Report on the 15th Round of NII Testbeds and Community for Information Access Research (NTCIR-15)

Makoto P. Kato
University of Tsukuba, Japan
mpkato@acm.org

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

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

Charles L.A. Clarke
University of Waterloo, Canada
claclark@gmail.com

Abstract

This is a report on the NTCIR-15 conference held online in December 2020. NTCIR is a sesquiannual research project designed to evaluate various information access technologies, including information retrieval, information recommendation, question answering, natural language processing, etc. 55 active research groups from 22 countries\regions have participated in one or more of the seven tasks in NTCIR-15. This report introduces the highlights of the conference, describes the scope and task designs of the seven tasks organized at NTCIR-15.

Date: 8–11 December, 2020.

Website: http://research.nii.ac.jp/ntcir/ntcir-15/.

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' continuous efforts over the course of 18 months. The fifteenth round of NTCIR, NTCIR-15, started in June 2019 and was concluded in

December 2020, with the NTCIR-15 conference held online[1]. Because of the COVID-19, NTCIR-15 was held virtually for the first time.

The main conference was initiated by an overview of NTCIR-15 and followed by two keynote speeches. The first keynote was given by Ben Carterette from Spotify, entitled *From Offline to Online Experimentation: Considerations from Experiences at Spotify*. The second keynote was given by Xiao-li Meng from Harvard University, entitled *Reproducibility, Replicability and Reliability: Reflections of a Statistician and a Data Science Editor*. Afterwards, each of the seven tasks was introduced by task organizers and further discussed at their own session, where task participants presented their approaches orally. Poster sessions were arranged during the conference at the 3rd and 4th day of the conference, where tasks participants could exchange information and ideas on these tasks. The conference was wrapped up with invited talks about TREC in 2020 from Ellen Voorhees and CLEF initiative during COVID-19 from Nicola Ferro, an introduction to a multimedia benchmarking initiative from Gareth Jones, and the experience on collaborative benchmarking for life-log from Cathal Gurrin. The details of the conference are reported in Section 2.

There were seven tasks organized in NTCIR-15: five core tasks (DialEval-1, FinNum-2, QA Lab-PoliInfo-2, SHINRA2020-ML and WWW-3) and two pilot tasks (DataSearch and MART). The NTCIR-15 tasks cover a broad range of IA topics, and can be summarized as follows [Liu et al., 2020]: (1) Fine-grained text understanding, (2) Modern retrieval tasks, (3) Dialogue analysis. A brief introduction to these tasks are provided in Section 3.

For more details, please refer to the online proceedings of NTCIR-15 [Clarke et al., 2020]:

- NTCIR-15 proceedings:
  http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings15/index.html

.

2 NTCIR-15 Conference

The NTCIR-15 conference was held from December 8 to 11 online, and attracted 277 participants from 22 countries\regions. The main conference started with a keynote presented by Ben Carterette, the ACM SIGIR chair from Spotify, and followed by the overview presentation from general chairs and program committee co-chairs. Lacking of sufficient knowledge on offline experiments, typically the Cranfield paradigm, to predict the online outcomes with real users in real-world conditions, a common framework was discussed for connecting both offline experiments and online A/B testing. He emphasized that understanding translation of offline experiments into online experiences will be the key as approaches to be adopted in real world continuously.

Each overview of seven tasks was then presented by task organizers, including task design, data and experimental results, with the purpose of familiarizing each task for all the participants in the NTCIR-15 conference.

Task organizers then hosted their own task session in a paralleling way in the next two days. Some participating groups were selected by task organizers and asked to have oral presentation in the corresponding sessions. Some were selected because they achieved the best performance in

[1] http://research.nii.ac.jp/ntcir/ntcir-15

the task while others were selected since they employed substantially different approaches from the other teams. Participants mainly presented their approaches and results with their own error analysis. These presentations largely triggered questions from the other participants, who tackled the same problem, and yielded deep discussions that may not be seen in ordinary conferences. All the participants had a chance to present their work at poster sessions set up at the 3rd and 4th day of conference.

At the 3rd day of the NTCIR-15 conference, the second keynote was presented by Xiao-li Meng, who is the Whipple V.N. Jones professor of Statistics at Harvard university and the Founding Editor-in-Chief of Harvard Data Science Review. He first differentiated the three key terms in scientific community, that is reproducibility, replicability and reliability. In the era of information exposure, reliability is guaranteed by the quality instead of the quantity of big data. Thereafter, he, as a statistician and an editor, reflected on the issue of quantifying the quality of big data and demonstrated that the big data is rather than big when taking the quality into consideration.

