Overview of 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

ABSTRACT

This is an overview of NTCIR-16, the sixteenth sesquiannual research project for evaluating information access technologies. NTCIR-16 involved various evaluation tasks related to information retrieval, natural language processing, question answering, etc. 10 tasks were organized in NTCIR-16. This paper describes an outline of NTCIR-16, which includes its organization, schedule, scope, and task designs. In addition, we introduce brief statistics of the NTCIR-16 participants. Readers should refer to individual task overview papers for their detailed descriptions and findings.

1 INTRODUCTION

Since 1997, the NTCIR project has promoted research efforts for enhancing Information Access (IA) technologies such as Information Retrieval, Question Answering, and Natural Language Processing technologies. Its general purposes are to (1) Offer a research infrastructure that allows researchers to conduct a 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 the effectiveness of novel IA techniques. Moreover, in the collaboration process, it is expected that deep insight into research problems is successfully shared among researchers. The ongoing NTCIR-16 aims to benefit all researchers who wish to advance their research efforts. For the details and characteristics of what has been proposed in NTCIR, readers should refer to the book [7].

2 OUTLINE OF NTCIR-16

2.1 Organization

The project of NTCIR-16 was directed by General Co-Chairs (GCCs): Charles Clarke (University of Waterloo), Noriko Kando (National Institute of Informatics), Makoto P. Kato (Tsukuba University), and Yiqun Liu (Tsinghua University). Under the supervision of GCCs, Program Committee (PC) reviews task proposals that were submitted according to a call for proposal and made acceptance decisions for NTCIR-16. The members of the PC are Ben Carterette (Spotify) Hsin-Hsi Chen (National Taiwan University), Nicola Ferro (University of Padova), Gareth Jones (Dublin City University), Yiqun Liu (Tsinghua University), Jian-Yun Nie (University de Montreal), Douglas Oard (University of Maryland), Tetsuya Sakai (Waseda University), Mark Sanderson (RMIT University), and Ian Soboroff (NIST). After the review by PC, organizers of accepted tasks have promoted research activities of NTCIR-16 under the coordination of the two Program Co-Chairs (PCCs).

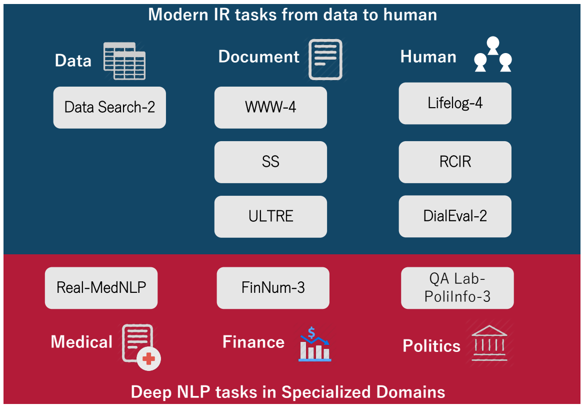

Figure 1: Overview of the NTCIR-16 task.

2.2 Schedule and Research Activities

A call for task proposals was released in October 2020, and 8 tasks (5 core tasks and 3 pilot tasks) were selected. To encourage more diverse tasks, a call for additional task proposals was released in December 2022, and two tasks (one core task and one pilot task) were selected. In total, 6 core tasks and four pilot tasks were organized in NTCIR-16. While Lifelog-4 and RCIR were selected in the additional call for task proposals, other tasks were selected in the call for task proposals. Actual NTCIR-16 activities started in January 2021, and a kickoff event was held in March 2021. 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 the organizers. New test collections were created based on the evaluation of results that were submitted by participants. The research outcome will be reported at the NTCIR-16 conference to be held online from June 14th to 17th, 2022.

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. Figure 1 summarizes the 6 core tasks (Data Search 2, DialEval-2, FinNum-3, Lifelog-4, QA Lab-PoliInfo-3, and WWW-4) and 4 pilot tasks (RCIR, Real-MedNLP, SS, and ULTRE) organized in the NTCIR-16.

- Information Retrieval: Modern IR tasks from data to human.
- Natural Language Processing: Deep language understanding in specialized domains such as finance, politics and medical treatment.

