Overview of NTCIR-14

Makoto P. Kato[1] and Yiqun Liu[2]

[1] University of Tsukuba mpkato@slis.tsukuba.ac.jp
[2] Tsinghua University yiqunliu@tsinghua.edu.cn

Abstract. This is an overview of NTCIR-14, the fourteenth sesquiannual research project for evaluating information access technologies. NTCIR-14 involved various evaluation tasks related to information retrieval, information recommendation, question answering, natural language processing, *etc.* (in total, seven tasks were organized at NTCIR-14). This paper describes an outline of the research project, which includes its organization, schedule, scope and task designs. In addition, we introduce brief statistics of participants in the NTCIR-14 Conference. Readers should refer to individual task overview papers for their detailed descriptions and findings.

Keywords: NTCIR · evaluation campaign · information access technology · benchmark

1 Introduction

Since 1997, NTCIR project has promoted research efforts for enhancing Information Access (IA) technologies such as Information Retrieval and Recommendation, Text Summarization, Information Extraction, and Question Answering techniques. Its general purposes are to: 1. Offer research infrastructure that allows researchers to conduct large-scale evaluation of IA technologies, 2. Form a research community in which findings from comparable experimental results are shared and exchanged, and 3. Develop evaluation methodologies and performance measures of IA technologies. Collaborative works in NTCIR have allowed us to create large-scale test collections that are indispensable for confirming effectiveness of novel IA techniques. Moreover, in the process of the collaboration, it is expected that deep insight into research problems is successfully shared among researchers. The on-going NTCIR-14 aims to be beneficial to all researchers who wish to advance their research efforts.

2 Outline of NTCIR-14

2.1 Organization

The project of NTCIR-14 was directed by General Co-Chairs (GCCs): Charles L. A. Clarke (Facebook, USA), and Noriko Kando (National Institute of Informatics, Japan). Under the supervision of GCCs, Program Committee (PC) reviewed task proposals that were submitted according to a call for proposals, and made acceptance decisions for NTCIR-14. The members of the PC are Ben Carterette (University of Delaware, USA), Hsin-Hsi Chen (National Taiwan University, Taiwan), Tat-Seng Chua (National University of Singapore, Singapore), Nicola Ferro (University of Padova, Italy), Kalervo Järvelin (University of Tampere, Finland), Gareth J. F. Jones (Dublin City University, Ireland), Makoto P. Kato (Co-chair, University of Tsukuba, Japan), Yiqun Liu (Co-chair, Tsinghua University, China), Mandar Mitra (Indian Statistical Institute, India), Douglas W. Oard (University of Maryland, USA), Maarten de Rijke (University of Amsterdam, the Netherlands), Tetsuya Sakai (Waseda University, Japan), Mark Sanderson (RMIT University, Australia), Ian Soboroff (NIST, USA), and Emine Yilmaz (University College London, United Kingdom). After the review by PC, organizers of accepted tasks have promoted research activities of NTCIR-14 under the coordination of the two Program Co-Chairs (PCCs).

2.2 Schedule and Research Activities

Call for task proposals was released in August 2017, and six tasks of NTCIR-14 were decided in November 2017, a month before the NTCIR-13 Conference. Accepted tasks were introduced by the organizers at the NTCIR-13 Conference. Actual NTCIR-14 activities started in January 2018, and a kick-off event was held in March 2018. In addition, call for additional pilot task proposals was released in April 2018, and an additional pilot task, FinNum, was accepted. In total, five core tasks and two pilot tasks (see below) were organized in NTCIR-14. According to the purpose and policy of each task, datasets for experiments (documents, queries and so on) were developed by the task organizers, and distributed to participants (*i.e.* research groups or teams participating in the task) by either the organizers or National Institute of Informatics. New test collections were created based on evaluation of results that were submitted by participants. The research outcome will be reported at the NTCIR-14 Conference to be held in Tokyo, from June 10th to 13th, in 2019.