At the last day of the NTCIR-15 conference, four invited speakers presented their projects and shared their experiences. Ellen Voorhees from the National Institute of Standards and Technology reported their new track (TREC-COVID) in TREC 2020. Impacted by the pandemic, both relevance assessment and the conference were required to be remote and the topics mostly or partially were related to COVID-19. She further highlighted the implementation and outcome of the eight tracks. Nicola Ferro from University of Padua talked about the achievements of European initiative and the COVID-10 MLIA @ Eval initiative in CLEF 2020. Gareth Jones from Dublin City University introduced the objectives of current activities and summarized them within MediaEval 2020. Cathal Gurrin from Dublin City University provided an overview of the outputs of the NTCIR-Lifelog task and highlighted the wider impact of NTCIR-lifelog.

3 NTCIR-15 Tasks

NTCIR-15 included five core tasks (DialEval-1, FinNum-2, QA Lab-PoliInfo-2, SHINRA2020-ML and WWWW-3) and two pilot tasks (DataSearch and MART). The former targeted the well-known IA problems while the latter targeted novel problems. There were 277 participants registered in NTCIR-15, of which 55 teams submitted their runs. Each NTCIR-15 task is briefly explained below (please refer to the NTCIR-15 overview paper [Liu et al., 2020] and overview paper of each task [Chen et al., 2020; Healy et al., 2020; Zeng et al., 2020; Sakai et al., 2020; Kato et al., 2020; Sekine et al., 2020; Kimura et al., 2020] for details).

3.1 Dialogue Evaluation Task (DialEval-1)

Automated help desk aims to answer customers' inquiries with intelligent agents instead of human custom services. To help improve this particular kind of dialogue systems, the DialEval task is designed to evaluate their performance both accurately and efficiently. DialEval-1 is the successor of Dialogue Quality (DQ) and Nugget Detection (ND) subtasks of Short Text Conversation (STC-3) task at NTCIR-14 in 2019. The dataset comes from the Chinese microblog platform Weibo and part of the collected posts were translated to English by the organizers as The English dataset. The training and development sets for Chinese data reuses DCH-1 dataset (3700 for training, 390 for development) which is from STC-3 task. Meanwhile, the task organizers developed a new test set with 300 dialogues. The English datasets contains 2,251 dialogues for training, 390 for development, and 300 for test.

3.2 Numeral Attachment in Financial Tweets (FinNum-2)

Existing works show that the processing methodologies for numerals may be rather different from those designed for ordinary textual information. The FinNum task aims to understand numeral information, especially for the numerals in financial social media data. This is the second year of the FinNum task and the focus is on the numeral attachment problem, which aims to connect given numeral with its corresponding cashtag (special tag for a particular stock) in a multi-numeral and/or multi-cashtag scenario. To evaluate performance on this task, the task provides a newly constructed dataset named NumAttach. Each instance in the dataset was annotated by two experts in the financial domain and only the instances with the same annotation results are included in the NumAttach dataset. Among the 10,340 instances in NumAttach, 7,187 were used for training, 1,044 were for development and the rest (around 20%) were for testing. The task deals with only English tweets and the task organizers plan to extend the task to deal with blog articles and formal financial documents in the future years.

3.3 Question Answering Lab for Political Information (QA Lab-PoliInfo-2)

QA Lab has been trying to tackle 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, since NTCIR-14, it has switched its focus on political information processing. Especially, in NTCIR-15, the task aims to extract summaries of the opinions of assembly members and the reasons and conditions for such opinions, from Japanese regional assembly minutes. This year the task reuses the Japanese Regional Assembly Minutes Corpus as in last round that collects minutes of plenary assemblies in 47 prefectures of Japan from April 2011 to March 2015. Four subtasks are proposed by the task organizers: stance classification, dialog summarization, entity linking and topic detection.

3.4 SHINRA 2020 Multi-lingual (SHINRA2020-ML)

Wikipedia contains a large amount of valuable information and is regarded as a great source of knowledge. However, the documents in Wikipedia is not well-structured and many valuable information cannot be directly adopted in knowledge-driven tasks. The final goal of SHINRA project is to structure all information in Wikipedia, while the SHINRA2020-ML task, as the first step, focuses on classification of Wikipedia pages in 30 languages into a well-defined category named Extended Named Entity (ENE). The task is defined as a multi-label classification problem and micro averaged F1 measure is adopted to evaluate system performance. Besides evaluation purposes, the organizers also aim to create the structured Wikipedia knowledge base using the outputs of the participated systems.

3.5 We Want Web with CENTRE (WWW-3)

The WWW task started at NTCIR-13 to keep addressing basic Web search problems in the IR community after the termination of the Web track at TREC 2014. The task is basically an ad-hoc retrieval task which deals with Web corpus in both English and Chinese. The task organizers want to quantify technical improvements in Web search performance over a several-year period so the task is supposed to last for at least three rounds. The third round of WWW Chinese subtask inherits the task design from the past two rounds: given a query set and a corpus, a system is required to retrieve and rank documents from the corpus for each query. As for the English subtask, it features the replicability and reproducibility experiments of CENTRE, launched since WWW-2 in NTCIR-14 in addition to traditional ad-hoc retrieval task. All participating runs were required to process 80 topics from WWW-2 and 80 new topics so that replicability and reproducibility can be studied. SogouT16-B and ClueWeb12-B13 were used as document collections for Chinese and English subtasks, respectively. Traditional result pooling technique is adopted in the result annotation process. The same evaluation metrics as in past rounds including nDCG@10 (MSnDCG@10), Q@10, and nERR@10 are adopted in assessing ad-hoc retrieval performances.

