Report on the 16th Round of NII Testbeds and Community for Information Access Research (NTCIR-16)

Takehiro Yamamoto
University of Hyogo
Japan
t.yamamoto@sis.u-hyogo.ac.jp

Zhicheng Dou
Renmin University of China
China
dou@ruc.edu.cn

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

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

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

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

Abstract

This is a report on the NTCIR-16 conference held online in June 2022. 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. 53 active research groups from 20 countries/regions participated in one or more of the 10 different tasks in NTCIR-16. This report introduces the highlights of the conference and describes the scope and task designs of 10 tasks organized at NTCIR-16.

Date: 14–17 June, 2022.

Website: https://research.nii.ac.jp/ntcir/ntcir-16/.

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 sixteenth round of NTCIR, NTCIR-16,

started in December 2020 and was concluded in June 2022, with the NTCIR-16 conference held online.

The conference began with a tutorial given by Tetsuya Sakai, entitled *Evaluating Evaluation Measures, Evaluating Information Access Systems, Designing and Constructing Test Collections, and Evaluating Again*. The main conference was initiated by an overview of NTCIR-16, and followed by the first keynote given by ChengXiang Zhai, entitled *Information Retrieval Evaluation as Search Simulation*. After that, each task was then introduced by task organizers. The second keynote was given by Ellen Voorhees entitled *Cranfield is Dead; Long Live Cranfield*. On the third and fourth days of the conference, task organizers organized their own sessions, where selected task participants had oral presentations on their approaches and results. Poster sessions were also arranged where tasks participants could exchange information and ideas on these tasks. The conference was wrapped up by the third keynote given by Falk Scholer entitled *The Impact of Query Variability and Relevance Measurement Scales on Information Retrieval Evaluation*, followed by three invited talks about CLEF 2022 from Nicola Ferro, TREC Neural CLIR Track from Douglas W. Oard, and MediaEval from Martha Larson and Gareth Jones. The details of the conference are reported in Section 2.

There were 10 tasks organized in NTCIR-16: six *core* tasks (Data Search 2, DialEval-2, FinNum-3, Lifelog-4, QA Lab-PoliInfo-3, and WWW-4) and four *pilot* tasks (RCIR, Real-MedNLP, SS, and ULTRE). The NTCIR-16 tasks cover a broad range of IA topics, and can be summarized as follows [Yamamoto and Dou, 2022]: (1) Information retrieval: modern IR tasks from data to human, and (2) Natural language processing: deep language understanding in specialized domains such as finance, politics and medical treatment. A brief introduction to these tasks are provided in Section 3.

For more details, please refer to the online proceedings[1] of NTCIR-16 [Kando et al., 2022].

2 NTCIR-16 Conference

The NTCIR-16 conference was held online from June 14 to 17, and attracted 303 participants from 35 countries/regions, which is an increase from 277 participants from 20 countries/regions at the NTCIR-15 conference.

The first day of the conference started with the tutorial given by Tetsuya Sakai. The tutorial was about the proper evaluation in information access systems. The tutorial covered a wide range of information access evaluations such as evaluation measures, test collection design, reliability of gold data, and reproducibility.

The second day of the conference started with an overview presentation from the general chairs and program committee co-chairs. ChengXiang Zhai from University of Illinois at Urbana-Champaign then gave a keynote presentation about how search simulation can be used to evaluate IR systems. He outlined the framework for the evaluation IR systems based on search simulation and discussed how the static IR test collections can support the search simulation based evaluation. He also suggested turning the static IR collections into collections of user simulations. After the keynote presentation, the overview of each task was presented by task organizers. There were ten tasks organized in NTCIR-16. The task organizers introduced their task design, data, and

[1] https://research.nii.ac.jp/ntcir/workshop/OnlineProceedings16/NTCIR/toc_ntcir.html

experimental results. The purpose of the overview presentation is to familiarize all the participants in the conference with each task. The second day ended with the second keynote presented by Ellen Voorhees from NIST. She reviewed the history of how Cranfield evaluation evolved, introduced the recent tracks in TREC, and discussed how the Cranfield paradigm impacted the community.

On the third and fourth days of the conference, the task organizers then hosted their task sessions in parallel. The task organizers selected some of the participating groups and asked them to have oral presentations in the corresponding 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. All the participants had a chance to present their work at poster sessions set up on the 3rd and 4th days of the conference. Participants mainly presented their approaches and results with their error analysis, which triggered questions from the other participants who tackled the same problem and resulted in deep discussions that may not be seen in ordinary conferences. Both the task sessions and poster sessions were hosted on Gather.

