Report on NTCIR-13: The Thirteenth Round of NII Testbeds and Community for Information Access Research

Yiqun LIU
Tsinghua University, China
yiqunliu@tsinghua.edu.cn

Makoto P. Kato
Kyoto University, Japan
mpkato@acm.org

Charles L.A. Clarke
Facebook, USA
claclark@gmail.com

Noriko Kando
National Institute of Informatics, Japan
kando@nii.ac.jp

Tetsuya Sakai
Waseda University, Japan
tetsuya@waseda.jp

Abstract

This is a report on the NTCIR-13 conference held in December 2017, in Tokyo, Japan. NTCIR is a series of parallel and collective evaluation efforts designed to enhance research on diverse information access technologies, including, but not limited to, cross-language and multimedia information access, question-answering, text mining and summarization, with an emphasis on East Asian languages such as Chinese, Korean, and Japanese, as well as English. 105 different research groups from 20 countries/regions participated in one or more of the nine different tasks in NTCIR-13, to compete and collaborate on a common ground and thereby advance the state of the art. This report introduces the highlights of the conference, describes the scope and task designs of nine tasks organized at NTCIR-13, and provides a brief introduction to NTCIR-14, which started from January 2018 and will be closed in June 2019, which will be the 20th anniversary since the first NTCIR Conference in the summer of 1999.

1 Introduction

Since 1997, the NTCIR project has promoted research efforts for enhancing information access (IA) technologies such as information retrieval (IR), text summarization, information extraction, and question answering techniques. Its general purposes are: (1) to offer research infrastructure that allows researchers to conduct a large-scale evaluation of IA technologies, (2) to form a research community in which findings from comparable experimental results are shared and exchanged, and (3) to develop evaluation methodologies and performance

measures of IA technologies. Collaborative works in the NTCIR allow 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.

Each NTCIR conference concludes the researchers? efforts over the course of 18 months or so, in the form of official results and future work items. The thirteenth round of NTCIR, NTCIR-13, started in June 2016 and was concluded in December 2017, with the NTCIR-13 conference held in Tokyo, Japan[1]. The conference began with a satellite workshop on EValuating Information Access (EVIA 2017)[2]. It is a one-day workshop for information access researchers to present their work on evaluation methods, measures, test collections, experimental designs, etc.

The main conference was initiated by an overview of NTCIR-13, and followed by a keynote given by Omar Alonso, about crowdsourcing in data labeling researches. After that, each task was then introduced by task organizers and further discussed at their own session, where task participants had oral presentations on their approaches. Poster sessions were arranged at lunch on the 2nd and 3rd days of the conference, which provided a place for task participants to exchange information and ideas on these tasks. This is the 20th anniversary since the NTCIR project started, so a special session was also arranged to celebrate the event. Fredric C. Gey and Douglas W. Oard were invited to give two talks about the history and future of NTCIR. The conference was wrapped up with invited talks about news from TREC and CLEF, and an introduction to NTCIR-14 tasks. The details of the conference are reported in Section 2.

There were nine tasks organized in NTCIR-13: five *core* tasks (Lifelog-2, MedWeb, OpenLiveQ, QALab-3 and STC-2) and four *pilot* tasks (AKG, ECA, NAILS and WWW). The NTCIR-12 tasks cover a broad range of IA topics, and can be summarized as follows [9]: (1) Answering complex questions and queries, (2) Mining knowledge from a large amount of human generated data, and (3) The application of (2) in (1). A brief introduction to these tasks are provided in Section 3.

Six tasks have already been accepted at NTCIR-14 [3]. NTCIR-14 keeps its diversity and offers a wide range of IA tasks: Lifelog-3, OpenLiveQ-2, QALab-4, STC-3, WWW-2, and CENTRE. These tasks are also introduced in Section 4.