It is intresting to see that the task covers a wide range of research topics in information retrieval such as ad-hoc retrieval (WWW-4), session search (SS), data retrieval (Data Search 2), lifelog retrieval

(Lifelog-4), dialogue evaluation (DialEval-2), and unbiased learning to rank (ULTRE). It is also worth mentioning that some tasks fucused on a specialized domain: FinNum-3 focuses on finantial documents, QA Lab-PoliInfo-3 focuses on political documents, and Real-MedNLP uses medical records.

3 OUTLINE OF NTCIR-16 TASKS

3.1 Data Search 2 (Core Task) [5]

As the open data movement increases, the need for a technique that retrieves the relevant data available on the Web is becoming more critical. Data Search 2, which is the second round of the Data Search task started from NTCIR-15, aims to explore and evaluate techniques for data search.

Data Search 2 consists of three subtasks, namely, the ad-hoc data retrieval (IR) subtask, question answering (QA) subtask, and user interface (UI) subtask. The collection was constructed from e-Stat (for Japanese subtask) and data.gov (for English subtask) as in the previous round. In the previous round, the topics were prepared from a community question answering service. In this round, the organizers collected web pages referring to the dataset and prepared the additional topics from these web pages. The performance of the submitted runs was evaluated with several metrics such as nDCG, nERR, and Q-measure.

3.2 DialEval-2 (Core Task) [9]

DialEval-2 is the successor of the DialEval-1 task at NTCIR-15 and the Short Text Conversation (STC) tasks which were organized from NTCIR-12 to NTCIR-14. As the automatic evaluation of the customer helpdesk dialogues is quite important, DialEval-2 aims to develop the techniques for automatically evaluating customer helpdesk dialogues. In DialEval-2, DCH-2 data collection, which is constructed from real human customer-helpdesk dialogues on Weibo, is used as the training and development set. Since the Chinese dialogues are translated into English, the dataset is provided as a parallel corpus of Chinese and English.

The dialogue quality subtasks require a system to evaluate the quality of a given dialogue in terms of three criteria task accomplishment, customer satisfaction, and dialogue effectiveness. Since the participants are asked to predict the distribution of each score, the evaluation metrics such as NMD and RSNOD, which are designed for cross-bin metrics, were used. The nugget detection subtask requires a system to identify turns that help towards problem-solving. RNSS and JSD are used for evaluating the nugget detection task.

3.3 FinNum-3 (Core Task) [1]

The FinNum-1 task started at NTCIR-14 aiming to better understand the numerals in financial documents. FinNum-3, which is the successor of the FinNum-1 and FinNum-2 tasks, aims to understand the claims in financial documents. The numerals often play an important role in a deep understanding of 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. The system is also asked to classify the relevant category for the numeral. The reports written by professional stock analysts are used for Chinese subtask whil ethe transcriptions of companies' earnings conference calls are used for English subtask. The participants' approaches to the task are diverse such as data augmentation, numerical representation, knowledge-based, and traditional machine learning.

3.4 Lifelog-4 (Core Task) [12]

LifeLog-4 is the successor of the LifeLog-1, LifeLog-2, and LifeLog-3 tasks which are organized in the NTCIR-12, 13, and 14, respectively. The aim of the Lifelog task is 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 the LSC'20 dataset, which contains 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 (LEST) subtask, which is similar to the traditional ad-hod document retrieval. Given a topic, the system is required to retrieve relevant images in the dataset. The topics contain ad-hoc topics and know-item topics. An example ad-hoc topic is “find examples of when was looking inside the refrigerator at home.“ The system is allowed to be either automatic or interactive. The interactive system allows a user actively interact with the system while the automatic system allows no interaction from the user except for query formulation.

3.5 QA Lab-PoliInfo-3 (Core Task) [6]

QA Lab-PoliInfo aims to explore the techniques for real-world complex question answering tasks. The spread of fake news is becoming a critical social problem. Precise understanding of the facts and opinions in political documents, which can be seen as a primary source of political information, are important to combat fake news.

QA Lab-PoliInfo-3 task is the third round of the QA Lab-PoliInfo task, which started at the NTCIR-14. Taking over the success of QA Lab-PoliInfo-1 and QA Lab-PoliInfo-2, the QA LabPoliInfo-3 task organized a variety of subtasks, namely, QA Alignment, Question Answering, Fact Verification, and Budget Argument Mining subtasks. Diverse political documents are also provided to the participants, such as the minutes of the Tokyo Metropolitan Assembly, newsletters of the Tokyo Metropolitan Government, budget information of the National Diet, and several prefectures and cities.