On the fourth day of the conference, the third keynote was presented by Falk Scholer from RMIT University. He talked about the differences between the test collections and user-based evaluation and discusses the relevance measurement scale and query variability. Before the closing session, three invited speakers presented their projects and shared their experiences. Nicola Ferro from University of Padua introduced the Labs organized in CLEF 2022. Douglas W. Oard from University of Maryland talked about the Neural CLIR (NeuCLIR) Track in TREC 2022. Martha Larson from Radboud University and Gareth Jones from Dublin City University presented the overview of MediaEval tasks.

3 NTCIR-16 Tasks

NTCIR-16 included six *core* tasks (Data Search 2, DialEval-2, FinNum-3, Lifelog-4, QA Lab-PoliInfo-3, and WWW-4) and four *pilot* tasks (RCIR, Real-MedNLP, SS, and ULTRE). The former targeted relatively well-known IA problems, while the latter targeted novel problems. Each NTCIR-16 task is briefly explained below (please refer to the NTCIR-16 overview paper [Yamamoto and Dou, 2022] and overview paper of each task for details).

3.1 Data Search 2 (Data Search 2) [Kato et al., 2022]

The Data Search task focuses on ad-hoc retrieval for governmental statistical data on the Web such as on data.gov and Japanese e-Stat. As the open data movement increases, the need for a technique that retrieves the relevant data is becoming more critical. For example, for a query "causes of death us 1999-2016", the system is expected to return the relevant statistical data about causes of death. The technical challenges focused on the Data Search task include matching the keyword query and metadata and handling numerical information in the statistical data.

3.2 Dialogue Evaluation 2 (DialEval-2) [Tao and Sakai, 2022]

The DialEval tasks are the successor of the Short Text Conversation (STC) tasks organized from NTCIR-12 to NTCIR-14. STC tasks focused on finding an appropriate response for a short dia-

logue. DialEval-2 focuses on developing techniques for automatically evaluating customer helpdesk dialogues.

The dialogue quality subtask requires a system to evaluate the quality of a given dialogue in terms of three criteria task accomplishment, customer satisfaction, and dialogue effectiveness. The nugget detection subtask requires a system to identify turns that help towards problem-solving. For both tasks, multiple assessors annotated the labels. The systems are asked to predict their distribution rather than predicting its mean or mode.

3.3 Investor's & Manager's Fine-grained Claim Detection (FinNum-3) [Chen et al., 2022a]

The FinNum tasks have started at NTCIR-14 aiming to better understand the numerals in financial documents. FinNum-3 aims to understand the claims in financial documents such as reports written by professional stock analysts and companies' earnings conference calls. The numerals often play an important role in deeply understanding the claims. For example, the claim "the sales growth rate may exceed 80%." makes a stronger estimation than "the sales growth rate may exceed 40%." Understanding the numerals in such claims gives us a fine-grained understanding of financial documents. FinNum-3 organized the claim detection task, in which a system is asked to identify whether the given numeral is an in-claim or out-of-claim.

3.4 Lifelog Access and Retrieval (Lifelog-4) [Zhou et al., 2022]

Lifelog-4 is the successor of the Lifelog-1, Lifelog-2, and Lifelog-3 tasks organized in the NTCIR-12, 13, and 14, respectively. The Lifelog tasks aim to foster comparative benchmarking of approaches to automatic and interactive information retrieval from multimodal lifelog archives. One of the characteristics of the Lifelog task is its dataset. Lifelog-4 uses four months of lifelog data from one active lifelogger. The dataset comprises (1) metadata such as time, location, and biometrics, (2) images recorded by the wearable camera, and (3) concepts annotated to these images.

Lifelog-4 organized one subtask called the Lifelog Semantic Access subtask, similar to the traditional ad-hoc document retrieval. Given a topic such as "find examples of when was looking inside the refrigerator at home", the system is asked to retrieve relevant images in the dataset. The system is allowed to be either automatic or interactive. The interactive system allows a user actively interact with the system.