2.3 Scope and Tasks

The core task explores problems that have been known well in the fields of IA, while the pilot task aims to address novel problems for which there are uncertainties as to how to evaluate them. The five core tasks (Lifelog-3, OpenLiveQ-2, QALab-PoliInfo, STC-3, and WWW-2) and two pilot task (CENTRE and FinNum) can be summarized as follows (illustrated in Figure 1):

1. Heterogeneous information access
2. Dialogue generation and analysis
3. Meta research on information access communities

It is interesting that three tasks are dealing with dialogue data in this round of NTCIR: QALab-PoliInfo focuses on regional assembly minutes, while STC-3

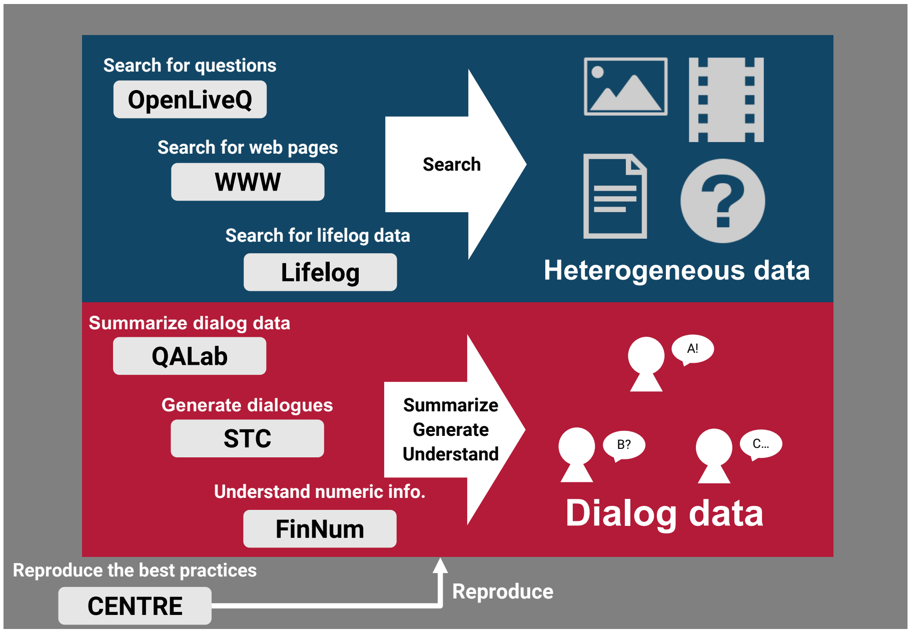

Fig. 1. Overview of the NTCIR-14 tasks.

and FinNum use microblog data. It is also worth mentioning that CENTRE is the first meta-task that operates across the three major information retrieval evaluation venues: CLEF, NTCIR, and TREC.

3 Outline of NTCIR-14 Tasks

3.1 Lifelog-3 (Core Task) [2]

Since NTCIR-12, Lifelog has promoted advances in information access systems for personal lifelog data, which are records of multiple aspects of one's life in digital form. At NTCIR-14, the Lifelog-3 task explored three different lifelog data access related challenges, namely, Lifelog Semantic Access sub-Task (LSAT; a known-item search task for lifelog data), Lifelog Activity Detection sub-Task (LADT; identification of Activities of Daily Living (ADLs) from lifelog data), and Lifelog Insight sub-Task (LIT; an exploratory task for knowledge mining and visualization of lifelog data).

The lifelog data were gathered by two lifeloggers who wore the lifelogging devices and recorded biometric data for several weeks. Wearable camera captured images for 12-14 hours per day (1,250 - 4,500 images per day). In addition, mobile devices gathered locations, physical movements and a history of music listening, and additional wearable sensors recorded health data such as continual heart-rate and continuous blood glucose level.

