Overview of NTCIR-15

Yiqun Liu
Tsinghua University
yiqunliu@tsinghua.edu.cn

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

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

ABSTRACT

This is an overview of NTCIR-15, the fifteenth sesquiannual research project for evaluating information access technologies. NTCIR-15 involved various evaluation tasks related to information retrieval, information recommendation, question answering, natural language processing, *etc.* (in total, seven tasks were organized at NTCIR-15). 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-15 Conference. Readers should refer to individual task overview papers for their detailed descriptions and findings.

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-15 aims to be beneficial to all researchers who wish to advance their research efforts.

2 OUTLINE OF NTCIR-15

2.1 Organization

The project of NTCIR-15 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-15. 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), Diane Kelly (University of Tennessee), Douglas Oard (University of Maryland), Maarten de Rijke (University of Amsterdam), 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-15 under the coordination of the two Program Co-Chairs (PCCs).

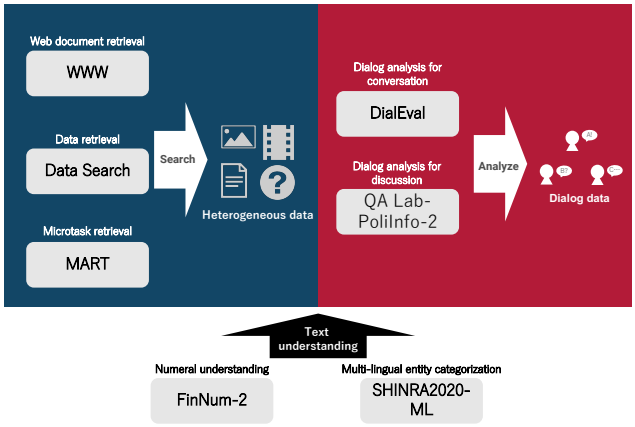

Figure 1: Overview of the NTCIR-15 tasks.

2.2 Schedule and Research Activities

Call for task proposals was released in May 2019, and seven tasks of NTCIR-15 were decided in June 2019. Accepted tasks were introduced by the organizers at the NTCIR-14 Conference. Actual NTCIR-15 activities started in July 2019, and a kick-off event was held in September 2019. In total, five core tasks and two pilot tasks (see below) were organized in NTCIR-15. 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-15 Conference to be held in Tokyo, from December 8th to 11th, in 2020.

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 (DialEval-1, FinNum-2, QA Lab-PoliInfo-2, SHINRA2020-ML, and WWW-3) and two pilot task (Data Search and MART) can be summarized as follows (illustrated in Figure 1):

(1) Modern retrieval tasks
(2) Dialogue analysis
(3) Finer-grained text understanding

It is interesting that two tasks, QA Lab-PoliInfo-2 and DialEval, are dealing with dialogue data in this round of NTCIR. The scope of search is extended to Web documents, open data, and microtasks. FinNum-2 and SHINRA2020-ML are fundamental NLP tasks, which can be potentially useful to address the other NTCIR-15 tasks.

3 OUTLINE OF NTCIR-15 TASKS

3.1 DialEval-1 (Core Task) [7]

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.

In both subtasks, a customer-helpdesk dialogue is defined as a multi-round textual dialogue that has two speakers: a Customer and a Helpdesk. In the DQ subtask, the participanting system is required to assign quality scores to each dialogue in terms of three subjective criteria: task accomplishment, customer satisfaction, and efficiency. In the ND subtask, the system needs to decide whether a dialogue turn helps solve the customer's problem.

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 FinNum-2 (Core Task) [1]

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. Given a target numeral and a cashtag, the task is defined as a binary classification problem to tell if the given numeral is related to the given cashtag.

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 10340 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 QA Lab-PoliInfo-2 (Core Task) [4]

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 polical 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.

The stance classification subtask aims at estimating politician's position by classifying their stance into two categories (agreement or disagreement) for each agenda. The dialog summarization task aims at summarizing the transcript of local assembly and ROUGE-1 Recall is used to evaluate system performance. The entity linking task is to extract the mentions of "law name" (a political term) from the local assembly member's utterances and link them to Wikipedia entries. The topic detection task is to make a list of argument topics from news-flashes of local assembly minutes and the task is not quantitatively evaluated.

3.4 SHINRA2020-ML (Core Task) [6]

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 but as a first step, The SHINRA2020-ML task focus on classification of Wikipedia pages in 30 languages into a well-defined category named Extended Named Entity (ENE).

In the task organizers' previous works, they have classified major Japanese Wikipedia pages (920K pages) into ENE categories. The multi-language editions of a classified entry are then adopted for training in their corresponding languages in the task. For example, out of the over two million Wikipedia pages in German, 275K pages have a language link from Japanese Wikipedia, which can potentially serve as a-bit-noisy training data for German Wikipedia classification. The ratio of training data for each language varies from 4.8% (Swedish) to 46.2% (Thai).

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 WWW-3 (Core Task) [5]

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 retreival 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. Since WWW-2 in NTCIR-14, a subtask named CENTRE was also introduced to evaluate the reproducibility of submitted runs in CLEF, NTCIR and TREC.

