Math Understanding Subtask.

12.2.2 NTCIR-11 Math-2 Task

The NTCIR-10 Math Pilot Task showed that participants considered the Math Retrieval Subtask more important. Therefore, the succeeding two tasks focused only on this subtask and made it as compulsory for all participants. In the NTCIR-11 Math-2 Task (Aizawa et al. 2014), based on the feedback from the participants in the pilot task, both the arXiv corpus and topics were reconstructed. Apart from this main subtask using the arXiv corpus, the NTCIR-11 Math-2 Task also provided an open free subtask using math-related Wikipedia articles. This optional subtask required an exact formula search (without any keywords) and complements the main subtask with an automated performance evaluation.

The NTCIR-11 Math-2 Task had eight teams participating (two new teams joined), most contributing to both subtasks .

12.2.3 NTCIR-12 MathIR Task

For the NTCIR-12 MathIR Task (Zanibbi et al. 2016), we reused the arXiv corpus we prepared for the NTCIR-11 Math-2 Task but with new topics. This subtask introduced a new formula query operator, *simto region*, that explicitly requires an approximate matching function for math formulae. We also created a new corpus of Wikipedia articles to provide a use case of math retrieval by nonexperts. The design of the subtask for the Wikipedia corpus was similar to that in the NTCIR-11 Math-2 Task except that a topic includes not only exact formula search but also formula+keyword search (Table 12.1).

Six teams participated in the NTCIR-12 MathIR Task.

12.3 NTCIR Math Datasets

In this section, we mainly describe the two datasets, arXiv and Wikipedia, designed for the Math Retrieval Subtasks during NTCIR-12. Each dataset consists of a corpus with mathematical formulae, a set of topics in which each query is expressed as

[1] https://kwarc.info/projects/arXMLiv/.

Table 12.1 Summary of NTCIR math subtasks

Subtasks		NTCIR-10	NTCIR-11	NTCIR-12
Math Retrieval Subtask for the ArXiv cor pus	Formula search	○		
	Formula+keyword search	○	○	○
	Formula+keyword search with "simto"			○
	Free-form query search	○		
Math Retrieval Subtask for the Wikipedia cor pus	Formula search		○	○
	Formula+keyword search			○
	Formula+keyword search with 'simto'			
Math understanding subtask		○		

a combination of mathematical formulae schemata and keywords, and relevance judgment results based on the submissions from participating teams.

12.3.1 Corpora

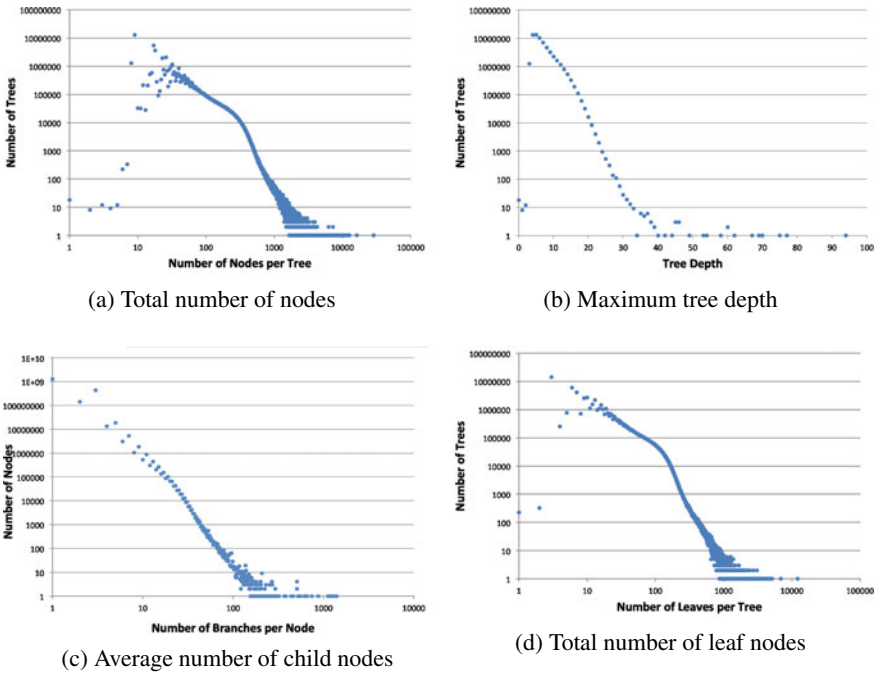
The arXiv corpus contains paragraphs from technical articles in the arXiv,[2] while the Wikipedia corpus contains complete articles from Wikipedia. Generally speaking, the arXiv articles (preprints of research articles) were written by technical experts for technical experts assuming a high level of mathematical sophistication from readers. In contrast, many Wikipedia articles on mathematics were written to be accessible for nonexperts at least in part.

12.3.1.1 ArXiv Corpus

The arXiv corpus consists of 105,120 scientific articles in English. These articles were converted from LaTeX sources available at http://arxiv.org to HTML5+MathML using the LaTeXML system[3] and include the arXiv categories math, cs, physics:math-ph, stat, physics:hep-th, and physics:nlin to obtain a varied sample of technical documents containing mathematics.

[2] http://www.arxiv.org.

[3] http://dlmf.nist.gov/LaTeXML/.



(a) Total number of nodes

(b) Maximum tree depth

(c) Average number of child nodes

(d) Total number of leaf nodes

Fig. 12.1 Math formulae statistics for the arXiv corpus

This subtask was designed for both formula-based search systems and document-based retrieval systems. In document-wise evaluation, human evaluators need to check all math formulae in the document. To reduce the cost of relevance judgment, we divided each document into paragraphs and used them as the search units ("documents") for the subtask. This produced 8,301,578 search units with roughly 60 million math formulae (including isolated symbols) encoded using LaTeX, Presentation MathML, and Content MathML Formulae[4]; 95% of the retrieval units had 23 or fewer math formulae, which is sufficiently small for document-based relevance judgment by human reviewers. Excerpts are stored independently in separate files, in both HTML5 and XHTML5 formats.

Figure 12.1 summarizes the basic statistics for the math formula trees in the ArXiv corpus. Figure 12.1a–d correspond to the distributions of the total number of nodes, maximum tree depth, average number of child nodes, and total number of leaf nodes in each math formula, respectively. These statistics show that the math trees in the arXiv corpus approximately follow the power-law distribution in their size. While there exists a vast amount of relatively simple trees, there also exists a non-negligible number of highly complex trees. This clearly shows that, as a benchmark for tree

[4]MathML (Ausbrooks et al. 2010) supplies two sub-languages: Presentation MathML encodes the visual (and possibly aural) appearance of the formulae in terms of a tree of layout primitives and Content MathML encodes the functional structure of formulae in terms of an operator tree.

structure search, the corpus is characterized by its large scale as well as the heterogeneity of the trees in it.

12.3.1.2 Wikipedia Corpus

The Wikipedia corpus contains 319,689 articles from English Wikipedia converted into a simpler XHTML format with images removed (5.15 GB uncompressed).[5] Unlike the arXiv corpus, articles were not split into smaller documents since they were simple/small enough for human annotation. Only 10% of the articles of the Wikipedia corpus contain explicit `<math>` tags that demarcate LaTeX, reflecting the small proportion of articles related to math in Wikipedia, while keeping the corpus size manageable for participants. All articles with a `<math>` tag were included in the corpus and the remaining 90% were sampled from articles that do not contain any `<math>` tag. These "text" articles act as distractors for keyword matching. There are over 590,000 formulae in the corpus with the same format as the arXiv corpus, i.e., encoded using LaTeX, Presentation MathML, and Content MathML. Note that untagged formulae frequently appear directly in HTML text (e.g. `'where x <sup> 2 ...'`). We made no attempt to detect or label these formulae embedded in the main text.

12.3.2 Topics

The Math Retrieval Subtasks were designed so that all topics include at least a single relevant document in the corpus, and ideally multiple relevant documents. In some cases, this is not possible, for example, with navigational queries where a specific document is sought after.

12.3.2.1 Topic Format

Details about the topic format are available in the documentation provided by the organizers (Kohlhase 2015). For participants, a math retrieval topic contains a (1) topic ID and (2) query (formula + keywords), but no textual description. The description is omitted to avoid participants biasing their system design toward the specific information needs identified in the topics. For evaluators, each topic also contains a narrative field that describes a user situation, the user's information needs, and relevance criteria. Formula queries are encoded in LaTeX, Presentation MathML, and Content MathML. In addition to the standard MathML notations, the following two subtask-specific extensions are adopted : *formulae query variables* and *formula simto regions* (see below).

[5] http://www.cs.rit.edu/~rlaz/NTCIR12_MathIR_WikiCorpus_v2.1.0.tar.bz2.

Formulae Query Variables (Wildcards). Formulae may contain query variables that act as wildcards, which can be matched to arbitrary subexpressions on candidate formulae. Query variables were represented using two different representations for the arXiv and Wikipedia topics. For the arXiv topics, query variables are named and indicated by a question mark (e.g., *?v*) while for the Wikipedia topics, wildcards are numbered and appear between asterisks (e.g., **1**).

This is an example query formula with the three query variables ?f, ?v, and ?d.

$$\frac{?f(?v + ?d) - ?f(?v)}{?d} \tag{12.1}$$

This query matches the argument of the limit on the right side of the equation below, substituting g for ?f, cx for ?v, and h for ?d. Note that each repetition of a query variable matches the same subexpression.

$$g'(cx) = \lim_{h \to 0} \frac{g(cx + h) - g(cx)}{h} \tag{12.2}$$

Formula Simto Regions. *Similarity regions* modify our formula query language, distinguishing subexpressions that should be identical to the query from those that are similar to the query in some sense. Consider the query formula below, which contains a *similarity region* called "a."

$$\frac{[a|\ g(cx + h) - g(cx)}{h} \tag{12.3}$$

The fraction operator and numerator h should match exactly, while the numerator may be replaced by a "similar" subexpression. Depending on the notion of similarity we choose to adopt, simto region "a" might match "$g(cx + h)+g(cx)$", if addition is similar to subtraction, or "$g(cx + h) - g(\mathbf{d}x)$", if c is somehow similar to d. The simto regions may also contain exact match constraints (see Kohlhase 2015).

12.3.2.2 ArXiv Topics

A total of 50 and 37 topics were provided during NTCIR-11 and NTCIR-12, respectively. Many of the topics in the arXiv subtask are sophisticated, for example, seeking to determine whether a connection exists between a factorial product and products starting with one. Some queries are simpler, such as looking for applications of operators, or loss functions used in machine learning. Eight out of the 37 topics during NTCIR-12 contained simto regions.

12.3.2.3 Wikipedia Topics

Topics for the Wikipedia subtask were designed with a less expert user population in mind. We imagined undergraduate and graduate students searching Wikipedia to locate or remember and relocate specific articles (i.e. navigational queries), browse math articles, learn/review mathematical concepts and notation they come across in their studies, find applications of concepts, or find information to help solve particular mathematical problems (e.g., for homework). A total of 30 topics were provided during NTCIR-12.

12.3.3 Relevance Judgment

The evaluation of the Math Retrieval Subtasks was pooling-based. First, all submitted results were converted into a `trec_eval` result file format. Next, for each topic, the top-20 ranked documents were selected from each run. Then, the set of pooled hits were evaluated by human assessors. After the pooling process, the selected retrieval units were fed into the SEPIA system[6] with MathML extensions developed by the organizers. Evaluators judged the relevance of each retrieval unit by comparing it to the query formulae and keywords, along with the described scenario provided with the topic, and selected one of the judgments `relevant` (R), `partially relevant` (PR), or `not-relevant` (N). The retrieval units were documents except for Wikipedia formula-only subtask, where the evaluation was based on individual formulae.

Evaluators had to rely on their mathematical intuition, the described information needs, and actual query to determine judgments. For the arXiv dataset, to ensure sufficient familiarity with mathematical documents, three evaluators were chosen from third-year and graduate students of (pure) mathematics. Each topic was evaluated by at least two evaluators. For the Wikipedia dataset, intended to represent mathematical information needs for nonexperts, ten students were recruited for evaluation: five undergraduates and five graduate (MSc) students. The Fleiss' κ values were 0.5615 and 0.5380 for the arXiv dataset and 0.3546 and 0.2619 for the Wikipedia dataset. Agreement between evaluators for the arXiv dataset was higher. This may be because of the greater mathematical expertise and shared background by these evaluators.

[6] https://code.google.com/p/sepia/.

12.4 Task Results and Discussion

12.4.1 Evaluation Metrics

In our evaluation, the judgment of each evaluator was converted into a relevance score using the mappings "Relevant" → 2, "Partially Relevant" → 1, and "Not Relevant" → 0. Then, the average score was binarized as follows:

- For "relevance" evaluation, the overall judgment is considered `relevant` if the average score is equal or greater than 1.5, and `not relevant` otherwise.
- For "partial relevance" evaluation, the overall judgment is considered `relevant` if the average score is equal or greater than 0.5, and `not relevant` otherwise.

Precision@k for $k = \{5, 10, 15, 20\}$ was used to evaluate participating systems. We chose these measures because they are simple to understand and characterize retrieval behavior as the number of hits increases. Precision@k values were obtained from `trec_eval` version 9.0, with which they were labeled `P_avgjg_5`, `P_avgjg_10`, `P_avgjg_15`, and `P_avgjg_20`, respectively.