3.6 WWW-4 (Core Task) [8]

The We Want Web with CENTRE (WWW-4) Task is the fourth round of the WWW task series which aims to evaluate the effectiveness of adhoc web search algorithms. Ranking is a core component of web search engines, and it has been studied for decades. In recent years, many new neural retrieval algorithms are proposed and it would be interesting to quantify the technical improvements of the recent approaches. WWW-4 focuses on the adhoc English web search task, and it keeps the same requirement of previous tasks:

given a query set and a document corpus, returning top ranked documents from the corpus for each query.

In WWW-4, there are two main changes. Firstly, a new English web corpus, namely Chuweb21, is introduced. Chuweb21 is a subset of the Common Crawl dataset and it contains 3,402,457 domains and 858,616,203 web pages. Secondly, two versions 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 RCIR (Pilot Task) [4]

RCIR is a pilot task that aims to understand the incorporation of reading comprehension measures and eye tracker signals into the process of document ranking. RCIR consists of two subtasks: a comprehension-evaluation task (CET) and a comprehension-based retrieval task (CRT). The former aims to predict a person's comprehension level by exploiting eye movement information when reading a passage, and the latter aims to explore the methods of integrating comprehension evidence into passage retrieval systems.

RCIR creates a dataset by collecting the 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.

In terms of evaluation measures, RCIR uses Spearman's correlation coefficient for the CET subtask, and uses normalized Discounted Cumulative Gain for the CRT subtask.

3.8 Real-MedNLP (Pilot Task) [10]

The pilot task Real-MedNLP is designed to explore the natural language processing techniques in medical fields. 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 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 SS (Pilot Task) [3]

SS is a pilot task aiming at exploring good ranking models for context-aware search (i.e., session search). Existing adhoc 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 proved 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 [2] 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 ULTRE (Pilot Task) [11]

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.

4 PARTICIPANTS

Figure 2 shows the number of *active* (those who submitted results) participants. In the figure, the numbers are given for all the tasks from NTCIR-1 to NTCIR-16. At NTCIR-16, 53 research groups have participated in the tasks. The number of participants is almost the same as in the previous round. Note that some research groups participated in multiple tasks, which were counted as different groups. Readers should refer to the individual task overview papers for getting the picture of participants' approaches to each task. Also, they should refer to the participants' individual papers for detailed descriptions of their methods.

5 CONCLUSIONS

This paper described the overview of the sixteenth cycle of NTCIR carried from January 202 to June 2022. NTCIR-16 has 10 evaluation tasks, which can be categorized into (1) traditional and novel information retrieval evaluation problems, and (2) natural language understanding in specialized domains. Most parts of the test collections developed by NTCIR-16 evaluation tasks will be released to non-participating research groups in the near future.

6 ACKNOWLEDGEMENTS

We would like to thank the organizers of all NTCIR-16 tasks for their tremendous amount of efforts devoted to running successful tasks, the task participants for their valuable contributions to the IA research community, and program committee members for their great suggestions for our accepted tasks. Finally, we would like to thank the current and past members of the NTCIR office for their continuous and careful support of our activity.

REFERENCES

[1] Chung-Chi Chen, Hen-Hsen Huang, Yu-Lieh Huang, Hiroya Takamura, and Hsin-Hsi Chen. 2022. Overview of the NTCIR-16 FinNum-3 Task: Investor's and Manager's Fine-grained Claim Detection. In *Proceedings of the NTCIR-16 Conference*.

[2] Jia Chen, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2019. 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, CIKM 2019, Beijing, China, November 3-7, 2019*, Wenwu

Zhu, Dacheng Tao, Xueqi Cheng, Peng Cui, Elke A. Rundensteiner, David Carmel, Qi He, and Jeffrey Xu Yu (Eds.). ACM, 2485–2488. https://doi.org/10.1145/3357384.3358158

[3] Jia Chen, Weihao Wu, Jiaxin Mao, Beining Wang, Fan Zhang, and Yiqun Liu. 2022. Overview of the NTCIR-16 Session Search (SS) Task. In *Proceedings of the NTCIR-16 Conference*.