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 CENTRE

replicability and reproducibility experiments in addition to traditional adhoc retreival task. Both subtasks have 80 new topics: 80 Chinese topics were created first, and these were then translated into English. All participating runs were required to process 160 topics (80 topics from WWW-2 and the new 80 topics) so that, at least in the English subtask, replicability and reproducibility can be studied.

SogouT16-B (about 18 million Web pages) and ClueWeb12-B13 (about 50 million Web pages) were used as document collections for Chinese and English subtasks, respectively. The task organizers also provide baseline results (by BM25), online retreival/page rendering services, and topics/qrels from past two rounds to help the participants. For the Chinese subtask, an additional corpus named Sogou-QCL is also provided, which consists of large-scale weak relevance labels generated by click models.

Traditional result pooling technique is adopted in the result annotation process. 32,375 English query-result pairs (pool depth = 15) and 11,712 Chinese ones (pool depth = 30) were annotated. The same evaluation metrics as in past rounds including nDCG@10 (MSnDCG@10), Q@10, and nERR@10 are adopted in assessing ad-hoc retreival performances. Replicability and reproducibility are also quantified for the CENTRE task.

3.6 Data Search (Pilot Task) [3]

With the rapid growth of data processing capacity, open data movement has attracted much attention in recent years to encourage collaboration among researchers all around the world. Many open-data government initiatives are launched and besides government portals, there are also a large number of Web-based data repositories. These services benefits researchers in various areas by providing supports from data science and this should be regarded as a major opportunity for both computer and information science communities.

The data search task aims to help locate data resources. In the first round of this pilot task, the task organizers focus on an adhoc 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 (Pilot Task) [2]

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 lifelog 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 lifelog camera, biosignal sensors (for EOG, HR), and computer plugin softwares (for mouse movements, screenshots, etc). 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 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-15. Task overview papers (see References) describe evaluation of the results submitted by the participants. At NTCIR-15, 52 research groups have participated in the tasks. The number of participants slightly increased from the previous round, but is still small compared to the past rounds. Note that some research groups participated in two tasks, which were counted as different groups.

5 CONCLUSIONS

This paper presented the overview of the 15th cycle of NTCIR carried out from July 2019 to December 2020. NTCIR-15 has seven evaluation tasks, which can be categorized into modern retrieval tasks, dialogue analysis, and finer-grained text understanding. Most parts of the test collections developed by NTCIR-15 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-15 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] Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen. 2020. Overview of the NTCIR-15 FinNum-2 Task: Numeral Attachment in Financial Tweets. In *NTCIR-15 Conference*.

[2] Graham Healy, Tu-Khiem Le, Hideo Joho, Frank Hopfgartner, and Cathal Gurrin. 2020. Overview of NTCIR-15 MART. In *NTCIR-15 Conference*.

[3] Makoto P. Kato, Hiroaki Ohshima, Ying-Hsang Liu, and Hsin-Liang Chen. 2020. Overview of the NTCIR-15 Data Search Task. In *NTCIR-15 Conference*.

[4] Yasutomo Kimura, Hideyuki Shibuki, Hokuto Ototake, Yuzu Uchida, Keiichi Takamaru, Madoka Ishioroshi, Teruko Mitamura, Masaharu Yoshioka, Tomoyosi Akiba, Yasuhiro Ogawa, Minoru Sasaki, Kenichi Yokote, Tatsunori Mori, Kenji Araki,

Table 1: Number of active participants (from NTCIR-1 to NTCIR-15)

Year	1999	2001	2002	2004	2005	2007	2008	2010	2011	2013	2014	2016	2017	2019	2020
Task/NTCIR round	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Total number	37	39	61	74	79	81	80	66	102	108	93	97	71	47	52
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, 15)											11	12	11	13	14
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) → We Want Web with CENTRE (WWW) (15)													5	4	8
Fine-Grained Numeral Understanding in Financial Tweet (FinNum) (14)														6	7
CLEF/NTCIR/TREC REproducibility (CENTRE) (14)														1	
Dialogue Evaluation (DialEval) (15)															7
SHINRA 2020 Multi-lingual (SHINRA2020-ML)															7
Data Search (Data Search)															4
Micro Activity Retrieval Task (MART)															5

Satoshi Sekine, and Noriko Kando. 2020. Overview of the NTCIR-15 QA Lab-PoliInfo-2 Task. In *NTCIR-15 Conference*.

[5] Tetsuya Sakai, Sijie Tao, Zhaohao Zeng, Yukun Zheng, Jiaxin Mao, Zhumin Chu, Yiqun Liu, Maria Maistro, Zhicheng Dou, Nicola Ferro, and Ian Soboroff. 2020. Overview of the NTCIR-15 We Want Web with CENTRE (WWW-3) Task. In *NTCIR-15 Conference*.

[6] Satoshi Sekine, Masako Nomoto, Kouta Nakayama, Asuka Sumida, and Koji Matsuda. 2020. Overview of SHINRA2020-ML Task. In *NTCIR-15 Conference*.

[7] Zhaohao Zeng, Sosuke Kato, Tetsuya Sakai, and Inho Kang. 2020. Overview of the NTCIR-15 Dialogue Evaluation (DialEval-1) Task. In *NTCIR-15 Conference*.