3.6 Data Search Task (DataSearch)

The data search task aims to help locate data resources. In the first round of this pilot task, the task organizers focus on an ad-hoc retrieval task on two statistical data collection published by the Japanese government (e-Stat, 1.3 million pages) and the US government (Data.gov, 0.2 million pages), respectively. The Japanese query topics are extracted from question-answer pairs containing data links on a Japanese CQA portal (Yahoo! Chiebukuro) and the English topics are translated from Japanese ones. The task organizers develop several baseline systems with traditional ad-hoc retrieval techniques such as BM25, LM, and BM25+RM3. Relevance judgments for top-ranked results for training queries from these baseline systems are annotated by crowd-sourcing services. Both baseline systems and these relevance judgments are provided to the participants for training purposes. After the participants submitted results, a traditional result pooling techniques (depth = 10) is adopted and nDCG, ERR, and Q-measure are used as evaluation metrics.

3.7 Micro Activity Retrieval (MART)

Since NTCIR-12, the Lifelog task has been promoting advances in information access systems for personal sensor data, which are records of multiple aspects of one's life in digital form. After three rounds of Lifelog task, the task organizers started the new succeeding task named MART which focuses on micro-activity detection and retrieval for life-log data. Different from previous tasks which pay attention to long-duration event segmentation tasks, micro-activities refer to activities that occur over short time-scales, such as minutes. The datasets used in the MART task are captured by instrumenting volunteers with a suite of multi-modal sensors alongside capturing computer interactions as they completed 20 pre-defined activities. It contains sensory information collected from 7 volunteers by life-log camera, bio-signal sensors (such as EOG, HR), and computer plugin softwares (such as mouse movements, screenshots). A total of 420 activities are collected and a third of them are adopted for testing purpose. The query topic set contains 20 queries (one for each activity). Each submitted result was a ranking list of the 140 activities in the test set. Average precision is adopted to evaluate the performance of submitted systems.

4 Summary

We reported the NTCIR-15 conference held in December 2020, which attracted 277 participants from 22 countries\regions. We also briefly introduced NTCIR-15 tasks (DialEval-1, FinNum-2, QA Lab-PoliInfo-2, SHINRA2020-ML, WWW-3, DataSearch and MART). We hope that readers are interested in NTCIR and participate in the NTCIR-16 tasks.

References

Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen. Overview of the ntcir-15 finnum-2 task: Numeral attachment in financial tweets. *Proceedings of the NTCIR-15 Conference*, 850(194):1–44, 2020.

Charles L. A. Clarke, Noriko Kando, Yiqun Liu, and editors. Makoto P. Kato. Proceedings of the 15th ntcir conference on evaluation of information access technologies. In *Proceedings of the NTCIR-15 Conference*, 2020.

Graham Healy, T-K Le, Hideo Joho, Frank Hopfgartner, and Cathal Gurrin. Overview of ntcir-15 mart. In *Proceedings of the NTCIR-15 Conference*. Sheffield, 2020.

Makoto P Kato, Hiroaki Ohshima, Ying-Hsang Liu, and Hsin-Liang Chen. Overview of the ntcir-15 data search task. In *Proceedings of the NTCIR-15 Conference*, 2020.

Yasutomo Kimura, Hideyuki Shibuki, Hokuto Ototake, Yuzu Uchida, Keiichi Takamaru, Madoka Ishioroshi, Teruko Mitamura, Masaharu Yoshioka, Tomoyoshi Akiba, Yasuhiro Ogawa, et al. Overview of the ntcir-15 qa lab-poliinfo-2 task. In *Proceedings of the NTCIR-15 Conference*, 2020.

Yiqun Liu, Makoto P. Kato, and Noriko Kando. Overview of NTCIR-15. In *In Proceedings of the NTCIR-15 Conference*, 2020.

Tetsuya Sakai, Sijie Tao, Zhaohao Zeng, Yukun Zheng, Jiaxin Mao, Zhumin Chu, Yiqun Liu, Maria Maistro, Zhicheng Dou, Nicola Ferro, et al. Overview of the ntcir-15 we want web with centre (www-3) task. In *Proceedings of the NTCIR-15 Conference*, 2020.

Satoshi Sekine, Masako Nomoto, Kouta Nakayama, Asuka Sumida, Koji Matsuda, and Maya Ando. Overview of shinra2020-ml task. In *Proceedings of the NTCIR-15 Conference*, 2020.

Zhaohao Zeng, Sosuke Kato, Tetsuya Sakai, and Inho Kang. Overview of the ntcir-15 dialogue evaluation (dialeval-1) task. In *Proceedings of the NTCIR-15 Conference*, 2020.