For more details, please refer to the online proceedings of NTCIR-13 and EVIA 2017:

- NTCIR-13 proceedings [6]:
  `http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings13/NTCIR/toc_ntcir.html`
- EVIA 2017 proceedings [2]:
  `http://ceur-ws.org/Vol-2008/`

2 NTCIR-13 Conference

The NTCIR-13 conference was held from December 5 to 8 in National Institute of Informatics (Tokyo, Japan), and attracted 201 participants from 20 countries/regions.

EVIA 2017 was organized by Nicola Ferro and Ian Soboroff, and held on the first day of the NTCIR-13 conference. The workshop consisted of a keynote presented by Douglas W.

[1] `http://research.nii.ac.jp/ntcir/ntcir-13/`
[2] `http://research.nii.ac.jp/ntcir/evia2017/`
[3] `http://research.nii.ac.jp/ntcir/ntcir-14/`

Oard and seven paper presentations. Douglas W. Oard from University of Maryland talked about their recent efforts in Cross Language Information Retrieval (CLIR) under a research program called MAchine Translation for English Retrieval of Information in Any Language (MATERIAL). Seven papers presented in EVIA 2017 were concerned with evaluation in diverse kinds of scenarios, e.g. evaluation methodology for Customer-Helpdesk and multi-turn dialogues, meta evaluation of evaluation metrics, and new evaluation metrics.

The main conference of NTCIR-13 started with the overview presentation from general chairs and program committee co-chairs. Omar Alonso from Microsoft in Silicon Valley gave a keynote presentation on "The practice of crowdsourcing: things to know about using humans and machines for labeling". From rich experiences in collecting data labels from crowdsourcing efforts, he outlined a number of methods that work in practice and describe a number of trade-offs when designing and implementing computation systems that use humans and machines.

After the keynote presentation, the overview of each task was presented by task organizers. There were nine tasks organized in NTCIR-13, and their data, task design, and experimental results were presented for familiarizing each task for all the participants in the NTCIR-13 conference.

Task organizers then hosted their own task sessions in parallel. Some participating groups were selected by task organizers and asked to have oral presentations in those sessions. Some were selected since they achieved the best performance in the task, while others were selected since they employed significantly different approaches from the other teams. Participants mainly presented their approaches and results, with their own error analysis, which triggered questions from the other participants who tackled the same problem and resulted in deep discussion that may not be seen in ordinary conferences. All the participants had a chance to present their work at poster sessions during lunch on the second and third day of the conference. Moreover, the task organizers of OpenLiveQ, AKG, STC and QALab-3 hosted *break-out sessions*, where organizers and participants discussed the current task and planned for the next round.

Since this is the twentieth year since NTCIR project started, the conference also featured a special "NTCIR 20th Anniversary Session" session. Two invited talks were given by Fredric C. Gey and Douglas W. Oard on the topic of "NTCIR from the beginning" and "NTCIR in the world", respectively. The talk given by Fredric introduced his research journey from 1997 on the topic of cross-language search and multilingual information access. In the second talk, Prof. Oard looked back over the history of NTCIR to highlight some of the impactful and innovative evaluation tasks. He also explored the impact of these tasks from national, regional, or global perspectives and provide valuable suggestions about the future development of NTCIR.

At the last day of the NTCIR-13 conferences, two invited speakers presented their projects. Ian Soboroff from the National Institute of Standards and Technology talked about the new tracks in TREC 2018. He also highlight some new and interesting evaluation activities hosted by NIST. Nicola Ferro from University of Padua introduced the activities conducted in the Evaluation Labs and the discussions held during the Conference of CLEF 2018. At the end of the conference, NTCIR-14 tasks were presented by task organizers.

3 NTCIR-13 Tasks

NTCIR-13 included five *core* tasks (Lifelog-2, MedWeb, OpenLiveQ, QALab-3 and STC-2) and four *pilot* tasks (AKG, ECA, NAILS and WWW). The former targeted relatively well-known IA problems, while the latter targeted novel problems. There were 115 teams registered in NTCIR-13, of which 71 teams submitted their runs in time. Each NTCIR-13 task is briefly explained below (please refer to the NTCIR-13 overview paper [9] and overview paper of each task [4, 12, 7, 11, 10, 1, 3, 5, 8] for details).