12.4.2 MIR Systems

The numbers of participating teams were 6, 8, 6 for the NTCIR 10, 11, 12 Math Tasks. Three teams participated in all three tasks. For NTCIR 11 and 12, there were one or two new participating teams. The architectures of the participating systems were quite diverse. For formula encodings, all the LaTeX, MathML Presentation Markup, MathML Content Markup formats were used by at least one system; Presentation Markup was the most popular notation. Also, the majority of systems used a general-purpose search engine for indexing.

The following common technical decisions should be considered in designing MIR systems.

12.4.2.1 How to Index Math Formulae?

Mathematical formulae are expressed as XML tree structures, which often become very complex. However, the search sometimes requires approximate matching to guarantee certain flexibility. There are two strategies for indexing math formulae: token-based and subtree-based. While token-based indexing takes into account math tokens, the same as words in a text, subtree-based indexing decomposes the XML structure into smaller fragments, i.e.,subtrees, and treats them as indexing units. In the NTCIR Math Tasks, the majority of systems took into account structural information for formulae.

12.4.2.2 How to Deal with Query Variables?

One of the prominent features of MIR is that a query formula can contain "variables", i.e., symbols that can serve as named wildcards. Since the unification operation is expensive, most participating systems used a re-ranking step, wherein one or more initial rankings are merged and/or reordered. This approach of obtaining an initial candidate ranking followed by a refined ranking is a common and effective strategy. To locate strong partial matches, all the automated systems used unification, whether for variables (e.g., "$x^2 + y^2 = z^2$" unifies with "$a^2 + b^2 = c^2$"), constants, or entire subexpressions (e.g., via structural unification or indirectly through *generalized terms* with wildcards for operator arguments).

12.4.2.3 Other Technical Decisions

Other issues include how to identify the importance of the keywords/math formulae in queries and documents; exploit context information; normalize math formulae with possibly many notation variations; deal with ambiguity in the original LaTeX notation; combine keyword-based search with math formula search; and deal with "simto"-type queries. To summarize, there can be many options for MIR system design, and they should be balanced with computation cost.

12.5 Further Trials

The NTCIR Math Tasks also contain several important trials that lead to further exploration in succeeding research, as detailed below.

12.5.1 ArXiv Free-Form Query Search at NTCIR-10

The NTCIR-10 Math Pilot Task contained 19 *open* queries from mathematicians expressed as free descriptions with natural language text and formulae. Here is an example (NTCIR10-OMIR-19):

> *Let X_n be a decreasing sequence of nonempty closed sets in a Banach space such that their diameter tends to 0. Is their intersection nonempty?*

These topics were collected from questions asked by mathematicians in related forums, which makes the task settings more realistic and general. Since converting the textual descriptions into "keyword+formula" queries requires deep natural language comprehension, we did not pursue this direction further in this task. However, real queries in forums are an important resource for analyzing user information needs in their retrieval (Mansouri et al. 2019; Stathopoulos and Teufel 2015).

The Answer Retrieval for Questions on Math (ARQMath) is a newly launched task for the 11th Conference and Labs of the Evaluation Forum (CLEF 2020).[7] Data from Math Stack Exchange,[8] a mathematics-dedicated question answering forum, are expected to be used for ARQMath. Such explorations are expected to give further insights into realistic information needs.

12.5.2 Wikipedia Formula Search at NTCIR-11

The NTCIR-11 Math-2 Task provided the first open platform for comparing formula search engines, based upon their ability to retrieve specific formula in Wikipedia articles (Schubotz et al. 2015). By using formula-only queries that require an exact match of the math tree structure, the platform enables automatic evaluation without any human intervention. Regardless of the simplicity of the task, the automatic evaluation framework was useful in verifying and tuning the formula search function of math search engines. This will enable us to establish leaderboard-style comparison of different strategies for complicated large-scale formula searches.

12.5.3 Math Understanding Subtask at NTCIR-10

The goal of the Math Understanding Subtask was to extract natural language definitions of mathematical formulae in a document for their semantic interpretation. The dataset for this subtask contains 10 manually annotated articles used in a dry run and an additional 35 used in a formal run.

A description is obtained from a continuous text region or concatenation of some discontinuous text regions. Shorter descriptions may also be obtained from a longer one. For instance, in the text "$log(x)$ is a function that computes the natural logarithm of the value x", the complete description of "$log(x)$" is "a function that computes the natural logarithm of the value x". Moreover, the shorter descriptions "a function" and "a function that computes the natural logarithm" can be obtained from the previous one. This corpus defines two types of possible descriptions of mathematical expressions, namely full description (contains the complete type) and short description (contains the short type). Participants could extract any type of description in their submission.

The training and test set consists of 35 and 10 annotated papers selected from the arXiv copus, respectively. Inter-annotator agreement was tested for the five papers taken from the corpus. There are three measurements to test the reliability of annotation: F1-score, Cohen's kappa, and Krippendorff's alpha. To compute the F1-score, the position of the annotated descriptions from two annotators is strictly matched.

[7] https://www.cs.rit.edu/~dprl/ARQMath/.

[8] https://math.stackexchange.com/.

The F1-score was 0.8670, Cohen's kappa was 0.8993, and Krippendorff's alpha was 0.7630 for full descriptions, and F1-score was 0.9014 for full and short descriptions). The evaluation was conducted by matching the position of the extracted descriptions against the positions of gold-standard descriptions, and precision, recall, and F1-score were used.

Math-description extraction is considered important to combine mathematical formulae with their textual descriptions for their interpretation. For example, Kristianto et al. (2017) combined the description extraction with formula dependency extraction and obtained consistent improvement in the Math Retrieval Subtasks in the succeeding NTCIR Math Tasks.

12.6 Further Impact of NTCIR Math Tasks

Several years after these NTCIR Math Tasks, we witnessed a number of valuable developments in mathematical content access studies. This section provides a brief introduction to some of these activities, although it is far less comprehensive.

12.6.1 Math Information Retrieval

Since these NTCIR Math Tasks, increasing attention has been paid to semantic retrieval of mathematical formulae. NLP techniques often play a critical role in bridging the gap between presentation and semantic representations of math formulae. Recent studies on this topic include variable typing (Stathopoulos et al. 2018), using the textual context for transformation from a presentation level to semantic level (Schubotz et al. 2018), and identifying declarations of mathematical objects (Lin et al. 2019).

Overall, there are several valuable approaches to MIR, including those we could not introduce in this book chapter. According to the number of citations on Semantic-Scholar,[9] the overview papers of the Math Tasks during NTCIR-10, 11, and 12 have 39, 39, 33 citations, respectively, as of December 2019. MIR is also characterized by the diversity of the conferences and journals of the related papers, including such fields as mathematics, information retrieval, image recognition, NLP, knowledge management, and document processing.

[9] https://www.semanticscholar.org.

12.6.2 Semantics Extraction in Mathematical Documents

Noteworthy recent work includes a general-purpose part-of-math tagger that performs semantic disambiguation and parsing of math formulae (Youssef 2017) and embeddings of math symbols (Mansouri et al. 2019; Youssef and Miller 2019). It has also been reported that image-based math-formula search is also capable of capturing semantic similarity without unification (Davila et al. 2019). Other related topics that were not addressed during the NTCIR Math Tasks include math document categorization (Barthel et al. 2013) using formulae information (Suzuki and Fujii 2017).

12.6.3 Corpora for Math Linguistics

The development work for the arXiv corpus (and the subsequent requests by the community) made it very clear that work on document understanding and information in Mathematics and STEM can only succeed based on large and shared document corpora. A single conversion run over the arXiv corpus (over 1.5 Million documents) is a multi-processor-year enterprise generating $10^8 - 10^9$ error reports in gigabytes of log files.

To support and manage this computational task, the corTEXsystem[10] has been developed as a general-purpose processing framework for corpora of scientific documents. The licensing issues involved in distributing the ensuing corpora have led to the recent establishment of *Special Interest group for Math Linguistics (SIGMathLing)*, [11] a forum and resource cooperative for the linguistics of mathematical and technical Documents. The problem is that many of the mathematical corpora (e.g., the arXiv corpus or the 3 Million abstracts of zbMATH[12]) are not available under a license that allows republishing. While the copyright owners are open towards research, they cannot afford to make the corpora public. SIGMathLing hosts such data sets in corpus cooperative: Researchers in mathematical semantics extraction and information retrieval sign a cooperative non-disclosure agreement, get access to the data sets and can deposit derived data sets in the cooperative. Data sets have dedicated landing pages so that they can be cited. A prime example of a data set is the XHTML5+MathML version of the arXiv corpus up to August 2019.[13]

[10] https://github.com/dginev/CorTeX.

[11] https://sigmathling.kwarc.info/.

[12] http://zbmath.org.

[13] The landing page is at https://sigmathling.kwarc.info/resources/arxmliv-dataset-082019/.

12.7 Conclusion

The NTCIR Math Tasks were an initial attempt in facilitating the formation of an interdisciplinary community of researchers interested in the challenging problems underlying MIR. The diversity of approaches reported at NTCIR shows that research in this field is active. We witnessed the progress of participating systems since the NTCIR-10 Pilot Task; improving scalability or addressing result ranking in new ways.

The design decision of the arXiv subask to exclusively concentrate on formula/keyword queries and use paragraphs as retrieval units made the retrieval task manageable but has also focused research away from questions such as result presentation and user interaction. In particular, few systems have invested in further semantics extraction from a corpus and used that in the search process to further address information needs. We feel that this direction should be further addressed in future tasks.

Ultimately, the success of MIR systems will be determined by how well they are able to accommodate user needs in terms of the adequacy of the query language, trade-off between query language expressiveness/flexibility, and answer latency on the one hand and learnability on the other. Similarly, the result ranking and monetization strategies for MIR are still a largely uncharted territory; we hope that future MIR tasks can help make progress on this front.

Acknowledgements The work reported in this chapter was partially supported by the Leibniz Association under grant SAW-2012-FIZ_KA-2, JSPS KAKENHI grant numbers 2430062 and 16H01756, and JST CREST number JPMJCR1513 and the National Science Foundation (USA) under grant no. HCC-1218801. We especially thank NTCIR Math co-organizers and collaborators, Iadh Ounis, Richard Zanibbi, Moritz Schubotz, and Goran Topić. We are also grateful to Kazuki Hayakawa and Takeshi Sagara for assisting with the task organization, Deyan Ginev who did most of the actual work in preparing the arXiv corpus for the NTCIR Math Tasks, Michal Růžička for generating XHTML files for the Wikipedia corpus, and Anurag Agarwal for assistance with designing the Wikipedia queries. Finally, we thank the students who evaluated the search hit pools at Jacobs University (arXiv) and Rochester Institute of Technology (Wikipedia).

References

Aizawa A, Kohlhase M, Ounis I (2013) NTCIR-10 Math pilot task overview. In: Kando N, Kato T (eds) Proceedings of the 10th NTCIR conference on evaluation of information access technologies, NTCIR-10, National Center of Sciences, Tokyo, Japan, National Institute of Informatics (NII). Accessed 18–21 June 2013

Aizawa A, Kohlhase M, Ounis I, Schubotz M (2014) NTCIR-11 Math-2 task overview. In: Kando N, Joho H, Kishida K (eds) Proceedings of the 11th NTCIR conference on evaluation of information access technologies, NTCIR-11, National Center of Sciences, Tokyo, Japan, National Institute of Informatics (NII). Accessed 9–12 Dec 2014

Ausbrooks R, Buswell S, Carlisle D, Chavchanidze G, Dalmas S, Devitt S, Diaz A, Dooley S, Hunter R, Ion P, Kohlhase M, Lazrek A, Libbrecht P, Miller B, Miner R, Sargent M, Smith B, Soiffer N, Sutor R, Watt S (2010) Mathematical markup language (MathML) version 3.0. W3C Recommendation, World Wide Web Consortium (W3C)

Barthel S, Tönnies S, Balke WT (2013) Large-scale experiments for mathematical document classification. In: Urs SR, Na JC, Buchanan G (eds) Digital libraries: social media and community networks. Springer International Publishing, Cham, pp 83–92

Davila K, Joshi R, Setlur S, Govindaraju V, Zanibbi R (2019) Tangent-V: Math formula image search using line-of-sight graphs. In: Advances in information retrieval - 41st European conference on IR research, ECIR 2019, Cologne, Germany, Proceedings, Part I, pp 681–695. https://doi.org/10.1007/978-3-030-15712-8_44. Accessed 14–18 Apr 2019

Guidi F, Sacerdoti Coen C (2016) A survey on retrieval of mathematical knowledge. Math Comput Sci 10(4):409–427. https://doi.org/10.1007/s11786-016-0274-0

Kohlhase M (2015) Formats for topics and submissions for the Math-2 task at NTCIR-12. Technical report, NTCIR. http://ntcir-math.nii.ac.jp/wp-content/blogs.dir/13/files/2014/05/NTCIR11-Math-topics.pdf

Kristianto GY, Topić G, Aizawa A (2017) Utilizing dependency relationships between math expressions in math IR. Inf Retriev J 20:132–167. https://doi.org/10.1007/s10791-017-9296-8

Lin J, Wang X, Wang Z, Beyette D, Liu JC (2019) Prediction of mathematical expression declarations based on spatial, semantic, and syntactic analysis. In: Proceedings of the ACM symposium on document engineering 2019. ACM, New York, NY, USA, DocEng '19, pp 15:1–15:10. https://doi.org/10.1145/3342558.3345399

Liska M, Sojka P, Ruzcka M, Mravec P (2011) Web interface and collection for mathematical retrieval: WebMIaS and MREC. In: Sojka P (ed) Towards Digital Mathematics Library, DML workshop. Masaryk University, Brno