3.2 OpenLiveQ-2 (Core Task) [3]

OpenLiveQ aims to provide an open live test environment of Yahoo Japan Corporation's community question-answering service (*Yahoo! Chiebukuro*) for question retrieval systems. The main task is an ad-hoc question retrieval task: given a query and a set of questions with their answers, return a ranked list of questions. The organizers released queries sampled from a query log of Yahoo! Chiebukuro search, and clickthrough data with demographics of search users.

Submitted runs were evaluated both offline and online. The offline evaluation uses an evaluation methodology used in ad-hoc retrieval evaluation, while the online evaluation was based on multileaved comparison. In the online evaluation, the task organizers used a multileaving algorithm: submitted ranked lists of questions were combined into a single SERP, presented to real users during the online test period, and evaluated on the basis of clicks observed.

3.3 QALab-PoliInfo (Core Task) [4]

QALab has tackled real-world complex question-answering problems since NTCIR-11, and had focused on solving problems in entrance examinations from NTCIR-11 to NTCIR-13. Motivated by increasing demand of fact-checking due to the fake news problem in the recent years, the NTCIR-14 QALab-PoliInfo task switched their focus to problems related to political information, and addressed three tasks, namely, segmentation, summarization, and classification of Japanese regional assembly minutes.

In the segmentation task, given summaries of a question and an answer to the question in an assembly, participants were expected to identify the original speech corresponding to each summary. In the summarization task, given a speech in an assembly, a system was expected to generated a summary of the speech within a length limit. In the classification task, participants were given a sentence in the minutes with a focused topic, and required to classify the sentence into three classes: support with fact-checkable reasons, against with fact-checkable reasons, and other.

3.4 STC-3 (Core Task) [7, 8]

The Short Text Conversation (STC-3) task took over the efforts in STC-1 and STC-2, which aimed to realize human-machine conversation, and involved three subtasks at NTCIR-14, namely, Chinese Emotional Conversation Generation (CECG) subtask, Dialogue Quality (DQ) subtask, and Nugget Detection (ND) subtask.

The CECG subtask is a task of generating a response to a short message (a microblog post in this task), where the reply should be coherent with a pre-specified emotion class (anger, disgust, happiness, like, sadness, or other). This subtask has been continued from the first round of STC, but required a generated response to be emotional at this round. The DQ and ND subtasks are concerned with automatic evaluation of human-machine customer-helpdesk dialogue systems. Participants were required to develop systems that predict (1) quality scores for each customer-helpdesk dialogue (DQ), and (2) nugget types for each customer-helpdesk dialogue turn (ND), where a nuggets type is either trigger, regular, goal, or *not* (not helpful towards problem solving) of the customer/helpdesk.

3.5 WWW-2 (Core Task) [5]

WWW started at NTCIR-13 to keep addressing basic Web search problems in the IR community. The second round of WWW inherited the task design from the first round: given a query set and a corpus, a system is required to retrieve and rank documents from the corpus for each query. WWW-2 additionally provided a short description for each query for more reliable relevance assessment, and a new dataset called Sogou-QCL, which consists of large-scale weak relevance labels generated by click models.

SogouT-16 (about 1.17 billion Web pages) and ClueWeb12-B13 (about 50 million Web pages) were used as document collections for Chinese and English subtasks, respectively. Based on a statistical approach for topic set size design, 80 queries were prepared for each of the subtasks. Runs were evaluated by standard evaluation metrics for graded relevance, namely, nDCG (normalized discounted cumulative gain), ERR (expected reciprocal rank), and Q-measure.

3.6 CENTRE (Pilot Task) [6]

CENTRE is a meta-task that aims to examine the replicability and reproducibility of results reported in the IR literature, and to establish methods for examining replicability and reproducibility. Each edition of CENTRE was organized at the three major information retrieval evaluation campaigns, *i.e.* CLEF, NTCIR, and TREC, and this task was the NTCIR edition of CENTRE.