Personal Lifelog Organisation & Retrieval Task (Lifelog-2)
Personal lifelogging is the process of capturing multiple aspects of one's life in digital form. This is the second round of Lifelog task in NTCIR and it has become a core task. Compared with the task in NTCIR-12, the task organizers develop a new (more semantically rich) test collection collected by real lifeloggers and a baseline search system. They also propose more subtasks including Lifelog Semantic Access (LSAT), Lifelog Event Segmentation (LES), Lifelog Insight (LIT) and Lifelog Annotation (LAT), among which LSAT and LIT originated from Lifelog-1. As for the task designing, they use a two-phase manner in which LTA is located in Phase I and the other three subtasks are in Phase II. Some subtasks, especially the LIT task, is organized as a forum for researchers to present their ideas on how to make good use of lifelog data to improve human life.

Medical Natural Language Processing for Web Document (MedWeb)
MedWeb is a newly proposed core task in NTCIR-13. Compared with previous efforts in dealing with medical related documents in MedNLP tasks (in NTCIR-10, 11 and 12), the new task mainly focuses on dealing with Web-based social media data. In this year, the organizers provide twitter contents and ask the participants to assign labels on whether or not a particular post contains symptoms. The assignment process can be regarded as a multi-label classification problem because a single post may contain descriptions of multiple symptoms. The original twitter texts are in Japanese and then translated into both English and Chinese to make it a cross-language task. Based on the corpus, the organizers evaluate the performance of systems in a three-step manner. They firstly distribute training corpus so that participants can develop their systems for a time period of around three months. After that, they release the test set and require result submission within two weeks. Finally, the submitted runs are annotated and final results are released. Different levels of matching are considered in the evaluation metric designing, which contains exact match accuracy, Hamming loss and precision/recall/F1-measure scores in both micro and macro levels.

Open Live Test for Question Retrieval (OpenLiveQ)
OpenLiveQ is a newly proposed core task in NTCIR-13. This task 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 simply defined as follows: 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 multi-leaved comparison. In the online evaluation, 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.

QA Lab for Entrance Exam (QALab-3)

Following the continuous efforts in NTCIR-11 and 12, QALab-3 also focus on developing question-answering systems that can solve university entrance exam questions. This year, the question set is composed of "world history" questions selected from both The National Center Test for University Admissions (multiple-choice questions) and secondary exams of the University of Tokyo (term and essay questions). The task organizers provide heterogeneous resources including high school textbooks, Wikipedia and World History Ontology. Participants can also use any other types of resources to construct their QA systems. Besides the traditional end-to-end task which aims at providing correct answers, three other subtasks are also proposed for the essay generation scenario, namely Extraction, Summarization and Evaluation-method. As for evaluation metrics, accuracy is adopted for multiple-choice questions. For term based questions, the accuracy based on exact matching (synonym is taken into consideration) is adopted. While for essay generation, the end-to-end subtask was assessed by human experts, using ROUGE method, Pyramid method and quality questions.

Short Text Conversation (STC-2)

Short Text Conversation (STC-2) task attempts to develop systems replying a short answer to the user in response to her/his short question. In STC-1, the pilot task was considered as an IR problem by maintaining a large repository of post-comment pairs, and then reusing these existing comments to respond to new posts. Besides this retrieval-based subtask, this round of STC also propose a new generation-based subtask which aims to generate "new" comments to answer questions. STC can be regarded as a simplified scenario of natural language conversational system in which multi-round conversation is not allowed and context information is not considered. This makes it focus on dig deeply into both IR and NLP techniques to find possible solutions. This year, the task also designed a transparent platform to compare the retreival-based and generation-based methods by comprehensive evaluations.