Mansouri B, Rohatgi S, Oard DW, Wu J, Giles CL, Zanibbi R (2019) Tangent-CFT: An embedding model for mathematical formulas. In: Proceedings of the 2019 ACM SIGIR international conference on theory of information retrieval, ICTIR 2019, Santa Clara, CA, USA, pp 11–18. https://doi.org/10.1145/3341981.3344235. Accessed 2–5 Oct 2019

Mansouri B, Zanibbi R, Oard DW (2019) Characterizing searches for mathematical concepts. In: 2019 ACM/IEEE joint conference on digital libraries (JCDL), pp 57–66. https://doi.org/10.1109/JCDL.2019.00019

Schubotz M, Youssef A, Markl V, Cohl HS (2015) Challenges of mathematical information retrieval in the NTCIR-11 Math Wikipedia Task. In: Proceedings of the 38th International ACM SIGIR conference on research and development in information retrieval. ACM, New York, NY, USA, SIGIR '15, pp 951–954. https://doi.org/10.1145/2766462.2767787

Schubotz M, Greiner-Petter A, Scharpf P, Meuschke N, Cohl HS, Gipp B (2018) Improving the representation and conversion of mathematical formulae by considering their textual context. In: Proceedings of the 18th ACM/IEEE on joint conference on digital libraries, JCDL 2018, Fort Worth, TX, USA, pp 233–242. https://doi.org/10.1145/3197026.3197058. Accessed 03–07 June 2018

Stathopoulos Y, Teufel S (2015) Retrieval of research-level mathematical information needs: A test collection and technical terminology experiment. In: Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (vol 2: Short Papers), Association for Computational Linguistics, Beijing, China, pp 334–340. https://doi.org/10.3115/v1/P15-2055

Stathopoulos Y, Baker S, Rei M, Teufel S (2018) Variable typing: Assigning meaning to variables in mathematical text. In: Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies, vol 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, pp 303–312. https://doi.org/10.18653/v1/N18-1028

Suzuki T, Fujii A (2017) Mathematical document categorization with structure of mathematical expressions. In: 2017 ACM/IEEE joint conference on digital libraries (JCDL), pp 1–10. https://doi.org/10.1109/JCDL.2017.7991566

Youssef A (2017) Part-of-math tagging and applications. In: Geuvers H, England M, Hasan O, Rabe F, Teschke O (eds) Intelligent computer mathematics. Springer International Publishing, Cham, pp 356–374

Youssef A, Miller BR (2019) Explorations into the use of word embedding in math search and math semantics. In: Intelligent computer mathematics - 12th international conference, CICM 2019, Prague, Czech Republic, Proceedings, pp 291–305. https://doi.org/10.1007/978-3-030-23250-4_20. Accessed 8–12 July 2019

Zanibbi R, Blostein D (2012) Recognition and retrieval of mathematical expressions. Int J Doc Anal Recognit (IJDAR) 15(4):331–357. https://doi.org/10.1007/s10032-011-0174-4

Zanibbi R, Aizawa A, Kohlhase M, Ounis I, Topic G, Davila K (2016) NTCIR-12 MathIR task overview. In: Kando N, Sakai T, Sanderson M (eds) Proceedings of the 12th NTCIR conference on evaluation of information access technologies, National Center of Sciences, Tokyo, Japan, National Institute of Informatics (NII). Accessed 7–10 June 2016

Chapter 13
Experiments in Lifelog Organisation and Retrieval at NTCIR

Cathal Gurrin, Hideo Joho, Frank Hopfgartner, Liting Zhou, Rami Albatal, Graham Healy, and Duc-Tien Dang Nguyen

Abstract Lifelogging can be described as the process by which individuals use various software and hardware devices to gather large archives of multimodal personal data from multiple sources and store them in a personal data archive, called a lifelog. The Lifelog task at NTCIR was a comparative benchmarking exercise with the aim of encouraging research into the organisation and retrieval of data from multimodal lifelogs. The Lifelog task ran for over 4 years from NTCIR-12 until NTCIR-14 (2015.02–2019.06); it supported participants to submit to five subtasks, each tackling a different challenge related to lifelog retrieval. In this chapter, a motivation is given for the Lifelog task and a review of progress since NTCIR-12 is presented. Finally, the lessons learned and challenges within the domain of lifelog retrieval are presented.

C. Gurrin (✉) · L. Zhou · G. Healy
Dublin City University, Dublin, Ireland
e-mail: cathal.gurrin@dcu.ie

L. Zhou
e-mail: Liting.zhou@insight-centre.org

G. Healy
e-mail: graham.healy@dcu.ie

H. Joho
University of Tsukuba, Tsukuba, Japan
e-mail: hideo@slis.tsukuba.ac.jp

F. Hopfgartner
University of Sheffield, Sheffield, UK
e-mail: f.hopfgartner@sheffield.ac.uk

R. Albatal
SeekLayer Ltd, Dublin, Ireland
e-mail: ramibatal@gmail.com

D.-T. D. Nguyen
University of Bergen, Bergen, Norway
e-mail: ductien.dangnguyen@uib.no

T. Sakai et al. (eds.), *Evaluating Information Retrieval and Access Tasks*,
The Information Retrieval Series 43,
https://doi.org/10.1007/978-981-15-5554-1_13

13.1 Introduction

Recent advances in computing technology and wearable sensors mean that individuals are now in a position to log data about their lives on a continual basis, with little manual effort required. These data logs are often called lifelogs, and the process of gathering them is referred to as lifelogging. Lifelogging typically occurs in a passive manner (i.e. using sensors and not relying on human input). A commonly used definition of lifelogging is as *'a form of pervasive computing, consisting of a unified digital record of the totality of an individual's experiences, captured multimodally through digital sensors and stored permanently as a personal multimedia archive'* (Dodge and Kitchin 2007). Lifelogging can generate enormous (potentially multi-decade) archives that are too large for manual organisation. What sets lifelogging apart from conventional personal data organisation challenges (e.g. photos or emails) is the fact that lifelogs, being captured passively, are typically continuous in nature and are non-curated archives. Hence, these lifelogs pose a significant challenge for researchers to develop appropriate information organisation and retrieval approaches.

In the past 15 years, lifelogging has been receiving increasing research attention across a range of domains, including multimedia analytics, event-based computing, pervasive computing, information retrieval, as well as various application domains such as memory-science, wellness and epidemiological studies. A detailed overview of the early research activities on lifelogging is provided by (Gurrin et al. 2014b), and we refer the reader to that overview. Prior to NTCIR-12, there was no forum that could support a comparative evaluation of approaches to lifelog data organisation and retrieval, nor were the suitable datasets, nor was there even consensus on which of the many potential research challenges were the most important. The Lifelog task at NTCIR-12 was proposed because the organisers identified that lifelogging had potential to become a relatively commonplace activity, thereby necessitating the development of new approaches to multimodal personal data analytics and retrieval. New personal sensors were coming to market, such as wearable cameras, AR glasses, various forms of fitness trackers and so on. These were generating large multimodal archives for individuals yet, as with many new technologies, the required organisation tools had not been considered. It is the belief of the organisers that such vast archives of personal data require search and automatic annotation as fundamental underlying technologies upon which various other applications can be built; hence, the Lifelog task was proposed.

13.2 Related Activities

Lifelog data has been used in many domains as a source of multimodal data logging the real-world activities of one, or more, individuals. From prior research, we note that lifelogging tools were applied in the domains of long-term memory understanding (Milton et al. 2011), supporting human recollection (Barnard et al. 2011),

supporting human memory (Berry et al. 2007; Harvey et al. 2016), facilitating large-scale epidemiological studies in healthcare (Signal et al. 2017), lifestyle monitoring at the individual level (Nguyen et al. 2016; Wilson et al. 2018), behaviour analytics (Everson et al. 2019), diet/obesity analytics (Zhou et al. 2019), or for exploring societal issues such as privacy-related concerns (Hoyle et al. 2014). For many of these domains of application, the lifelog data was gathered and analysed by humans in order to draw conclusions for their research tasks.

In terms of actual functional retrieval systems for lifelog data, a number of early retrieval engines had been developed prior to NTCIR-12, such as the MyLifeBits system (Gemmell et al. 2002) or the Sensecam Browser (Lee et al. 2008). Both of these systems were browsing engines, rather than search engines, and relied on a database metaphor to support access. Subsequently, it was found that a faceted-multimodal search engine (even a simple one) was many times faster and more effective than browsing systems at finding known items from large lifelogs (Doherty et al. 2012), yet there were few search engines designed for lifelog data and no means of comparing their effectiveness. This means that prior to the Lifelog task at NTCIR-12, there were no comparative benchmarking activities and comparative and reproducible research on lifelogging was rather sparse. The main reason for this was the lack of publicly available lifelog datasets, which was due to the highly personal nature of lifelog data and the related requirement to guarantee people's privacy when releasing such datasets for widespread use.

The NTCIR-12 Lifelog pilot task (Gurrin et al. 2016) introduced the first shared test collection for lifelog data and attracted the first cohort of participants to, what was at the time, a very novel and challenging task. Since this initiative at NTCIR-12, there have been two related activities at alternative venues; one at ImageCLEF (Dang-Nguyen et al. 2017a, 2018) which focused on a series of image-retrieval and summarisation focused benchmarking initiatives since 2017, and the Lifelog Search Challenge (LSC) (Gurrin et al. 2019b) which was modelled on the successful Video Browser Showdown (Lokoc et al. 2018). The LSC encourages participants to develop interactive search engines for lifelog data and evaluate them in a public forum. The LSC has run at the annual ACM ICMR conference since 2018.

Specifically in relation to standalone retrieval efforts, early research on lifelog retrieval has focused on using *images* as unit of retrieval (e.g. Lee et al. 2008) with some early work in supporting user browsing these image collections (Doherty et al. 2011), or on the use of maps metadata, such as GPS locations, to organise content visually (Chowdhury et al. 2016). Once again, we refer the reader to (Gurrin et al. 2014b) for an overview of early efforts at lifelog search and retrieval. Significant efforts also went into the development of graphical user interfaces to visualise the data and also provide a positive user experience. Many good examples of interactive interfaces can be seen in the systems developed for the interactive Lifelog Search Challenge since 2018.

13.3 Lifelog Datasets Released at NTCIR

Over the course of the three most recent NTCIR workshops, the Lifelog task introduced three new datasets. The datasets were developed to represent a multimodal digital surrogate of the life activities of a number of individuals as they go about their daily lives, over an extended period of time (weeks or months). These datasets represented unprecedented data-rich archives for a number of individuals, pushing the boundaries of what was feasible to collect and distribute in an ethically and legally acceptable manner. Each dataset was gathered by either two or three lifeloggers, who wore/carried with them various lifelogging devices and gathered activity/biometric data for most (or all) of the waking hours in the day. The three datasets contained images from passive-capture wearable cameras as the core of each dataset. The passive-capture wearable camera was either clipped to clothing or worn on a lanyard around the neck, which captured images (from the wearer's viewpoint) and operated for 12–14 h per day (1,250–4,500 images per day—depending on capture frequency, camera type, or length of waking day). For examples of images captured by such wearable cameras, see Fig. 13.1. Additionally, mobile phone apps gathered contextual data such as location or physical movements and additional sensors (e.g. smartwatches or biometric-testing sensors) provided health and wellness data.

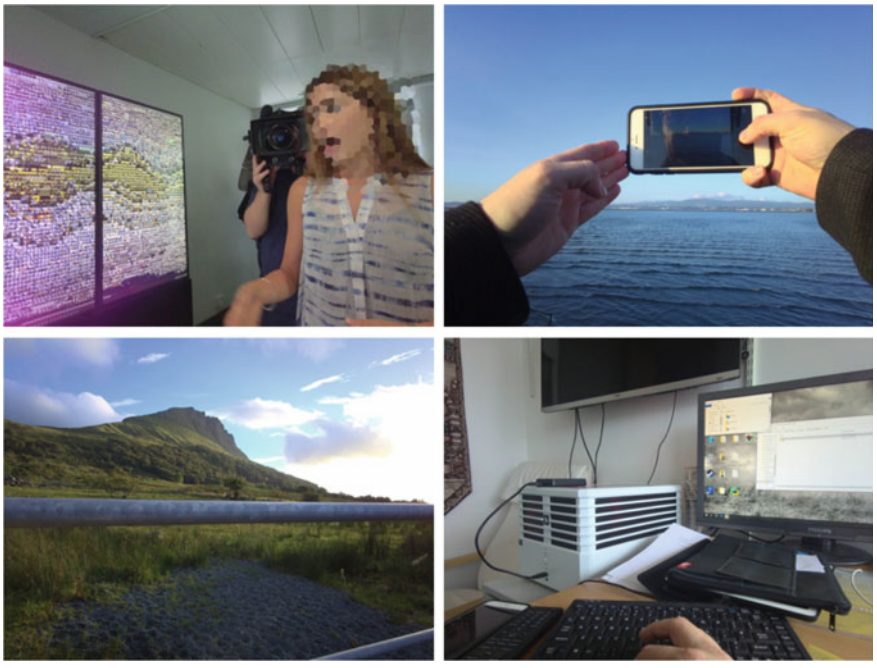


Fig. 13.1 Examples of Wearable Camera Images (Narrative Clip from NTCIR-13)

Typically, the datasets consist of:

- **Multimedia Content**: Wearable camera images captured at a rate of about two images per minute and worn from breakfast to sleep. Accompanying this image data for NTCIR-13/14 was a time-stamped record of music listening activities sourced from `Last.FM`[1] and (for NTCIR-14) an archive of all conventional (active-capture) digital photos taken by the lifelogger.
- **Biometrics Data**: Using off-the-shelf fitness trackers,[2] the lifeloggers gathered 24×7 heart rate, caloric burn and steps. In addition, for NTCIR-2014, continuous blood glucose monitoring was added which captured readings every 15 min using the Freestyle Libre wearable sensor.[3]
- **Human Activity Data**: The daily activities of the lifeloggers were captured in terms of the semantic locations visited, physical activities (e.g. walking, running, standing) from the Moves app,[4] along with (for NTCIR-14) a time-stamped diet log of all food and drink consumed.
- **Enhancements to the Data**: The wearable camera images were annotated with the outputs of various visual concept detectors which described in textual form the content of the lifelog images.