[4] Graham Healy, Tu-Khiem Le, Mai Boi Quach, Minh-Triet Tran, Thanh-Binh Nguyen, and Cathal Gurrin. 2022. Overview of the NTCIR-16 RCIR Task. In *Proceedings of the NTCIR-16 Conference*.

[5] Makoto P. Kato, Hiroaki Ohshima, Ying-Hsang Liu, Hsin-Liang Chen, and Yu Nakano. 2022. Overview of the NTCIR-16 Data Search 2 Task. In *Proceedings of the NTCIR-16 Conference*.

[6] 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. 2022. Overview of the NTCIR-16 QA Lab-PoliInfo-3 Task. In *Proceedings of the NTCIR-16 Conference*.

[7] Tetsuya Sakai, Douglas W Oard, and Noriko Kando. 2021. *Evaluating Information Retrieval and Access Tasks: NTCIR's Legacy of Research Impact.* Springer Nature.

[8] Tetsuya Sakai, Sijie Tao, Zhumin Chu, Maria Maistro, Yujing Li, Nuo Chen, Nicola Ferro, Junjie Wang, Ian Soboroff, and Yiqun Liu. 2022. Overview of the NTCIR-16 WeWantWeb with CENTRE (WWW-4) Task. In *Proceedings of the NTCIR-16 Conference*.

[9] Sijie Tao and Tetsuya Sakai. 2022. Overview of the NTCIR-16 Dialogue Evaluation (DialEval-2) Task. In *Proceedings of the NTCIR-16 Conference*.

[10] Shuntaro Yada, Yuta Nakamura, Shoko Wakamiya, and Eiji Aramaki. 2022. Real-MedNLP: Overview of REAL document-based MEDical Natural Language Processing Task. In *Proceedings of the NTCIR-16 Conference*.

[11] Yurou Zhao, Zechun Niu, Feng Wang, Jiaxin Mao, Qingyao Ai, Yang Tao, Junqi Zhang, and Yiqun Liu. 2022. Overview of the NTCIR-16 Unbiased Learning to Rank Evaluation (ULTRE) Task. In *Proceedings of the NTCIR-16 Conference*.

[12] Liting Zhou, Cathal Gurrin, Graham Healy, Hideo Joho, Thanh-Binh Nguyen, Rami Albatal, and Frank Hopfgartner. 2022. Overview of the NTCIR-16 Lifelog-4 Task. In *Proceedings of the NTCIR-16 Conference*.

Year	**1999**	**2001**	**2002**	**2004**	**2005**	**2007**	**2008**	**2010**	**2011**	**2013**	**2014**	**2016**	**2017**	**2019**	**2020**	**2022**
Task/NTCIR round	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Total number	37	39	61	74	79	81	80	66	102	108	93	97	71	47	52	53
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) –> Real-MedNLP(16)										12	12	8		9		10
QA Lab for Entrance Exam (QALab) (11, 12, 13) -> QA Lab for Political Information (QALab-PoliInfo) (14, 15, 16)											11	12	11	13	14	12
Temporal Information Access (Temporalia) (11, 12)											8	14				
Cooking Recipe Search (RecipeSearch) (11)											4					
Personal Lifelog Organisation & Retrieval (Lifelog) (12, 13, 14, 16)												8	4	6		3
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) -> We Want Web with CENTER (WWW) (15, 16)													5	4	8	3
Fine-Grained Numeral Understanding in Financial Tweet (FinNum) (14,15,16)														6	7	7
CLEF/NTCIR/TREC REproducibility (CENTRE) (14)														1		
Dialogue Evaluation (DialEval) (15, 16)															7	4
SHINRA 2020 Multi-lingual (SHINRA 2020-ML) (15)															7	
Data Search (Data Search) (15, 16)															5	6
Micro Activity Retrieval Task (MART) (15)															5	
Session Search (SS) (16)																3
Reading Comprehension for Information Retrieval (RCIR) (16)																3
Unbiased Learning to Ranking Evaluation Task (ULTRE) (16)																2

Figure 2: Number of active participants (from NTCIR-1 to NTCIR-16).