Actionable Knowledge Graph (AKG)

AKG aims to help users (especially search engine users) to gain actionable suggestions after submitting a query containing entities. The definition of "Actionalble Knowledge Graph" is "a specialized version of KG that contains data on the range of possible actions and their related information in relation to particular entity types and their instances." This task is inspired by the increasing number of knowledge graph results on search engine result pages and may contribute to better search user experiences. As the first round, the AKG task is composed of two subtasks. The first one is named Action Mining (AM) subtask which requires both IE and IR techniques to return relevant actions for input entities. The second subtask is Actionable Knowledge Graph Generation (AKGG) in which participating systems are required to assign properties for the combination of entity and one of its actions. Traditional IR based methods such as nNDCG@N and nERR@N are adopted to evaluate system performance.

Emotion Cause Analysis (ECA)

ECA aims to locate the stimuli, or the cause of emotions besides just identifying the emo-

tions. Considering that there is an increasing demand in finding the emotion causes both from researchers (to better understand users) and from businesses (to understand why their products/services are liked/disliked by customers), the pilot task may advance existing techniques and lead to novel interesting research directions. In this round of the task, the task organizers invest much effort in the construction of a first-of-its-kind corpus which contains English and Chinese news articles, emotions annotated based on the context and direct causes that stimulate the emotions. They designed two subtasks including a coarse-grained subtask which requires detecting causes at the clause level and a fine-grained one which requires detection at the phrase level. Precision, recall and F-measure are adopted as the main evaluation metric in this task.

Neurally Augmented Image Labelling Strategies (NAILS)
NAILS is described by the task organizers as a "data challenge". The participants are expected to make predictions on whether one image is relevant or not based on human volunteer's neural responses to high-speed image search tasks. The neural response data collected by task organizers is the P300 oddball signal which is a well-known signal in Electroencephalography (EEG) studies. Basically, the participants should build their own machine learning models based on a number of training samples and then the organizers will test the proposed models' performances using withheld ground truth data. Besides the main evaluation metric of prediction accuracy, the organizers also encourage participating teams to contribute better solutions in terms of speed, model complexity, neurophysiological interpretability and/or cross-task applicability.

We Want Web (WWW)
With straight ad hoc web search tasks disappearing from NTCIR and TREC, the organizers believe that it is necessary to maintain a search task for the whole IR community. The WWW task can be regarded as a recent forum in the continuous efforts to tackle basic Web search problems. As the first round of the task, WWW contains both a Chinese subtask and an English subtask. Both subtasks have similar settings, share overlapped query topics but use different corpuses. The Chinese subtask chooses a new version of SogouT while English subtask uses the traditional ClueWeb-12 dataset. The organizers propose to continue the task for at least three rounds to monitor the progress of IR algorithms in a relatively long time period.

Lastly, languages of each task are shown in Table 1 for highlighting the linguistic diversity of the NTCIR-13 tasks. English was used in many of the tasks, and Japanese and Chinese were both used in four tasks. It can be seen that NTCIR-13 provided opportunities to address tasks specific to Asian languages such as Japanese and Chinese, as well as English tasks that could be tackled by a wide range of researchers.

4 NTCIR-14 Tasks

Proposals for NTCIR-14 tasks were submitted in September 2017, and reviewed by the NTCIR-14 program committee members. Six accepted tasks were announced at the NTCIR-13 conference, and are looking for participants as of April 2018. NTCIR-14 includes five core tasks (Lifelog-3, OpenLiveQ-2, QALab-PoliInfo, STC-2, and WWW-2), and a pilot task

Table 1: Languages used in each NTCIR-13 task.

Task	English	Japanese	Chinese
Lifelog-2	✓		
MedWeb	✓	✓	✓
OpenLiveQ		✓	
QALab-3		✓	
STC-2	✓		✓
AKG		✓	
ECA	✓		✓
NAILS	✓		
WWW	✓		✓
	6	4	4

(CENTRE). The objectives of NTCIR-14 can be summarized as follows: (1) information retrieval for a wide variety of contents such as documents, lifelog data, and questions, (2) analysis and generation of conversation, and (3) meta research focusing on replicability and reproducibility.