Readers who are interested in more information on the three lifelog datasets are referred to the task overview papers for NTCIR-12 (Gurrin et al. 2016), NTCIR-13 (Gurrin et al. 2017) and NTCIR-14 (Gurrin et al. 2019a). See Table 13.1 for a summary comparison of the three datasets.

What makes lifelog dataset generation a challenging task is the personal nature of real lifelog data (Chaudhari et al. 2007; Dang-Nguyen et al. 2017b) which must be gathered and released in a carefully organised process. One, or more, individuals must be willing to share a digital representation of their real-world activities with both researchers and the community. Aside from the difficulties of finding lifeloggers willing to share, various legal and institutional requirements needed to be met, such as passing review by an institutional ethics board, and for NTCIR-14, the preparation of a Data Protection Impact Assessment (to meet European GDPR requirements). Datasets were made available via the NTCIR-Lifelog website[5] and were password protected and secured by HTACCESS with username/password pairs generated for each participant. Additionally, in a style similar to TREC, each participating organisation needed an appropriate representative to sign an organisational agreement form and send it to the task organiser. Individual agreement forms were maintained by the participating organisation on behalf of each task participant within that organisation.

Prior to release, each dataset was subject to a detailed multi-phase redaction process to anonymise the dataset in terms of the lifelogger's identity as well as the identity of bystanders in the data. While many approaches have been proposed to

[1] Last.FM Music Tracker—https://www.last.fm/.

[2] For example, the Fitbit Fitness Tracker (FitBit Versa)—https://www.fitbit.com/.

[3] Freestyle Libre wearable glucose monitor—https://www.freestylelibre.ie/.

[4] Moves App for Android and iOS—http://www.moves-app.com/.

[5] NTCIR-Lifelog website—http://ntcir-lifelog.computing.dcu.ie/.

Table 13.1 Statistics of NTCIR lifelog datasets

Criteria	NTCIR-12	NTCIR-13	NTCIR-14
Number of Lifeloggers	3	2	2
Number of Days	90 days	90 days	43 days
Collection Size	18 GB	26 GB	14 GB
Number of Images	88,124 images	114,547 images	81,474 images
Number of Locations	130 locations	138 locations	61 locations
Physical Activities	Moves app	Moves app	Moves app
Calorie Burn	-	Fitness Watch	Fitness Watch
Step Count	-	Fitness Watch	Fitness Watch
Heart Rate	-	Chest Strap	Fitness Watch
Blood Glucose	-	Daily	Continuous
Music Listening	-	Last.FM	Last.FM
Cholesterol	-	Weekly	-
Uric Acid	-	Weekly	-
Diet Log	-	Manual	Manual
Conventional Photos	-	-	Smartphone

supporting privacy preservation in lifelog data (Gurrin et al. 2014a; Memon and Tanaka 2014), it was realised that none were effective enough to be deployed in an automated manner over lifelog data. Hence, a multi-step process was put in place that relied on manual (or semi-manual) redaction, and is summarised as follows:

- **Data Filtering:** Given the personal nature of lifelog data, it was necessary to allow the lifeloggers to remove any lifelog data that they may have been unwilling to share. This sharable data was then reviewed by a trusted member of the organising team and further deletions occurred where deemed prudent.
- **Privacy Protection:** Privacy-by-design (Cavoukian 2010) was a requirement for the test collection. Consequently, faces, readable screens and personal details (e.g. bank cards, passports) were blurred in either a fully manual or semi-automated process. Additionally, every image was resized down to 1024×768 resolution which had the effect of rendering most textual content illegible. Following this, a validation check was performed on the redaction outputs.

The overall data redaction and release process is summarised in Fig. 13.2, which shows the steps taken by the lifelogger (1), the organisers (2) and the responsibility on the task participants (3) who use the data for their experiments. As can be seen, the lifelogger gets the opportunity to review, filter and clean their data before the organisers carry out a secondary data review and cleaning, followed by the execution of a number of processes to ensure privacy of individuals associated with the dataset, followed by a final validation of the data before it is released for interested researchers who sign up to access the data.

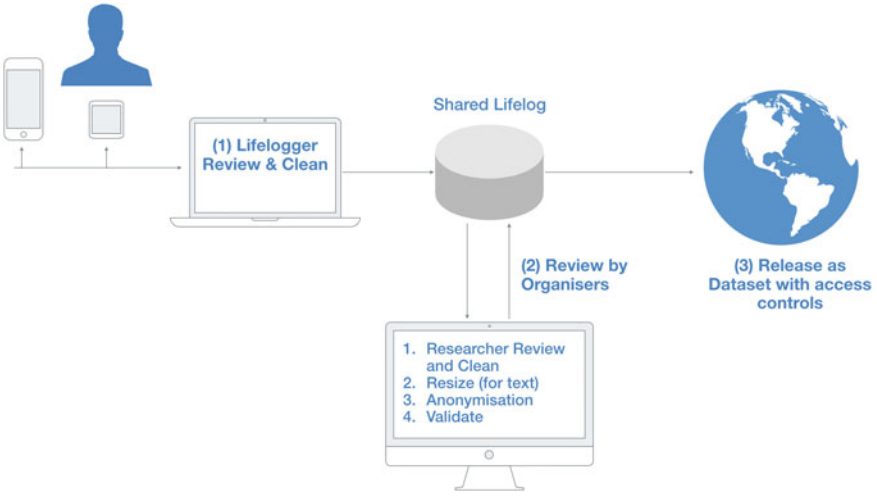


Fig. 13.2 Overview of the Redaction Process for the NTCIR Collections

13.4 Lifelog Subtasks at NTCIR

Based on the use cases described previously and guided by the human memory-access applications of Sellen and Whittaker (2010), five different challenges were explored at NRCIR-Lifelog. In this section, we focus on the two main subtasks that ran for all three Lifelog instances and we briefly describe the other three subtasks.

13.4.1 Lifelog Semantic Access Subtask

The Lifelog Semantic Access subtask (LSAT) was the core task of the three editions of the Lifelog task. The aim of the task was to explore ad hoc search and retrieval from lifelogs, which the organisers believe to be a fundamental enabling technology to make lifelogs a useful tool for individuals. In this subtask, the participants were required to retrieve a number of specific moments in a lifelogger's life in response to a topic description, as shown in Fig. 13.3. There were either 24 or 48 topics prepared for each instance of the task. For the purposes of evaluation, the organisers took the simplifying assumption that an image (point-in-time) is an appropriate document for retrieval. The task can best be compared to a known-item search task with one (or more) relevant items per topic. Evaluation was by means of standard evaluation measures and calculated using treceval.[6] For NTCIR-12 & NTCIR-13, full relevance judgements were prepared, but for NTCIR-14, pooled relevance judgements were

[6] https://trec.nist.gov/trec_eval/.

TITLE: Icecream by the Sea
DESCRIPTION: Find the moment when U_1 was eating icecream beside the sea.
NARRATIVE: To be relevant, the moment must show both the icecream with cone in the hand of u1 as well as the sea clearly visible. Any moments by the sea, or eating an icecream which do not occur together are not considered to be relevant.

EXAMPLES OF RELEVANT MOMENTS FOUND BY PARTICIPANTS

Fig. 13.3 LSAT Topic Example, including example results

used. Participants were allowed to undertake the LAST subtask in an interactive or automatic manner. For interactive submissions, a maximum of five minutes of search time was allowed per topic.

Over the three instances of the LSAT Task, we note that task participants took many different approaches to the development of retrieval systems. Given that there are no standardised baselines that can be applied, this is not surprising. Participating teams developed many different experimental systems, both interactive and automatic in nature. We look firstly at interactive retrieval engines over the three editions of NTCIR. At NTCIR-12, the participating team from University of Barcelona (Spain) developed the only interactive retrieval engine that integrated a semantic-content tagging tool to enhance the quality of the annotations (de Oliveira Barra et al. 2016). At NTCIR-13, the DCU team (Ireland) employed a human-in-the-loop to translate the provided queries into system queries for their retrieval engine, in one of their runs (Duane et al. 2017). However, at NTCIR-14, we note that three of the participants developed interactive systems and a fourth participant also integrated the human-in-the-loop query enhancement. NTU (Taiwan) developed an interactive lifelog retrieval system that automatically suggested to the user a list of candidate query words and adopted a probabilistic relevance-based ranking function for retrieval (Fu et al. 2019). They enhanced the official concept annotations and pre-processed the visual content to remove poor quality images and to offset the fish-eye nature of the wearable camera data. DCU (Ireland) developed an interactive retrieval engine for lifelog data (Ninh et al. 2019) that was designed for novice users and relied on an extensive list of facet filters over provided metadata. Finally, the VNU-HCM (Vietnam) group developed an interactive retrieval system (Nguyen et al. 2019) that used enhanced metadata

and visual enrichment, sometimes including human annotations. Their scalable and user-friendly interface to this system significantly outperformed competing systems at NTCIR-14, due primarily to the enhanced annotations. As expected, all interactive runs significantly outperformed the automatic runs at each edition of NTCIR-Lifelog.

In terms of approaches to automatic retrieval, at NTCIR-12, the VTIR (USA) team hypothesised that location was a very important component in the information retrieval process (Xia et al. 2016), and thus enhanced location semantic descriptions were used with the BM25 retrieval model. The authors comment that this approach worked well for some of the topics, which were location dependent. The IDEAS Institute for Information Industry (Taiwan) took a textual approach to retrieval (Lin et al. 2016) utilising word2vec to better match visual concepts to user queries (an approach referred to as bridging the lexical gap) via query expansion. The QUT group took an approach to retrieval that generated long, descriptive paragraphs of text to annotate the lifelog content, as opposed to the conventional tag-based approach (Scells et al. 2016); however, this was not shown to be successful. Finally, the LIG-MRM group (France) performed significantly better of all other approaches at NTCIR-12, by focusing on enhancing the performance of the visual concept detectors to be used for retrieval, and not relying on the provided (Caffe) classifier output (Safadi et al. 2016). The Caffe classifier provides a modifiable framework for state-of-the-art deep learning algorithms and a collection of reference models (Jia et al. 2014).

AT NTCIR-13, three participating groups took part in the LSAT subtask in an automated manner. DCU (Ireland) took part with their baseline search engine (Duane et al. 2017) that indexed the provided metadata and concepts using BM25 as the retrieval model, with both automated query runs and human-enhanced query runs. VCI2R (Singapore) proposed a general framework to bridge the semantic gap between lifelog data and the event-based LSAT topics (Lin et al. 2017) by enhancing the visual annotations and employing temporal smoothing of annotations, which proved to be the most successful approach at NTCIR-13. Finally, the PGB group (Japan) focused on the image and location data and enhanced the visual annotations (including people counting) and indexed locations using point-stay detection (D-Star algorithm) and integrated important location detection using the DBSCAN algorithm (Yamamoto et al. 2017). It performed better than the baseline, but not as well as the VCI2R and the human-in-the-loop run by DCU.

At NTCIR-14, NTU (Taiwan) submitted both interactive and automatic runs, and their automatic run (the top-ranked automatic run) included a query enhancement process using the top 10 nearest concepts to the query terms to expand the query before submitting the query (Fu et al. 2019). QUIK (Japan) from Kyushu University integrated online visual WWW content in the search process and operated based on an underlying assumption that a lifelog image of an activity would be similar to images returned from a WWW search engine for similar activities (Suzuki and Ikeda 2019). The approach operated using only the visual content of the collection and used the WWW data to train a visual classifier with a convolutional neural network for each topic. Although an automated process, a human-in-the-loop mechanism was employed to filter the WWW examples.

After NTCIR-14, the main approaches that the organisers consider to be valuable for lifelog access are the use of enhanced visual concept detectors to improve indexing, which has been continually shown to be effective both at NTCIR and the Lifelog Search Challenge (Gurrin et al. 2019b), as well as the application of approaches to bridging the lexical gap, either via some form of index term expansion or query expansion. Given the interest in developing interactive systems, the Lifelog Search Challenge is now the main venue for the comparative benchmarking of interactive lifelog retrieval systems.

13.4.2 Lifelog Insight Subtask

The Lifelog Insight subtask (LIT) also ran at all three editions of NTCIR-Lifelog and was designed to explore knowledge mining from lifelogs, with particular application in epidemiological studies. The LIT subtask was exploratory in nature, and the aim of this subtask was to gain insights into the lifelogger's daily life activities. It followed the idea of the Quantified Self movement that focuses on the visualisation of knowledge mined from self-tracking data to provide 'self-knowledge through numbers'. Participants were requested to provide insights that support the lifelogger in the act of reflecting upon their life, facilitate filtering, or provide for efficient/effective means of lifelog data visualisation. The LIT subtask was not evaluated in the traditional sense, rather all participants were asked to write about and bring their demonstrations or reflective output at the NTCIR conference.