The task had three subtasks: T1 (Replicability), T2TREC (Reproducibility), and T2OPEN (Reproducibility). In CENTRE at NTCIR, a pair of runs is said to be replicated/reproduced if the improvement of a run over the other is also found in the same test collection where the improvement was originally reported (replicability), or a different test collection (reproducibility). The T1 subtask is to replicate a run pair from the NTCIR-13 WWW-1 task, while the T2TREC subtask is to reproduce a run pair from the TREC 2013 Web track on the NTCIR-13 WWW-1 test collection. The T2OPEN subtask is to reproduce any existing algorithms on the WWW-1 test collection.

3.7 FinNum (Pilot Task) [1]

FinNum aims at fine-grained numeral understanding in microblogs towards a better understanding of documents containing numeric information such as financial reports or national statistics. Numerals in microblogs were classified into seven categories: Monetary, Percentage, Option, Indicator, Temporal, Quantity, Product/Version. Some of the categories had subcategories. For example, the Monetary class is further classified into money, quote, change, buy price, sell price, forecast, stop loss, and support or resistance. There are in total fourteen subcategories under seven categories. The task of FinNum is to classify numerals in microblogs into the seven categories (Subtask 1) or seventeen categories including Indicator, Quantity, Product/Version, and all the subcategories (Subtask 2).

Microblog data were gathered form StockTwits, a social trading platform for investors. There were 8,868 annotated numerals in the dataset of FinNum, which was released under CC BY-NC-SA 4.0 license. Submitted results were evaluated by micro-averaged and macro-averaged F-scores.

4 Participants

Table 1 shows the numbers of *active* participants (those who submitted results). In this table, the numbers are given for all the tasks from NTCIR-1 to NTCIR-14. Task overview papers (see References) describe evaluation of the results submitted by the participants. At NTCIR-14, 47 research groups have participated in the tasks and the number of participants drastically decreased from NTCIR-13 (*i.e.* 71 groups) and NTCIR-12 (*i.e.* 97). Note that some research groups participated in two tasks, which were counted as different groups.

5 Conclusions

This paper presented the overview of the 14th cycle of NTCIR carried out from January 2018 to June 2019. NTCIR-14 has seven evaluation tasks, which can be categorized into heterogeneous information access, dialogue generation and analysis, and meta research on information access communities. Most parts of the test collections developed by NTCIR-14 evaluation tasks will be released to non-participating research groups in the near future.

6 Acknowledgments

We would like to thank the organizers of all NTCIR-14 tasks for their tremendous amount of efforts devoted to run successful tasks, the task participants for their valuable contributions to the IA research community, and program committee members for their great suggestions to our accepted tasks. Finally, we would like to thank the current and past members of the NTCIR office for their continuous and careful support to our activity.

References

1. Chen, C.C., Huang, H.H., Takamura, H., Chen, H.H.: Overview of the ntcir-14 finnum task: Fine-grained numeral understanding in financial social media data. In: NTCIR-14 Conference (2019)

Table 1. Number of active participants (from NTCIR-1 to NTCIR-14)