Below, we briefly introduce the NTCIR-14 tasks based on the current task description (see `http://research.nii.ac.jp/ntcir/ntcir-14/` for details).

Lifelog Search Task (Lifelog-3)
Lifelog-3 provides lifelog data that consist of images taken by wearable cameras, human biometrics, information access logs, and human activities, and offers three subtasks, namely, Lifelog Semantic Access Task (LSAT), Lifelog Insights Task (LIT), and Lifelog Activity Detection Task (LADT). LSAT is a known-item search task that was also run at NTCIR-12 and -13. LIT is a free-style task where participants are expected to develop a system that helps users gain insights into the lifelogger's life. LADT is a new task at NTCIR-14, for automatically annotating the lifelog data with human activities.
Website: `http://ntcir-lifelog.computing.dcu.ie/`

Open Live Test for Question Retrieval (OpenLiveQ-2)
OpenLiveQ provides an open live test environment of Yahoo Japan Corporations community QA service for question retrieval systems, and aims to provide an opportunity for a more realistic evaluation to address problems specific to a production environment. OpenLiveQ-2 is the second round of the OpenLiveQ task at NTCIR-13.
Website: `http://www.openliveq.net/`

Question Answering Lab for Political Information (QALab-PoliInfo)
QALab-PoliInfo task aims at complex real-world question answering (QA) technologies, to extract structured data on the opinions of assemblymen, and the reasons and conditions for such opinions, from Japanese regional assembly minutes. Three subtasks related to the assembly minutes are proposed in this task: Segmentation subtask: given a pair of an utterance in the assembly minutes and its quote, extract a segment in the minutes to

precisely understand the quote. Summarization subtask: given an utterance in the assembly minutes, generate a summary that clearly conveys the intent of the speaker of the utterance. Classification subtask: given an utterance in the assembly minutes, classify the utterance into four classes: merit, demerit, both, and N/A.
Website: `https://poliinfo.github.io/`

Short Text Conversation (STC-3)
STC-3 offers three subtasks that focus on human-machine conversation: Chinese Emotional Conversation Generation (CECG), Dialogue Quality, and Nugget Detection subtasks. CECG: given a Chinese Weibo post and an emotion category (e.g. anger, disgust, happiness), return an appropriate response that matches the category. Dialogue Quality: given a helpdesk-customer dialogue, estimate the distribution of overall subjective scores (e.g. customer satisfaction and task accomplishment) as rated by multiple assessors. Nugget Detection: given a helpdesk-customer dialogue, estimate, for each utterance, the distribution of multiple assessors' labels over pre-defined classes (e.g. Not-A-Nugget, Regular Nugget, Trigger Nugget, Goal Nugget).
Website: `http://sakailab.com/ntcir14stc3/`

We Want Web-2 (WWW-2)
WWW runs an ad hoc Web task for at least three rounds at NTCIR for quantifying the progress of Web search system development. Systems are evaluated not only with traditional measures but also with more advanced measures that better reflect user experiences. WWW-2, the second round of WWW, will offer a new resource for Chinese subtask, Sogou-QCL, which contains weak relevance labels for millions of query-doc pairs.
Website: `http://www.thuir.cn/ntcirwww2`

CLEF/NTCIR/TREC REproducibility (CENTRE)
CENTRE is a unique task that started from NTCIR-14, and is a collaboration across CLEF, NTCIR, and TREC. At NTCIR, participants are expected to replicate/reproduce best practices in the past rounds of these evaluation campaigns. Replicability subtask aims to develop the same system as that reported in the past NTCIR and to examine the performance with the NTCIR data, while Reproducibility subtask aims to develop the same system as that reported in the past CLEF or TREC and to examine the performance with the NTCIR data.
Website: `http://sakailab.com/ntcir14centre/`

5 Summary

We reported the NTCIR-13 conference held in December 2017, which attracted attracted 201 participants from 20 countries/regions. We also briefly introduced NTCIR-13 tasks (Lifelog-2, MedWeb, OpenLiveQ, QALab-3, STC-2, AKG, ECA, NAILS, and WWW), and on-going NTCIR-14 tasks (Lifelog-3, OpenLiveQ-2, QALab-4, STC-3, WWW-2, and CENTRE). We hope that readers are interested in NTCIR and participate in the NTCIR-14 tasks.