At NTCIR-12, the Sakai Lab at Waseda University (Japan) developed a prototype smartphone application called Sleepflower, which was designed to improve the sleep cycles of a group of users (Iijima and Sakai 2016). A flower metaphor was displayed on the smartphone screen to represent the current sleepiness of a particular user, based on a manual analysis of the habits of the lifeloggers. Participants from Toyohashi University (Japan) examined repeated pattern discovery from lifelog image sequences, by applying a Spoken Term Discovery technique (Yamauchi and Akiba 2016) and a variant of Dynamic Time Warping was used in an experimental approach to extract meaningful patterns from the lifelog data. DCU (Ireland) introduced an interactive lifelog interrogation system which allowed for manual interrogation of the lifelog dataset for the occurrence of visual concepts that were assumed to match the information needs (Duane et al. 2016). The results of this manual interrogation were then used to generate insights and infographics.

At NTCIR-13, Tsinghua University (China) developed an approach to give insights into the big-five personality traits, moods, music moods, style detection and sleep-quality prediction (Soleimaninejadian et al. 2017). The team augmented the provided dataset with lifelog data gathered by other volunteers. The team found that their approaches achieved objective results with a high degree of accuracy, and noted the implications for improving traditional psychological research by employing lifelog data. Participants from the Institute for Infocomm Research (Singapore) presented a method for finding insights from the lifelog data by creating a topic-

focused minute-by-minute annotation of the user's activities (Xu et al. 2017). This was achieved by applying deep learning approaches for image analytics and then fusing the multimodal sensor data to generate insights into patterns and associations between lifelogger activities. The team from DCU (Ireland) introduced a new interactive lifelog interrogation system which was implemented for access in a Virtual Reality Environment (Duane et al. 2017). The system was designed to allow a user to explore visual lifelog data in an interactive and highly visual manner. Finally, the PGB group (Japan) developed an approach to automatically label the lifelog images with 15 concept labels (Yamamoto et al. 2017) using a DNN model with a fusion layer of tri-modal data (image, location and biometric).

At NTCIR-14, only one group took part in the LIT subtask. THUIR (China) developed a number of detectors for the lifelog data to automatically identify and visualise the status/context of a user (Nguyen et al. 2019) and a comparison between the various approaches showed that the visual features were significantly better than non-visual (metadata) features.

13.4.3 Other Subtasks (LEST, LAT and LADT)

A number of additional exploratory subtasks were run once (or twice) only. We will briefly describe these and comment on why they were not run in all three instances of the Lifelog task. The Lifelog Event Segmentation subtask (LEST) ran at NTCIR-13, the aim of which was to examine approaches to event segmentation from continual lifelog stream data (Gurrin et al. 2017). Event segmentation had been the typical approach to generation of indexable and retrievable documents (events) from lifelog collections. Given that the definition of an event is inherently subjective to the experience of the individual lifelogger, the organisers defined 15 types of events for the segmentation process, based on the 15 common lifestyle activities defined by Kahneman et al. (2004). The PGB group (NTT, Japan) participated in the LEST and developed a number of alternative approaches to event segmentation, included temporal visual similarity, user-linger-points, the use of LDA to reduce dimensionality and identify boundaries, and a multi-feature approach that used cosine similarity between segments (Yamamoto et al. 2017). The user-linger-points approach proved to be the most successful for event segmentation.

At NTCIR-14, this LEST morphed into the Lifelog Activity Detection subtask (LADT) at NTCIR-14 (Gurrin et al. 2019a), which required the classification of the multimodal lifelog data into one or more human activities that were identified as occurring in the lifelog collection. The NTU group (Taiwan) developed a new approach for the multi-label classification of lifelog images (Fu et al. 2019). In order to train the classifier, the authors manually labelled 4 days, which were chosen because they covered most of the activities that the lifeloggers were involved in.

However, the organisers note that there was little interest from the community in this task. This was surprising, since many of the previous applications of lifelog data to solve real-world challenges (e.g. healthcare or epidemiological studies) would

require the detection of human activities as a fundamental building block. Perhaps, this task will become very relevant and interesting at a later date, once lifelogging becomes a more commonplace activity for personal use or scientific enquiry.

It is worth noting that one outcome of this subtask was a new pilot task at NTCIR-15, which has a micro-activity detection/retrieval task (called MART) that extends this early work by focusing on the identification of short activities of daily life (e.g. writing an email, making a cup of coffee) and is targeted at the generation of rich and detailed semantic logs of everyday activities.

Finally, another exploratory subtask that ran at NTCIR-13 was the Lifelog Annotation subtask (LAT), which aimed to develop approaches for annotation of the multimodal lifelog data (images) with a fixed set of 15 high-level labels/concepts chosen from a manually generated ontology of lifelogging activities (Gurrin et al. 2017). These concepts were based on both the activities (facets of daily life) of the individual and the environmental settings (contexts) of the individual. Motivated by the realisation from NTCIR-12 that high-quality annotations are important for the retrieval process, the aim of this task was to provide various sets of high-quality shared annotations for all other uses to use in the LSAT subtask. However, only one group participated, so this annotation sharing did not occur. The PGB group (Yamamoto et al. 2017) developed a DNN model, with a fusion layer of tri-modal data (image, location and biometrics) to perform the content annotation. It was found that visual and biometric features can enhance the automatic annotation process, yet location actually was found to reduce annotation quality. Once again, this task was not attractive to NTCIR participants, so the Lifelog Activity Detection subtask (LADT) at NTCIR-14 replaced it.

13.5 Lessons Learned

Since NTCIR-12, 18 different research groups have taken part in the Lifelog task, some of them multiple times and across multiple tasks. Uptake on the subtasks suggests that the community is interested in the retrieval challenge and, to a lesser extent, the insights challenge. The other three challenges have not attracted much interest at this point. At the end of the NTCIR-Lifelog tasks, we can identify some lessons learned from the three editions of the NTCIR-Lifelog task:

- Novel Datasets: Eighteen participants submitted official runs to NTCIR, but at least three times as many downloaded the datasets. Even 4 years after starting the NTCIR-Lifelog task, requests for the datasets are still being received by the organisers. There is clearly an interest in the community to develop retrieval and analytics tools over such datasets, so there is significant potential for others in the community to define and release novel datasets of human life-experience data.
- Richer metadata: Repeatedly, we have seen that the best performing retrieval systems enhanced the provided metadata by relying on additional visual concept detectors, or seeking additional sources of metadata to enhance the retrieval performance. There is clearly a need to develop new approaches to the creation of

semantically rich metadata for multimodal lifelogs, in order to facilitate more effective retrieval algorithms.

- Bridge the lexical gap: Many participants found that there was a lexical gap between the terms used by the lifeloggers in their topic descriptions, and the indexed textual content and annotations. This suggests a need for term or query expansion, and the current consideration is that this could be achieved using approaches such as conventional query expansion or word embedding.
- Integrate external WWW content: This has been used by some participants with positive results. The external content helps to enhance the quality of content annotations or can be used as a form of query enhancement.
- There is an observed interest in the generation of insights or knowledge from lifelog data, as seen by the participation in the LIT subtask. This seems best suited to addressing the reflection and reminiscence use case of human memory as outlined by Sellen and Whittaker (2010).
- Document segmentation of the lifelog data into indexable content is as of yet an unsolved challenge. Initial attempts at lifelog 'event segmentation' (Lee et al. 2008) generated static documents for retrieval using an early sensor-based approach to segmentation. As with any information retrieval system, the concept of a document needs to be clearly defined and understood, which is not yet the case for lifelog data.
- Interactive search: Finally, interactive systems have been increasing in interest since NTCIR-12 and the Lifelog Search Challenge (Gurrin et al. 2019b) has been started to specifically explore this challenge. This appears to be the current hot topic for lifelog search and retrieval.

13.5.1 Conclusions and Future Plans

Over the course of the three instances of the NTCIR-Lifelog task, the uptake by participants was not as high as the organisers had hoped. One reason for this may be the emergence of a suite of parallel activities to motivate research into lifelogging and personal data analytics, such as the previously introduced interactive Lifelog Search Challenge (Gurrin et al. 2019b) and the ImageCLEF-Lifelog activities (Ionescu et al. 2018). The Lifelog Search Challenge in particular his been attracting 8–10 groups annually who come together to partake in a real-time interactive search challenge, which provides an open forum for all ACM ICMR conference attendees to partake as either observers or even as novice users in the competition. The ImageCLEF-Lifelog task tends to attract researchers more focused on the computer vision aspects of insight generation and data organisation and as such, it is targeting a slightly different audience. Regardless of the reasons, the uptake of the task and the level of interest in the dataset, along with the other related activities suggests a keen level of interest in the community for lifelog retrieval and the organisers note that this interest is likely to grow as volumes of personal multimodal data increase in society. The organisers understand that lifelog retrieval is a challenging activity,

and the future of the Lifelog task at NTCIR is perhaps in the refinement of the task to address key challenges in the domain, such as privacy-aware retrieval from personal multimodal data, epidemiological-scale analytics studies that analyse large lifelogs from multiple participants, targeted healthcare tasks of interest to concerned individuals and medical professions (e.g. finding medicine-taking events), or novel related-domains such as neural data retrieval.

It is an inevitable fact that the main challenge for any organisers of such tasks is the effort required to generate appropriate and real-world datasets and release them in an ethically and legally complaint manner. The three lifelog datasets released by the task organisers at NTCIR represent about a year of effort in total from a number of researchers and lifeloggers; this naturally incurs significant expenses in terms of organisers time and resources. Real-world use cases are likely to either focus on retrieval from longitudinal archives donated by one individual, or across large populations (as in epidemiological studies) and the data gathering and release methodology employed for this task was not ideal, due to the large overhead of effort required to ensure privacy preservation. The evaluation-as-a-service model proposed by Hopfgartner et al. (Hopfgartner et al. 2020) is one potential way forward, which brings the algorithms to the data, rather than the conventional data-to-algorithm approach. Another potential next step is to encourage more comparative evaluation of interactive systems, since a user of a lifelog tool (either an individual or a professional analyst) is most likely to be using such tools in an interactive manner. In any case, the organisers of the NTCIR-Lifelog tasks consider that this book chapter marks the end-of-the-beginning of research into lifelog data organisation and retrieval, rather than the conclusion of a short-lived sub-topic of IR. It is our belief that lifelogging as a topic will continue to become more popular for IR researchers and that the availability of relevant datasets and challenges will increase in the coming years.

Acknowledgements Many thanks to the editors and all authors of this book, and to the present and past organisers and participants of the NTCIR tasks. We also wish to acknowledge the data gatherers who kindly donated their time and data to the dataset releases. Finally, we acknowledge the support of Science Foundation Ireland under grant nos SFI 13/TIDA/I2875 and 11/RFP.1/CMS3283, and JSPS KAKENHI Grant Number 18H00974.

References

Barnard PJ, Murphy FC, Carthery-Goulart MT, Ramponi C, Clare L (2011) Exploring the basis and boundary conditions of SenseCam-facilitated recollection. Memory 19(7):758–767

Berry E, Kapur N, Williams L, Hodges S, Watson P, Smyth G, Srinivasan J, Smith R, Wilson B, Wood K (2007) The use of a wearable camera, SenseCam, as a pictorial diary to improve autobiographical memory in a patient with limbic encephalitis: a preliminary report. Neuropsychol Rehabil 17(4–5):582–601

Cavoukian A (2010) Privacy by design: The 7 foundational principles. implementation and mapping of fair information practices. Information and Privacy Commissioner of Ontario, Canada

Chaudhari J, Cheung SS, Venkatesh M (2007) Privacy protection for life-log video. In: IEEE workshop on signal processing applications for public security and forensics, 2007 SAFE '07, pp 1–5

Chowdhury S, Ferdous MS, Jose JM (2016) A user-study examining visualization of lifelogs. In: 2016 14th International Workshop on Content-Based Multimedia Indexing (CBMI), pp 1–6

Dang-Nguyen DT, Piras L, Riegler M, Boato G, Zhou L, Gurrin C (2017a) Overview of ImageCLE-Flifelog 2017: lifelog retrieval and summarization. In: CLEF2017 working notes, CEUR-WS.org, Dublin, Ireland, CEUR workshop proceedings

Dang-Nguyen DT, Zhou L, Gupta R, Riegler M, Gurrin C (2017b) Building a disclosed lifelog dataset: Challenges, principles and processes. In: Proceedings of the 15th international workshop on content-based multimedia indexing. ACM, New York, NY, USA, CBMI '17, pp 22:1–22:6

Dang-Nguyen DT, Piras L, Riegler M, Zhou L, Lux M, Gurrin C (2018) Overview of ImageCLE-Flifelog 2018: daily living understanding and lifelog moment retrieval. In: CLEF2018 working notes. CEUR workshop proceedings, Avignon, France. Accessed 10–14 Sept 2018

Dodge M, Kitchin R (2007) 'Outlines of a world coming into existence': pervasive computing and the ethics of forgetting. Environ Plann B: Plann Des 34(3):431–445

Doherty AR, Moulin CJa, Smeaton AF, (2011) Automatically assisting human memory: A Sense-Cam browser. Memory 19(7):785–795

Doherty AR, Pauly-Takacs K, Caprani N, Gurrin C, Moulin CJa, O'Connor NE, Smeaton AF (2012) Experiences of aiding autobiographical memory using the sensecam. Human-Comput Interac 27(1–2):151–174

Duane A, Gurrin C, Zhou L, Gupta R (2016) Visual insights from personal lifelogs - insight at the NTCIR-12 lifelog LIT task. In: Proceedings of NTCIR-12, Tokyo, Japan