3.5 QA Lab for Politics Information-3 (QA Lab-PoliInfo-3) [Kimura et al., 2022]

QA Lab-PoliInfo aims to explore the techniques for real-world complex question-answering tasks, especially in the political domain. QA Lab-PoliInfo-3 task mainly focused on the understanding of natural languages in local assembly minutes. For example, the QA alignment subtask asks the system to identify the relevant answer for a given question from the minutes, The question answering subtask asks the system to generate answers for questions from the minutes. Diverse political documents are provided to the participants, such as the minutes of the Tokyo Metropolitan As-

sembly, newsletters of the Tokyo Metropolitan Government, budget information of the National Diet, and several prefectures and cities.

3.6 We Want Web 4 with CENTRE (WWW-4) [Sakai et al., 2022]

Document ranking is one of the core components of a web search engine. It has been studied for decades and learning to rank algorithms have been widely applied to solve this problem in traditional search engines. In recent years, many neural retrieval and ranking algorithms have been proposed. It would be interesting to quantify the technical improvements of the recent ad-hoc ranking approaches, especially the neural ones.

The We Want Web (WWW) task series was designed to evaluate the effectiveness of ad-hoc search algorithms. Given a document corpus and a set of queries, the participants are required to return top ranked documents from the corpus for each query. The We Want Web 4 with CENTRE (WWW-4) Task is the fourth round of the WWw series. There are two main changes in WWW-4. Firstly, Chuweb21, a subset of the Common Crawl dataset, is introduced for this task. Chuweb21 contains 3,402,457 domains and 858,616,203 English web pages. Secondly, two types of relevance assessment are introduced: the Gold version given by the topic creators, and the Bronze version labeled by "normal" assessors who are neither topic creators nor topic experts.

3.7 Reading Comprehension for Information Retrieval (RCIR) [Healy et al., 2022]

The RCIR task is a new pilot task that aims to explore whether the reading comprehension measures and eye tracker signals are useful to document ranking. There are two subtasks: the comprehension-evaluation task (CET) and the comprehension-based retrieval task (CRT). The CET subtask is designed to evaluate the prediction of a person's comprehension level based on eye movement when reading a passage, and the CRT subtask aims to explore models integrating comprehension evidence into passage retrieval systems.

In the RCIR task, the task organizers created a dataset comprised of eye movements of experimental participants during their reading tasks with different constraints and manipulations, and the corresponding answers of the multiple-choice questions presented to experimental participants. The questions are used to measure the comprehension level of the participants. With such a dataset, it is possible to discover the connection between the eye movement and the comprehension level.

3.8 Real document-based Medical Natural Language Processing (Real-MedNLP) [Yada et al., 2022]

The pilot task Real-MedNLP aims to explore natural language processing techniques in the medical field. It is the successor of the four previous MedNLP tasks: MedNLP-1, MedNLP-2, MedNLPDoc, and MedWeb. Different from these previous tasks, Real-MedNLP introduces two real clinical text datasets: the MedTxt-CR corpus containing case reports and the MedTxt-RR corpus containing radiology reports. The original datasets are in Japanese and are translated into English.

With the support of the real clinical text corpus, Real-MedNLP offers subtasks on few-resource named entity recognition, and adverse drug event extraction.

3.9 Session Search (SS) [Chen et al., 2022b]

SS is a pilot task aiming at exploring good ranking models for context-aware search (i.e., session search). Existing ad-hoc search models assume each query submitted to a search engine is standalone. However, in a real search scenario, a user may issue multiple queries to a search system within a short time interval, to find the information they need. Utilizing the contextual information, such as the preceding queries and their clicks, has been shown beneficial for generating better ranking results for the current query.

SS consists of two subtasks, namely the Fully Observed Session Search (FOSS) task and the Partially Observed Session Search (POSS) task. SS uses the TianGong-ST dataset [Chen et al., 2019] for training, and merges the TianGong-SS-FSD and TianGong-Qref datasets for testing. The over 100k training sessions in TianGong-ST are sampled from real web search sessions from query logs of the Sogou search engine. Among these sessions, 2,000 are manually assessed by humans. Differently, the test sessions are extracted from field studies conducted by users.

3.10 Unbiased Learning to Ranking Evaluation Task (ULTRE) [Zhao et al., 2022]

The ULTRE task is motivated by the advances in the trending research topic "Unbiased Learning to Rank" which aims to learn a stable ranking model from the noisy and biased user behaviour data. It consists of two subtasks: the offline ULTR subtask and the online ULTR subtask.

ULTRE constructs a dataset constructed based on SogouSRR. The dataset includes 1,200 queries sampled from Sogou.com and HTML sources of their top 10 search results. ULTRE uses real click logs to train and calibrate click models. These models are then used to generate synthetic user clicks for training queries for both subtasks. Human relevance labels are used to evaluate the performance over the test queries.