6 Acknowledgements

Contributions of the NTCIR-13 program committee and organizing committee members, task organizers, data providers, student volunteers, and NTCIR office staffs were vital for the success of NTCIR-13 and the NTCIR-13 conference. We would like to express our gratitude to the following organizations for sponsoring the NTCIR-13 conference: Yahoo! JAPAN Corporation, kizasi Company, Inc., Japan Patent Information Organization and IR-Advanced Linguistic Technologies Inc. We also would like to thank ACM SIGIR for providing the SIGIR Friends fund which is used in the travel support for international participating students and younger researchers..

References

[1] R. Blanco, H. Joho, A. Jatowt, H.-T. Yu, and S. Yamamoto. Overview of ntcir-13 actionable knowledge graph (akg) task. In *Proceedings of the NTCIR-13 Conference*, 2017.

[2] N. Ferro and I. Soboroff, editors. *Proceedings of the Eighth International Workshop on Evaluating Information Access (EVIA 2016), a Satellite Workshop of the NTCIR-12 Conference*. National Institute of Informatics, 2017. `http://ceur-ws.org/Vol-2008/`.

[3] Q. Gao, H. Jiannan, X. Ruifeng, G. Lin, Y. He, K.-F. Wong, and Q. Lu. Overview of NTCIR-13 ECA task. In *Proceedings of the NTCIR-13 Conference.*, 2017.

[4] C. Gurrin, H. Joho, F. Hopfgartner, L. Zhou, D.-T. Dang-Nguyen, R. Gupta, and R. Albatal. Overview of NTCIR-13 Lifelog-2 task. In *Proceedings of the NTCIR-13 Conference*, 2017.

[5] G. Healy, T. Ward, C. Gurrin, and A. Smeaton. Overview of the NTCIR-13 NAILS task. In *Proceedings of the NTCIR-13 Conference*, 2017.

[6] N. Kando, C. L. A. Clarke, T. Sakai, M. P. Kato, and Y. Liu, editors. *Proceedings of the 13th NTCIR Conference on Evaluation of Information Access Technologies*. National Institute of Informatics, 2017. `http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings13/NTCIR/toc_ntcir.html`.

[7] M. P. Kato, T. Yamamoto, T. Manabe, A. Nishida, and S. Fujita. Overview of the NTCIR-13 OpenLiveQ task. In *Proceedings of the NTCIR-13 Conference*, 2017.

[8] C. Luo, T. Sakai, Y. Liu, Z. Dou, C. Xiong, and J. Xu. Overview of the NTCIR-13 We Want Web task. In *Proceedings of the NTCIR-13 Conference*, 2017.

[9] M. P. Kato and Y. Liu. Overview of NTCIR-13. In *Proceedings of the NTCIR-13 Conference*, 2017.

[10] L. Shang, T. Sakai, H. Li, R. Higashinaka, Y. Miyao, Y. Arase, and M. Nomoto. Overview of the NTCIR-13 Short Text Conversation task. In *Proceedings of the NTCIR-13 Conference*, 2017.

[11] H. Shibuki, K. Sakamoto, M. Ishioroshi, Y. Kano, T. Mitamura, T. Mori, and N. Kando. Overview of the NTCIR-13 QALab-3 task. In *Proceedings of the NTCIR-13 Conference*, 2017.

[12] S. Wakamiya, M. Morita, Y. Kano, T. Ohkuma, and E. Aramaki. Overview of the ntcir-13: MedWeb task. In *Proceedings of the NTCIR-13 Conference*, 2017.