Duane A, Zhou L, Dang-Nguyen DT, Gurrin C (2017) DCU at the NTCIR-13 lifelog-2 task. In: Proceedings of NTCIR-13, Tokyo, Japan

Everson B, Mackintosh KA, McNarry MA, Todd C, Stratton G (2019) Can Wearable Cameras be Used to Validate School-Aged Children's Lifestyle Behaviours? Children 6(2):20

Fu MH, Chia-Chun C, Huang GH, Chen HH (2019) Introducing external textual knowledge for lifelog retrieval and annotation. In: Proceedings of NTCIR-14, Tokyo, Japan

Gemmell J, Bell G, Lueder R, Drucker S, Wong C (2002) Mylifebits: Fulfilling the memex vision. In: Proceedings of the Tenth ACM International Conference on Multimedia, ACM, New York, NY, USA, ACM Multimedia '02, pp 235–238

Gurrin C, Albatal R, Joho H, Ishii K (2014a) A privacy by design approach to lifelogging. In: O Hara K, Nguyen C, Haynes P (eds) Digital enlightenment yearbook 2014. IOS Press, pp 49–73

Gurrin C, Smeaton AF, Doherty AR (2014b) LifeLogging: personal big data. Found Trends® Inf Retriev 8(1):1–125

Gurrin C, Joho H, Hopfgartner F, Zhou L, Albatal R (2016) Overview of NTCIR-12 lifelog task. In: Proceedings of NTCIR-12, Tokyo, Japan

Gurrin C, Joho H, Hopfgartner F, Zhou L, Gupta R, Albatal R, Dang-Nguyen DT (2017) Overview of NTCIR-13 lifelog-2 task. In: Proceedings of NTCIR-13, Tokyo, Japan, pp 6–11

Gurrin C, Joho H, Hopfgartner F, Zhou L, Ninh VT, Le TK, Albatal R, Dang-Nguyen DT, Healy G (2019a) Overview of NTCIR-14 lifelog-3 task. In: Proceedings of NTCIR-14, Tokyo, Japan

Gurrin C, Schoeffmann K, Joho H, Zhou L, Duane A, Leibetseder A, Riegler M, Piras L (2019b) Comparing approaches to interactive lifelog search at the lifelog search challenge (LSC2018). ITE Trans Media Technol Appl 7(2):46–59

Harvey M, Langheinrich M, Ward G (2016) Remembering through lifelogging: a survey of human memory augmentation. Pervasive Mobile Comput 27:14–26

Hopfgartner F, Gurrin C, Joho H (2020) Rethinking the test collection methodology for personal self-tracking data. In: 26th international conference on multimedia modeling 2020. Springer International Publishing

Hoyle R, Templeman R, Armes S, Anthony D, Crandall D, Kapadia A (2014) Privacy behaviors of lifeloggers using wearable cameras. In: Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing - UbiComp '14 Adjunct, pp 571–582

Iijima S, Sakai T (2016) SLLL at the NTCIR-12 lifelog task: sleepflower and the LIT subtask. In: Proceedings of NTCIR-12, Tokyo, Japan

Ionescu B, Müller H, Villegas M, de Herrera AGS, Eickhoff C, Andrearczyk V, Cid YD, Liauchuk V, Kovalev V, Hasan SA et al (2018) Overview of ImageCLEF 2018: Challenges, datasets and evaluation. International Conference of the Cross-Language Evaluation Forum for European Languages. Springer, Cham, pp 309–334

Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T (2014) Caffe: convolutional architecture for fast feature embedding. arXiv:1408.5093

Kahneman D, Krueger AAB, Schkade DA, Schwarz N, Stone AAA (2004) A survey method for characterizing daily life experience: the day reconstruction method. Science 306(5702):1776–1780

Lee H, Smeaton AF, O'Connor NE, Jones G, Blighe M, Byrne D, Doherty A, Gurrin C (2008) Constructing a SenseCam visual diary as a media process. Multimed Syst 14(6):341–349

Lin HL, Chiang TC, Chen LP, Yang PC (2016) Image searching by events with deep learning for NTCIR-12 lifelog. In: Proceedings of NTCIR-12, Tokyo, Japan

Lin J, del Molino AG, Xu Q, Fang F, Subbaraju V, Lim JH (2017) VCI2R at the NTCIR-13 lifelog semantic access task. In: Proceedings of NTCIR-13, Tokyo, Japan

Lokoc J, Bailer W, Schoeffmann K, Münzer B, Awad G (2018) On influential trends in interactive video retrieval: video browser showdown 2015–2017. IEEE Trans Multimed 20(12):3361–3376

Memon MA, Tanaka J (2014) Ensuring privacy during pervasive logging by a passerby. J Inf Process 22(2):334–343

Milton F, Muhlert N, Butler CR, Smith A, Benattayallah A, Zeman A (2011) An fMRI study of long-term everyday memory using SenseCam. Memory 19(7):733–744

Nguyen IVK, Shrestha P, Zhang M, Liu Y, Ma S (2019) THUIR at the NTCIR-14 lifelog-3 task: how does lifelog help the users status recognition. In: Proceedings of NTCIR-14, Tokyo, Japan

Nguyen THC, Nebel JC, Florez-Revuelta F (2016) Recognition of activities of daily living with egocentric vision: a review. Sensors 16(1):

Ninh VT, Le TK, Zhou L, Healy G, Venkataraman K, Tran MT, Dang-Nguyen DT, Smith S, Gurrin C (2019) A baseline interactive retrieval engine for the NTICR-14 Lifelog-3 semantic access task. In: Proceedings of NTCIR-14, Tokyo, Japan

de Oliveira Barra G, Ayala AC, Bolaos M, Dimiccoli M, Aghaei M, Carn M, Giro-I-Nieto X, Radeva P (2016) LEMoRe: a lifelog engine for moments retrieval at the NTCIR-lifelog LSAT task. In: Proceedings of NTCIR-12, Tokyo, Japan

Safadi B, Mulhem P, Quénot G, Chevallet JP (2016) MRIM-LIG at NTCIR lifelog semantic access task. In: Proceedings of NTCIR-12, Tokyo, Japan

Scells H, Zuccon G, Kitto K (2016) QUT at the NTCIR lifelog semantic access task. In: Proceedings of NTCIR-12, Tokyo, Japan

Sellen AJ, Whittaker S (2010) Beyond total capture: a constructive critique of lifelogging. Commun ACM 53(5):70–77

Signal LN, Stanley J, Smith M, Barr MB, Chambers TJ, Zhou J, Duane A, Gurrin C, Smeaton AF, McKerchar C, Pearson AL, Hoek J, Jenkin GLS, Ni Mhurchu C (2017) Children's everyday exposure to food marketing: an objective analysis using wearable cameras. Int J Behav Nutr Phys Activity 14(1):137

Soleimaninejadian P, Wang Y, Tong H, Feng Z, Zhang M, Liu Y, Ma S (2017) Bridging technology and psychology through the lifelog. In: Proceedings of NTCIR-13, Tokyo, Japan

Suzuki T, Ikeda D (2019) Smart lifelog retrieval system with habit-based concepts and moment visualization. In: Proceedings of NTCIR-14, Tokyo, Japan

Wilson G, Jones D, Schofield P, Martin DJ (2018) The use of a wearable camera to explore daily functioning of older adults living with persistent pain: methodological reflections and recommendations. J Rehabil Assist Technol Eng 5(205566831876):541

Xia L, Ma Y, Fan W (2016) VTIR at the NTCIR-12 2016 lifelog semantic access task. In: Proceedings of NTCIR-12, Tokyo, Japan

Xu Q, Subbaraju V, del Molino AG, Lin J, Fang F, Lim JH (2017) Visualizing personal lifelog data for deeper insights at the NTCIR-13 lifelog task. In: Proceedings of NTCIR-13, Tokyo, Japan

Yamamoto S, Nishimura T, Akagi Y, Takimoto Y, Inoue T, Toda H (2017) PBG at the NTCIR-13 lifelog-2 LAT, LSAT, and LEST tasks. In: Proceedings of NTCIR-13, Tokyo, Japan

Yamauchi K, Akiba T (2016) Repeated event discovery from image sequences by using segmental dynamic time warping: experiment at the NTCIR-12 lifelog task. In: Proceedings of NTCIR-12, Tokyo, Japan

Zhou Q, Wang D, Mhurchu CN, Gurrin C, Zhou J, Cheng Y, Wang H (2019) The use of wearable cameras in assessing children's dietary intake and behaviours in China. Appetite. https://doi.org/10.1016/J.APPET.2019.03.032

Chapter 14
The Future of Information Retrieval Evaluation

Douglas W. Oard

Abstract Looking back over the storied history of NTCIR that is recounted in this volume, we can see many impactful contributions. As we look at the future, we might then ask what points of continuity and change we might reasonably anticipate. Beginning that discussion is the focus of this chapter.

14.1 Introduction

In his book *The Third Wave*, Alvin Toffler placed what many have called the Information Age alongside the two most consequential transformations in human society, the introduction of agriculture, and the industrial revolution (Toffler 1980). That information retrieval will continue to play a central role in the coming years thus seems undeniable. One point of continuity between the current era and the flowering of science that helped to foster the industrial revolution is Lord Kelvin's admonition that "if you can not measure it, you can not improve it." Hence, the central role of information retrieval evaluation seems assured as well. That is not to say, however, that we will continue to measure our results in the same ways. Indeed, it seems reasonable to expect that information retrieval evaluation will continue to co-evolve along with changes in the information ecosystems that it serves. This chapter reflects on both the emergence of shared task evaluation and on present trends in information retrieval evaluation.

D. W. Oard (✉)
University of Maryland, College Park, MD, USA
e-mail: oard@umd.edu

T. Sakai et al. (eds.), *Evaluating Information Retrieval and Access Tasks*,
The Information Retrieval Series 43,
https://doi.org/10.1007/978-981-15-5554-1_14

14.2 First Things First

Shared task evaluation arose in information retrieval from the convergence of two broad lines of work. The first was the test collection tradition in information retrieval that dates back to the early Cranfield collections of the 1960s (Cleverdon 1991). The central idea in a test collection is to model the behavior of a user by selecting some representative set of documents[1] to be searched, generating representative search topics, generating representative queries for those search topics, and finally generating relevance judgments for some useful set of query-document pairs.

It was the need for relevance judgments that ultimately led to the creation of shared task evaluation for information retrieval. Many early collections were exhaustively judged (i.e., all query-document pairs had a relevance judgment), but as the document collections became larger exhaustive judgments proved to be infeasible. The challenge of larger collections was compounded by the emergence of search topics for which relatively few documents in the collection would be relevant. It was those topics seeking rare documents that made random sampling unsuitable as a means of dealing with increasing collection sizes. The approach that was ultimately adopted, pooling, relied on a form of purposeful sampling in which samples were drawn only from document sets in which existing retrieval systems had difficulty distinguishing between documents that were relevant and documents that were not. Ranked retrieval was becoming an increasingly widespread object of study at the time the idea of pooling was first tried in the Text Retrieval Conference, so this approach to sampling was generally operationalized as merging sets of documents that were highly ranked by one or more of several representative ranked retrieval systems (Voorhees and Harman 2005). It was this need for contributions of results from a number of representative systems that led to the emergence of shared task information retrieval evaluation.

In the movie *The Right Stuff* about the early American space program, one of the characters observes the importance of financial support with the pithy quote "No bucks, no Buck Rogers." Shared task evaluation requires resources for planning and coordination, but most essentially for creating the relevance judgments. This side of the equation came from the Defense Advanced Research Projects Agency (DARPA) in the United States, where the voice of Lord Kelvin was strong. The competition for funding within DARPA was adjudicated in part using the "Heilmeier Catechism," a set of questions to be answered by any new program, one of which is "What are the mid-term and final 'exams' to check for success?" DARPA had started a human language technology program, focusing initially on speech recognition, in 1986. Central to that program was a focus on evaluation. By 1990, DARPA was ready to expand its focus to include information retrieval. Hence was born the TIPSTER program, which in turn supported the early years of the Text Retrieval Conference (TREC).

As is sometimes the case when innovating, shared task evaluation rapidly evolved well beyond its initial focus on measurement. TREC did indeed produce test collec-

[1] Although it is conventional to refer to documents, the term is often used inclusively to refer to other types of information objects as well.

tions. Importantly those collections were shown to be reusable to a useful degree, thus permitting test collections developed in one year to be used in subsequent years as a basis for testing refinements to the system design. This approach, which came to be called evaluation-guided research, emerged in parallel in several research communities (e.g., information retrieval, speech recognition, and named entity recognition). It would be well recognized by machine learning researchers today as an early instance of supervised learning (albeit one with substantial human intervention in the early days). A second important thing that TREC did was that it produced baseline results to which future results could be compared. This facilitated the entry of new research teams, who could compare their systems against established baselines. A third innovation was the emergence in 1996 of TREC's more narrowly focused "tracks" to support specific research goals. These three innovations—collections, comparisons, and communities—together serve as a useful frame for examining not just shared task evaluation in TREC, but approaches to information retrieval evaluation more generally.

Much has been written about the benefits of shared task evaluation, but when considering alternatives it is equally important to consider its limitations as well. Perhaps most obviously, shared task evaluation is expensive. For example, the cost of the first 18 years of TREC, was calculated to be $29 million USD (Tassey et al. 2010), which is clearly well beyond what many individual researchers could support on their own. Two natural results of this are that some process for making investment decisions is needed, and those decisions must initially be made before seeing what the results will be. Those facts, in turn, tend to result in multi-year commitments to a research program so that insights generated in one year can be capitalized upon in the subsequent years. As a result, shared task evaluations have a limited capacity to start on new lines of work. Perhaps even more importantly, the need for some decision process, whether centralized or consensus-based, results in there being some gatekeeper role beyond the individual researcher that must judge whether a broad line of research merits the community's attention. Moreover, schedule considerations result in proposals needing to be made early—typically more than a year before the first results will become available. None of these limitations are show stoppers for research problems that require large-scale "team science" experimentation, but there are many settings (e.g., commercial research on problems with immediate operational implications, or a single student working alone on a novel problem in a 3-year Ph.D. program) for which shared task evaluation is not sufficiently responsive.