Lastly, languages of each task are shown in Table 1 for highlighting the linguistic diversity of the NTCIR-16 tasks. English is the most covered language, followed by Chinese and Japanese. It can be seen that NTCIR-16 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 Summary

We reported the NTCIR-16 conference held online in June 2022, which attracted 303 participants from 35 countries/regions. We also briefly introduced NTCIR-16 tasks (Data Search 2, DialEval-2, FinNum-3, Lifelog-4, QA Lab-PoliInfo-3, WWW-4, RCIR, Real-MedNLP, SS, and ULTRE). We hope that readers are interested in NTCIR and participate in the NTCIR-17 tasks[2].

[2] https://research.nii.ac.jp/ntcir/ntcir-17/

Table 1: Languages used in each NTCIR-16 task.

Task	Chinese	Japanese	English
Data Search 2		✓	✓
DialEval-2	✓		✓
FinNum-3	✓		✓
Lifelog-4			✓
QA Lab-PoliInfo-3		✓	
WWW-4			✓
RCIR			✓
Real Med-NLP		✓	✓
SS	✓		
ULTRE	✓		
	4	3	7

References

Chung-Chi Chen, Hen-Hsen Huang, Yu-Lieh Huang, Hiroya Takamura, and Hsin-Hsi Chen. Overview of the NTCIR-16 FinNum-3 task: Investor's and manager's fine-grained claim detection. In *Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies*, 2022a.

Jia Chen, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. Tiangong-st: A new dataset with large-scale refined real-world web search sessions. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 2485–2488, 2019.

Jia Chen, Weihao Wu, Jiaxin Mao, Beining Wang, Fan Zhang, and Yiqun Liu. Overview of the NTCIR-16 session search (SS) task. In *Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies*, 2022b.

Graham Healy, Tu-Khiem Le, Mai Boi Quach, Minh-Triet Tran, Thanh-Binh Nguyen, and Cathal Gurrin. Overview of the NTCIR-16 RCIR task. In *Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies*, 2022.

Noriko Kando, Charles L. A. Clarke, Makoto P. Kato, Yiqun Liu, Takehiro Yamamoto, and Zhicheng Dou, editors. *Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies*, 2022.

Makoto P. Kato, Hiroaki Ohshima, Ying-Hsang Liu, Hsin-Liang Chen, and Yu Nakano. Overview of the NTCIR-16 Data Search 2 task. In *Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies*, 2022.

Yasutomo Kimura, Hideyuki Shibuki, Hokuto Ototake, Yuzu Uchida, Keiichi Takamaru, Madoka Ishioroshi, Kazuma Kadowaki, Masaharu Yoshioka, Tomoyosi Akiba, Yasuhiro Ogawa, Minoru Sasaki, Ken-Ichi Yokote, Tatsunori Mori, Kenji Araki, Teruko Mitamura, and Satoshi Sekine.

Overview of the NTCIR-16 QA Lab-PoliInfo-3 task. In *Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies*, 2022.

Tetsuya Sakai, Sijie Tao, Zhumin Chu, Maria Maistro, Yujing Li, Nuo Chen, Nicola Ferro, Junjie Wang, Ian Soboroff, and Yiqun Liu. Overview of the NTCIR-16 WeWantWeb with CENTRE (WWW-4) task. In *Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies*, 2022.

Sijie Tao and Tetsuya Sakai. Overview of the NTCIR-16 dialogue evaluation (DialEval-2) task. In *Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies*, 2022.

Shuntaro Yada, Yuta Nakamura, Shoko Wakamiya, and Eiji Aramaki. Real-MedNLP: Overview of real document-based medical natural language processing task. In *Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies*, 2022.

Takehiro Yamamoto and Zhicheng Dou. Overview of NTCIR-16. In *Proceedings of the NTCIR-16 Conference*, 2022.

Yurou Zhao, Zechun Niu, Feng Wang, Jiaxin Mao, Qingyao Ai, Yang Tao, Junqi Zhang, and Yiqun Liu. Overview of the NTCIR-16 unbiased learning to rank evaluation (ULTRE) task. In *Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies*, 2022.

Liting Zhou, Cathal Gurrin, Graham Healy, Hideo Joho, Thanh-Binh Nguyen, Rami Albatal, and Frank Hopfgartner. Overview of the NTCIR-16 Lifelog-4 task. In *Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies*, 2022.