Year	**1999**	**2001**	**2002**	**2004**	**2005**	**2007**	**2008**	**2010**	**2011**	**2013**	**2014**	**2016**	**2017**	**2019**
Task/NTCIR round	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Total number	37	39	61	74	79	81	80	66	102	108	93	97	71	47
Automatic Term Recognition and Role Analysis (TMREC) (1)	9													
Ad hoc/Crosslingual IR (1) → Chinese/English/Japanese IR (2) → CLIR (3-6)	28	30	20	26	25	22								
Text Summarization Challenge (TSC) (2-4)		9	8	9										
Web Retrieval (WEB) (3-5)			7	11	7									
Question Answering Challenge (QAC) (3-6)			16	18	7	8								
Patent Retrieval [and Classification] (PATENT) (3-6)			10	10	13	12								
Multimodal Summarization for Trend Information (MUST) (5-7)					13	15	13							
Crosslingual Question Answering (CLQA) (5, 6) → Advanced Crosslingual Information Access (ACLIA) (7, 8)					14	12	19	14						
Opinion (6) → Multilingual Opinion Analysis (MOAT) (7, 8)						12	21	16						
Patent Mining (PAT-MN) (7, 8)							12	11						
Community Question Answering (CQA) (8)								4						
Geotemporal IR (GeoTime) (8, 9)								13	12					
Interactive Visual Exploration (Vis-Ex) (9)									4					
Patent Translation (PAT-MT)(7, 8) → Patent Machine Translation (PatentMT)(9, 10)							15	8	21	21				
Crosslingual Link Discovery (Crosslink) (9, 10)									11	10				
INTENT(9, 10) → Search Intent and Task Mining (IMine) (11, 12)									16	11	12	9		
One Click Access (1CLICK)(9, 10) → Mobile Information Access (MobileClick) (11, 12)									4	8	4	11		
Recognizing Inference in Text (RITE)(9,10) → Recognizing Inference in Text and Validation (RITE-VAL)(11)									24	28	23			
IR for Spoken Documents (SpokenDoc) (9, 10) → Spoken Query and Spoken Document Retrieval (SpokenQuery&Doc) (11, 12)									10	12	11	7		
Mathematical Information Access (Math) (10, 11) → MathIR (12)										6	8	6		
Medical Natural Language Processing (MedNLP) (10, 11) → MedNLPDoc (12) → MedWeb (13)										12	12	8	9	
QA Lab for Entrance Exam (QALab) (11, 12, 13) → QA Lab for Political Information (QALab-PoliInfo) (14)											11	12	11	13
Temporal Information Access (Temporalia) (11, 12)											8	14		
Cooking Recipe Search (RecipeSearch) (11)											4			
Personal Lifelog Organisation & Retrieval (Lifelog) (12, 13, 14)												8	4	6
Short Text Conversation (STC) (12, 13, 14)												22	27	13
Open Live Test for Question Retrieval (OpenLiveQ) (13, 14)													7	4
Actionable Knowledge Graph (AKG) (13)													3	
Emotion Cause Analysis (ECA) (13)													3	
Neurally Augmented Image Labelling Strategies (NAILS) (13)													2	
We Want Web (WWW) (13, 14)													5	4
Fine-Grained Numeral Understanding in Financial Tweet (FinNum) (14)														6
CLEF/NTCIR/TREC REproducibility (CENTRE) (14)														1

2. Gurrin, C., Joho, H., Hopfgartner, F., Zhou, L., Ninh, V.T., Le, T.K., Albatal, R., Dang Nguyen, D.T., Healy, G.: Overview of the ntcir-14 lifelog-3 task. In: NTCIR-14 Conference (2019)
3. Kato, M.P., Nishida, A., Manabe, T., Fujita, S., Yamamoto, T.: Overview of the ntcir-14 openliveq-2 task. In: NTCIR-14 Conference (2019)
4. Kimura, Y., Shibuki, H., Ototake, H., Uchida, Y., Takamaru, K., Sakamoto, K., Ishioroshi, M., Mitamura, T., Kando, N., Mori, T., Yuasa, H., Sekine, S., Inui, K.: Overview of the ntcir-14 qa lab-poliinfo task. In: NTCIR-14 Conference (2019)
5. Mao, J., Sakai, T., Luo, C., Xiao, P., Liu, Y., Dou, Z.: Overview of the ntcir-14 we want web task. In: NTCIR-14 Conference (2019)
6. Sakai, T., Ferro, N., Soboroff, I., Zeng, Z., Xiao, P., Maistro, M.: Overview of the ntcir-14 centre task. In: NTCIR-14 Conference (2019)
7. Zeng, Z., Kato, S., Sakai, T.: Overview of the ntcir-14 short text conversation task: Dialogue quality and nugget detection subtasks. In: NTCIR-14 Conference (2019)

8. Zhang, Y., Huang, M.: Overview of the ntcir-14 short text generation subtask: Emotion generation challenge. In: NTCIR-14 Conference (2019)