A second critique of shared task evaluation is that it can generate a tendency toward convergence in methods, perhaps thereby delaying the exploration of important alternative approaches. To see an example of this, we need to look no further than the current widespread interest in neural "deep learning" methods. This sort of bursty convergence in which new techniques are rapidly explored by the community has benefits, but the degree of convergence that in engenders has risks as well. Importantly, this risk is not unique to shared task evaluations—it is simply the flip side of any approach in which researchers come together as a community to compare results in an evaluation-guided research setting.

14.3 The Shared Task Evaluation Ecosystem

In the two decades that followed TREC's creation, shared task evaluation expanded at an impressive pace. Some notable examples (with the year in which they started) include the following:

- TDT (1996): The Topic Detection and Tracking (TDT) evaluation formed as a parallel evaluation venue to TREC to focus on streaming news content in text and speech (Wayne 2000).
- NTCIR (1999): The focus of this volume, NTCIR formed as a counterpart to TREC with a focus on East Asia.
- CLEF (2000): Initially called the Cross-Language Evaluation Forum, CLEF initially spun out from the TREC CLIR track (Braschler and Peters 2004).
- INEX (2002): The Initiative for Evaluation of XML Retrieval (INEX) formed independently to focus on retrieval of structured documents, and ultimately became a task in CLEF (Lalmas and Tombros 2007).
- TRECVID (2003): The TREC Video Retrieval Evaluation (TRECVID) is a separate evaluation venue that initially spun out from the TREC Video Track (Smeaton et al. 2006).
- MIREX (2005): The Music Information Retrieval Evaluation eXchange (MIREX) implemented a large-scale infrastructure for evaluation, using algorithm deposit to accommodate copyright concerns (Downie et al. 2014).
- FIRE (2008): The Forum for Information Retrieval Evaluation (FIRE) has a focus on South Asia (Majumder et al. 2018).
- MediaEval (2010): The MediaEval Benchmarking Initiative for Multimedia Evaluation initially spun out from the CLEF VideoCLEF Task (Larson et al. 2017).

No such list could ever be complete, since shared task evaluation exists any time two or more research groups come together around an evaluation task. For example, several evaluations have been conducted in a national context, including in China, France, Russia, and South Korea. Moreover, the boundaries between information retrieval and the cognate disciplines of natural language processing and speech processing are porous, and there have been evaluations in those communities that certainly bear on information retrieval research. For example, there have been evaluations of both event detection and summarization in the Text Analysis Conference (TAC),[2] and there has been evaluation of spoken term detection in the Open Keyword Search evaluation,[3] both of which are, like TREC, organized by the National Institute of Standards and Technology (NIST).

All of those are TREC-like, in that they are evaluation venues independent of any larger event, in which participants actually come together in a workshop-like setting to discuss their results. There are, however, numerous additional examples in which

[2] https://tac.nist.gov/

[3] https://www.nist.gov/itl/iad/mig/open-keyword-search-evaluation

one or both of those characteristics are not present. Cases in which a shared task evaluation is organized in conjunction with a larger event are sometimes called "data challenges." The granddaddy of these data challenges was perhaps SensEval, named for its focus on Word Sense Disambiguation. SensEval initially formed independently in 1998, but then associated itself with a workshop starting in 2001 (and later changed its name to SemEval in 2007, reflecting its broader interest in semantics).[4] The Conference on Computational Natural Language Learning (CoNLL) started a shared task in 1999,[5] followed in 2001 by the Document Understanding Conference (DUC, which despite its name was actually a workshop series, initially held at SIGIR). SemEval and the CoNLL shared task continue as data workshops to this day, having been joined by many others (e.g., the Big Data Cup[6]); DUC ultimately became a standalone venue (as TAC).

If data challenges are one step away from independent shared task evaluations such as NTCIR and TREC, prize-based competitions represent an even further departure from the independent conference paradigm. Perhaps the best known members of this genre of shared task evaluation are Kaggle[7] and the Netflix Prize (Bennett et al. 2007). The Netflix Prize started in 2007 with the goal of advancing research on large-scale recommender system. Netflix, a provider of streaming video services, offered participants access to a large collection of anonymized usage data, offering a $1 million USD reward for achieving a 10% improvement over the company's best current algorithm. Kaggle was founded in 2010 to capitalize on similar opportunities for a broad range of problems, acting as a forum within which communities could form around specific challenges. Kaggle has in turn given rise to other similar venues, including Tianchi[8] and Innocentive.[9] Prize competitions often operate as a market in which sponsors define the task and then pay the prize in exchange for a license to commercially use the technique that wins the competition. This stands in sharp contrast to the non-commercial ethos of many of the independent shared task evaluations listed at the start of this section, which focus principally on pre-competitive basic research. Indeed, some of the independent shared task evaluation venues actively seek to minimize the competitive aspect of shared task evaluation, in part because of concerns that a "winner-take-all" perspective might depress participation by teams who would otherwise be able to contribute diversity to the document pools that will be judged for relevance.

[4] https://aclweb.org/aclwiki/SemEval_Portal

[5] http://www.conll.org/previous-tasks

[6] http://cci.drexel.edu/bigdata/bigdata2019/BigDataCupChallenges.html

[7] https://www.kaggle.com/

[8] https://tianchi.aliyun.com/

[9] https://www.innocentive.com/

14.4 A Brave New World

In the movie *The Wizard of Oz*, Dorothy observes at one point that "we're not in Kansas anymore." So it is with information retrieval evaluation as well—there are now many more things under the sun than just shared task evaluation. At least four alternatives can be discerned, each of which has its own strengths and weaknesses.

The first to emerge were project data repositories. Perhaps the best known of these is the Linguistic Data Consortium (LDC) at the University of Pennsylvania,[10] which was founded in 1992 with support from DARPA to serve as a repository for the human language technology community. LDC and similar organizations around the globe (e.g., the European Language Resources Association, ELRA,[11] or the Linguistic Data Consortium for Indian Languages, LDC-IL[12]) permit researchers to deposit test collections that they have created that may in the future be of use to others. In this way, what were once internal evaluations on data generated within a project can become shared, and over time can emerge as a shared task reference to which future work can be compared. Perhaps the most successful example of this general approach is the University of California Irvine Machine Learning Repository (Dua and Graff 2017), which provides test collections that serve as standard references among machine learning researchers (notably including some text classification researchers).

Project data repositories help with community formation and with providing a basis for comparisons, but (at least when serving solely as repositories) they do not create collections. That's where crowdsourcing comes in. Shared task evaluations in the TREC heritage predate the World Wide Web, but as user-generated content became more pervasive in what came to be called Web 2.0, crowdsourcing emerged as an alternative way of obtaining relevance judgments (Alonso 2019). Crowdsourcing can be used in many ways in the evaluation of information retrieval systems, but perhaps the most obvious alternative to the approach used in shared task evaluation is to simply pay crowdworkers to create relevance judgments. Because queries are often treated as independent in information retrieval test collections, the relevance judgment task is easily distributable across multiple crowdworkers. At least two concerns arise when this is done. First, crowdworkers may be less well trained or less attentive to their task than relevance assessors who work at a central facility as their primary job would be. This concern has spawned a line of work on assessing the accuracy of crowdworkers. Second, one common approach to managing those risks, having several crowdworkers vote on the correct relevance label, has the effect of subtly redefining relevance (for purposes of evaluation) away from the opinion of an individual and toward the consensus of a group. Balanced against these concerns, however, are the speed, scalability, and relative affordability of crowdsourcing. Moreover, the diversity of available crowdworkers can provide access to people with needed skills (e.g., language skills or some types of topic expertise) that simply might not be available otherwise. For these reasons, crowdsourcing can offer transforma-

[10] https://www.ldc.upenn.edu/

[11] http://www.elra.info/en/

[12] http://www.ldcil.org/

tional advantages to isolated researchers who, for reasons of location, funding, or problem specificity simply cannot plausibly create a shared task evaluation. Note, however, that crowdsourced test collections need not remain isolated once they have been created, since they can be shared through data repositories.

Creating test collections is, however, just one of at least two ways in which crowdsourcing can be used for information retrieval evaluation. An alternative is to study the actual use of a system using crowdworkers. Test collections have many desirable attributes, but no test collection captures every important aspect of actual information retrieval tasks. Evaluating information retrieval systems in actual use has traditionally been a focus of user studies, and crowdsourcing offers an opportunity to extend the user study beyond the researcher's laboratory across the Internet to meet the users where they are. This opens new opportunities to intermix research using test collections (which are optimized for affordably repeatable evaluation under controlled conditions) and user studies (which offer higher fidelity evaluation, but at incremental cost each time an experiment is run).

There are, of course, limits to the user studies that can be run with crowdworkers. In addition to the obvious limits imposed by affordability considerations, fidelity is always a concern when paying a user to perform a task that you have designed. One way of addressing both of these concerns is to perform what has come to be called online evaluation (Radlinski and Craswell 2010). The basic approach is simple. First, build a system that becomes so popular that there will be a large number of users whose behavior you can study. Then design experiments in which some aspect of the system (the independent variable) is changed, and the effect is observed by observing some behavioral signal (the dependent variable). Variants on this idea include A-B testing and interleaving. Of course, the first step there—creating systems that have a large user population—can be a tad expensive! But once such a system is available, a very large number of experiments can be run at low cost. Naturally, this approach is popular among commercial services that have a large user base. Batch evaluation measures have also been tuned using query logs, thus more closely linking online and offline (i.e., batch) evaluation (Ferrante et al. 2014).

14.5 Trendlines

One thing that should be clear from the story to this point is that independent shared task evaluations such as NTCIR are now just one part of an increasingly diverse and specialized evaluation ecosystem. But that is just one of many trendlines that together will continue to reshape the future of information retrieval evaluation. This section reviews several others.

It is fashionable today in many contexts to remark on convergence. What used to be separate devices (e.g., phones, computers, and televisions) now are one. What used to be stored on separate media (video, images, documents, datasets) are now all stored as digital files. What used to be separate functions (computing and communication) are now becoming nearly inseparable. All of these are examples of

convergence. We are seeing examples as well of convergence across fields. Information retrieval researchers use speech and language technologies that in an earlier time would have been thought of as separate fields. Database researchers work with semi-structured data that the information retrieval community would recognize as structured documents. Data scientists analyze interaction patterns to help optimize the user experience. Interactive information retrieval research draws in equal measure on insights from information retrieval and human–computer interaction. Work on fairness, accountability, and transparency in machine learning finds application in designs of information retrieval systems that are informed as much by social as by technical goals. This convergence of disciplines creates new opportunities, but at the same time it challenges the notions we have developed over time about what is, and what is not, information retrieval.

If convergence disrupts what it is we think we do, the Internet is perhaps even more disruptive because it changes where we can do it. In an earlier era, information retrieval research suffered from what we might call the tyranny of geography. There were a few places in the world where top flight information retrieval research was going on, and it was much easier to get into the field if you could get to one of those places. Today, information retrieval is taught in many places, and indeed well over half the world's population has access to free online courses on the topic. Cloud computing has gone some distance toward democratizing access to high-end computing, and the widely available low-end computing infrastructure has capabilities that were unavailable anywhere on Earth just a few decades ago. We have by no means completely erased the tyranny of geography at this point in history, but it is quite clearly on the wane.

Solving one problem often reveals another, and so it is with the competition for our attention. For essentially all of human history, and with rare exception, information was scarce and human attention was relatively abundant. No one with an Internet connection can fail to notice that the situation today has sharply reversed, and that it is information that is abundant, while it is human attention that is now scarce. If we view our job as helping to separate the wheat from the chaff, it should be clear that this trendline suggests that we'll have no shortage of important problems to work on.

Another trendline worthy of remark is that the nature of gatekeeping is shifting. Long ago we had to choose between a Web track, a filtering track, an interactive track, or whatever other ideas were put forward, because venues like NTCIR simply could not do everything. It's still not possible to do everything, but the emergence of options such as crowdsourcing and online evaluation greatly expand the range of information retrieval evaluations that can be conducted. That's not to say that there will be no gatekeepers. Peer review, for example, will continue to play some role with regard to what gets published. But to the extent that some of the gatekeeping can be shifted from before the work is done to after the results become available, that could help to enhance the diversity of the research ecosystem.

One foundational assumption in information retrieval is that information wants to be found, and that our job is to find it. That's actually probably not true for much of the information in the world, however. Examples abound of information that should not be found. In Europe, the right to be forgotten is a right not to have specific infor-

mation about you found. In many countries with legislation that promotes freedom of access to government information, specific exemptions identify types of information that should not be disclosed. We have debates today about which types of information governments or commercial entities should be allowed to use, and for what purposes. Article 12 of the Universal Declaration of Human Rights declares privacy to be a human right, with all of the complexity that operationalizing the meaning of such a statement entails. In an earlier era, information retrieval research encountered restrictions on access from time to time, and in such cases the response of researchers was generally to focus instead on the many cases in which access control was not a problem.

We are perhaps now nearing the limits of that strategy. Consider the fact that almost all of the words produced on the planet—probably upward of 99%—are spoken, not written. Couple that with the fact that well over half that speech is produced in the presence of a networked recording device (e.g., a mobile phone). And couple that with the fact that both the speed and accuracy of technology for automatically transcribing that speech has improved by leaps and bounds in recent years. At present, we are largely disregarding all of that content simply because we have no idea how to protect those parts that need to be protected. This has implications for research, of course, but it has implications for evaluation design as well. We have grown up in an era in which we all learned to respect copyright when dealing with test collections. We now need to learn how to deal with sensitive content that will in some cases prevent us from distributing test collections. That does not mean that we won't be able to do shared task evaluations, but it does mean that we'll need to think anew about how best to do them. The Netflix Prize, for example, ended because of a privacy lawsuit.

It has been said that "data is the new oil," a catchy phrase intended to illustrate that there is money to be made. At one time, most information retrieval researchers worked in universities. Today, the balance has shifted very strongly in favor of industry. That's good news, because that's where the money is, so there is now vastly more research on information retrieval being published than ever before. It is also good news because industry has access to evaluation opportunities that simply can't be replicated elsewhere, most notably with online evaluation. And it is also good news because all this commercial activity is helping to bring new problems to the attention of the information retrieval research community.

14.6 An Inconclusion

It is traditional to end a chapter with a conclusion, but when writing about the future perhaps it would be wise to recognize that the evidence we see today is not sufficiently conclusive to allow us to see that future with clarity. Herewith, therefore, some inconclusive remarks. Josef Schumpeter is best known for his description of creative destruction, a process by which innovations result in the displacement of earlier enterprises that had been built to leverage earlier innovations (Schumpeter 1942). As the convergence examples above indicate, creative destruction is at least

as vibrant today as it was when Schumpeter was writing. Independent shared task evaluations such as NTCIR were created in an earlier era, to fill a role that has since been augmented, and perhaps partially replaced, by other approaches to information retrieval evaluation. It therefore seems timely to consider the question of what role NTCIR, and other independent shared task evaluations, may play in the future. Fortunately, the very name of NTCIR, the NII Testbeds and Community for Information Access Research, can help to guide that discussion.

N is for NII, the National Institute of Informatics. NII, like NACSIS before it, has been a source of leadership, not just in information retrieval evaluation, but in the emergence of a vibrant information retrieval research community in Japan specifically, and in East Asia more generally. Ultimately, NII is made up of people, and it is the choices made by those people that will define the future leadership role of that institution. With wise choices, that N will remain a capital letter.

T is for Testbeds. As explained throughout this chapter, the testbeds of the sort NTCIR has created (principally, test collections) are one part of what is now a rich ecosystem of evaluation methods. There will surely continue to be demand for test collections, but shared task evaluations like NTCIR are no longer the only affordable way in which test collections can be created, and we now live in a world in which a broader range of testbeds can be affordably constructed. We therefore may see the T in NTCIR decline somewhat in its impact, perhaps becoming a lower case t.

C is for Communities. For all the trendlines that portend change, one thing that seems unlikely to change any time soon is human nature. Humans are social animals, and research is a social enterprise. We need ways of bringing people together around new problems, ways of helping new people to join those communities, ways of creating the kinds of shared understanding that are needed to learn from each other how best to solve those problems, and ways of defining what it would mean to succeed at solving those problems. Shared task evaluations like NTCIR serve all of those functions. The C in NTCIR seems destined to remain a capital letter.

I is for Information access. As noted at the start of this chapter, we live in an information age, and it therefore seems unlikely that the focus of NTCIR on information would diminish. The same might not be said for access, however, since we are now seeing some convergence of research on (at least) information access, information creation, information understanding, information manipulation, and information policy. So the I in NTCIR seems sure to remain capitalized, but we may see some shifts in what it stands for.

R is for Research. We might think of research in three ways. The most obvious is to think narrowly in terms of some specific type of research, such as evaluation-guided research or statistical hypothesis testing. An alternative is to think of research more inclusively, as any systematic way of generating new and generalizable knowledge. And a third alternative would be to think even more broadly about research, as an undergraduate student might, as self-directed learning about new things. Many people who do not see themselves as researchers in the first or second sense need to do research in the third sense. One way or another, the R seems likely to remain since it is central to the self-image of NTCIR, but perhaps the meaning of that R will shift somewhat over time.

Well, there we have it. It seems that we can look forward to a world in which NtCIR remains, and all we will need to do is to figure out what it actually stands for!

Acknowledgements This work has been supported in part by NSF grant 1618695. The author is grateful to Bonnie Dorr, Nicola Ferro, Donna Harman, Gareth Jones, Noriko Kando, Martha Larson, Prasenjit Majumder, Paul McNamee, Jim Mayfield, Mandar Mitra, Carol Peters, Mark Sanderson, Ian Soboroff, Ellen Voorhees, and Charles Wayne for discussions over the years that have helped to shape his thinking on this topic.

References

Alonso O (2019) The practice of crowdsourcing. Synth Lect Inf Concepts Retr Serv 11(1):1–149

Bennett J, Lanning S, et al (2007) The Netflix prize. In: Proceedings of KDD cup and workshop, New York, NY, USA., vol 2007, p 35

Braschler M, Peters C (2004) Cross-language evaluation forum: objectives, results, achievements. Inf Retr 7(1–2):7–31

Cleverdon CW (1991) The significance of the Cranfield tests on index languages. In: Bookstein A, Chiaramella Y, Salton G, Raghavan VV (eds) Proceedings of the 14th annual international ACM SIGIR conference on research and development in information retrieval. Chicago, Illinois, USA, October 13–16, 1991 (Special Issue of the SIGIR Forum), ACM, pp 3–12. https://doi.org/10.1145/122860.122861

Downie JS, Hu X, Lee JH, Choi K, Cunningham SJ, Hao Y (2014) Ten years of MIREX (music information retrieval evaluation exchange): reflections, challenges and opportunities. In: Wang H, Yang Y, Lee JH (eds) Proceedings of the 15th international society for music information retrieval conference, ISMIR 2014, Taipei, Taiwan, October 27–31, 2014, pp 657–662. http://www.terasoft.com.tw/conf/ismir2014/proceedings/T119_342_Paper.pdf

Dua D, Graff C (2017) UCI machine learning repository. http://archive.ics.uci.edu/ml

Ferrante M, Ferro N, Maistro M (2014) Injecting user models and time into precision via Markov chains. In: Geva S, Trotman A, Bruza P, Clarke CLA, Järvelin K (eds) The 37th international ACM SIGIR conference on research and development in information retrieval, SIGIR '14, Gold Coast, QLD, Australia - July 06–11, 2014, ACM, pp 597–606. https://doi.org/10.1145/2600428.2609637

Lalmas M, Tombros A (2007) Evaluating XML retrieval effectiveness at INEX. SIGIR Forum 41(1):40–57. https://doi.org/10.1145/1273221.1273225

Larson M, Soleymani M, Gravier G, Ionescu B, Jones GJ (2017) The benchmarking initiative for multimedia evaluation: mediaeval 2016. IEEE MultiMedia 24(1):93–96

Majumder P, Mitra M, Sankhavara J, Mehta P (eds) (2018) Proceedings of the 10th annual meeting of the Forum for Information Retrieval Evaluation, FIRE 2018, Gandhinagar, India, December 06–09, 2018. ACM. https://doi.org/10.1145/3293339

Radlinski F, Craswell N (2010) Comparing the sensitivity of information retrieval metrics. In: Crestani F, Marchand-Maillet S, Chen H, Efthimiadis EN, Savoy J (eds) Proceeding of the 33rd international ACM SIGIR conference on research and development in information retrieval, SIGIR 2010, Geneva, Switzerland, July 19–23, 2010, ACM, pp 667–674. https://doi.org/10.1145/1835449.1835560

Schumpeter JA (1942) Capitalism, socialism and democracy. Routledge, Abingdon

Smeaton AF, Over P, Kraaij W (2006) Evaluation campaigns and trecvid. In: Proceedings of the 8th ACM international workshop on multimedia information retrieval, ACM, pp 321–330

Tassey G, Rowe BR, Wood DW, Link AN, Simoni DA (2010) Economic impact assessment of NIST's Text REtrieval Conference (TREC) Program. RTI International

Toffler A (1980) The third wave. Morrow

Voorhees EM, Harman DK (2005) TREC: experiment and evaluation in information retrieval. MIT Press, Cambridge

Wayne CL (2000) Multilingual topic detection and tracking: successful research enabled by corpora and evaluation. In: Proceedings of the second international conference on language resources and evaluation, LREC 2000, 31 May–June 2, 2000, Athens, Greece, European Language Resources Association. http://www.lrec-conf.org/proceedings/lrec2000/pdf/168.pdf

Index

A
Abstractive summarization, 43
Automatic lifelog retrieval systems, 196
Average precision, 3

B
Bigram, 34
Bilingual information retrieval, 25
Biometrics, 192, 193, 199, 200
Bpref, 8

C
Claim, 51
1CLICK, 154
Combinational relevance score, 6
Community question answering, 10
Comparable corpus, 23
Condensed lists, 16
Conference and Labs of the Evaluation Forum (CLEF), 210
Content MathML, 176
Cranfield collections, 3, 208
Cross-language information retrieval, 5, 9, 24
Cross-lingual sentiment analysis, 92

D
Data fusion, 37
Deep NLP, 94
Discounted cumulative gain, 3
Diversified search, 11
D♯-measures, 11

E
Evaluation-as-a-service, 202
Evaluation measures, 4, 33
Evaluation metric, 31, 33, 43, 116–119, 156, 158, 163–165, 167, 180
Expected reciprocal rank, 14
Exploratory search, 75
Extractive summarization, 43
Extrinsic evaluation, 43

F
File forming term, 53
F-measure, 44
Formulae query variable, 178
Formula similarity region, 178
Forum for Information Retrieval Evaluation (FIRE), 210

G
Generalised average precision, 9
Geographical information retrieval, 129
Graded relevance, 3

H
H-measure, 11

T. Sakai et al. (eds.), *Evaluating Information Retrieval and Access Tasks*,
The Information Retrieval Series 43,
https://doi.org/10.1007/978-981-15-5554-1

I
Indicative summaries, 43
Information extraction, 79
Information retrieval for question answering, 9
Informativeness, 44
Informative summaries, 43
Initiative for Evaluation of XML Retrieval (INEX), 210
Interactive lifelog retrieval systems, 196, 198
Interactive track, 77
International patent classification, 53
Intrinsic evaluation, 43
Invalidity, 53
Invalidity search, 52
IR4QA, 130
IUnit, 157

K
κ coefficient, 89

L
Language modeling, 26
Lexical gap, 197, 198, 201
Lifelog activity detection subtask, 199, 200
Lifelog annotation subtask, 199, 200
Lifelog datasets, 191–194, 198, 202
Lifelog event segmentation subtask, 199
Lifelogging, 189–192, 200–202
Lifelog insight subtask, 198, 199, 201
Lifelog search challenge, 191, 198, 201
Lifelog semantic access subtask, 195–197, 200
Lifelog topic, 195, 197, 198, 201, 202

M
Machine translation, 25
MathML, 176
Mean average precision, 33
MediaEval, 210
M-measure, 164
MobileClick, 154
Multi-document summarization, 43
Multilingual information retrieval, 25
Multi-modal summarization, 74
Music Information Retrieval Evaluation eXchange (MIREX), 210

N
Neural machine translation, 37, 104
Normalised discounted cumulative gain, 9
Normalised sliding ratio, 4
Novelty, 53
NTCIR, 3, 210
NTCIREVAL, 9
Nugget, 157

O
Opinion holder, 88
Opinion target, 88
Out-of-vocabulary problem, 32, 35

P
P+, 14
Parallel corpus, 23, 27
Patent application, 51
Patent retrieval, 5
Patent translation, 97–102, 104, 105
Pivot language, 34
11-point average precision, 3
Polarity, 88
Pooling, 179, 208
Pooling method, 31
Precision, 44
Preference judgments, 16
Presentation MathML, 176
Prior art search, 52
Privacy-by-design, 194
Pseudo relevance feedback, 26, 35

Q
Q-measure, 9
Quality, 44
Quantified self, 198
Query-based summarization, 43, 154
Query expansion, 34
Query variable, 178
Question-answering, 43

R
Reading comprehension, 43
Recall, 44
Recall-precision curves, 5
Reciprocal rank, 9
Redaction process, 193, 195
Relaxed relevance, 5
Relevance assessments, 3
Relevance judgement, 31, 32, 43
Reproducibility, 15
Rigid relevance, 5
R-precision, 3

S
Sentiment analysis, 85
SEPIA, 179
Short text conversation, 13
Single-document summarization, 42
S-measure, 163
Spoken document retrieval, 6
Statistical machine translation, 37, 97, 98
Summary generation, 81

T
Technical trend map, 61
Technology survey, 52
Temporal expression, 132
Temporal information access, 129, 130, 133
Temporal information retrieval, 129, 130, 133, 135, 139
Test collection, 3, 30
Text Analysis Conference (TAC), 210
Text mining, 82
Text REtrieval Conference (TREC), 3
Text summarization, 41
Topic Detection and Tracking (TDT), 210
Translation probability, 26, 27
Transliteration, 35
Trec_eval version 9.0, 180
TRECVID, 210
Trend information, 75
Tweet sentiment analysis, 93

U
U-measure, 165

V
Visualization, 80

W
Wearable cameras, 190, 192, 196
Web search, 8, 14
Weighted reciprocal rank, 9
WIPO-alpha, 67
Word sense disambiguation, 26, 35