Overview of NTCIR-13

Makoto P. Kato
Kyoto University
mpkato@acm.org

Yiqun Liu
Tsinghua University
yiqunliu@tsinghua.edu.cn

ABSTRACT

This is an overview of NTCIR-13, the thirteenth sesquiannual research project for evaluating information access technologies. NTCIR-13 presents a diverse set of evaluation tasks related to information retrieval, question answering, natural language processing, etc (in total, nine tasks have been organized at NTCIR-13). 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-13 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, 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 ongoing NTCIR-13 aims to be beneficial to all researchers who wish to advance their research efforts.

2. OUTLINE OF NTCIR-13

2.1 Organization

The overall project of NTCIR-13 was directed by General Co-Chairs (GCCs): Charles L. A. Clarke (Facebook, USA), Noriko Kando (National Institute of Informatics, Japan), and Tetsuya Sakai (Waseda University, 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-13. The members of the PC are Tat-Seng Chua (National University of Singapore, Singapore), Nicola Ferro (University of Padua, Italy), Kal Jarvelin (University of Tampere, Finland), Gareth Jones (Dublin City University, Ireland), Makoto P. Kato (Co-chair, Kyoto University, Japan), Chin-Yew Lin (Microsoft Research Asia, China), Yiqun Liu (Co-chair, Tsinghua University, China), Maarten de Rijke (University of Amsterdam, the Netherlands), Mark Sanderson (RMIT University, Australia), and Ian Soboroff (NIST, USA). After the review by PC, organizers of accepted tasks have promoted research activities of NTCIR-13 under the coordination of two Program Co-Chairs (PCCs), which are authors of this paper.

2.2 Schedule and Research Activities

Call for task proposals was released on March 2016, and the tasks of NTCIR-13 were finally determined in May 2016, a month before the NTCIR-12 Conference. Accepted tasks were introduced by the organizers at the NTCIR-12 Conference. Actual NTCIR-13 activities started on July 2016, and a kick-off event was held on August 2016. In addition, call for additional pilot task proposals was released on September 2016. Unfortunately, however, no task proposal was submitted at that round. In total, five core tasks and four pilot tasks (see below) have been organized in NTCIR-13. 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 have been created based on evaluation of results that were submitted by participants. The research outcome will be reported at the NTCIR-13 Conference to be held in Tokyo, from December 5th to 8th in 2017.

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 (Lifelog-2, MedWeb, OpenLiveQ, QALab-3, STC-2) and four pilot tasks (AKG, ECA, NAILS, WWW) can be summarized as follows (illustrated in Figure 1):

1. Answering complex questions and queries through deep understanding of text and user intents;
2. Mining knowledge from a large amount of human generated data; and
3. Application of knowledge extracted from big data to intelligent IA technologies.

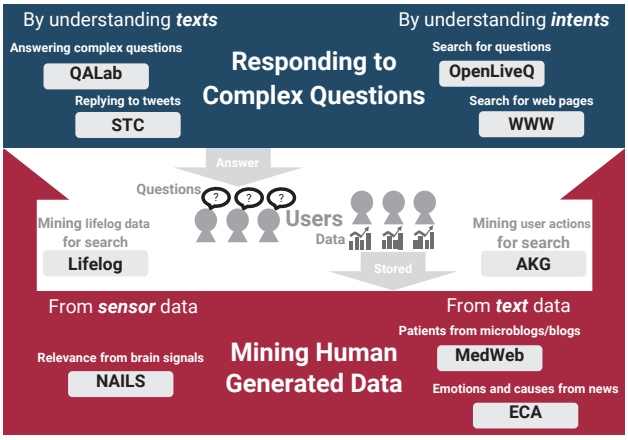

Figure 1: Illustration of NTCIR-13 tasks.

3. OUTLINE OF NTCIR-13 TASKS

3.1 Lifelog-2 (Core Task) [3]

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.

Dealing with lifelogging data is non-trivial task and should involve collaborative researches with both multimedia analytics and IR research communities. To prompt such collaborations, the task organizers also organize a workshop in ACM MM2017 conference. 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.

3.2 MedWeb (Core Task) [9]

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.

3.3 OpenLiveQ (Core Task) [5]

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 multileaved 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.

3.4 QALab-3 (Core Task) [8]

This is the third round of QALab task in NTCIR. 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 (synonyms are 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.

3.5 STC-2 (Core Task) [7]

Short Text Conversation (STC-2) task follows the efforts of STC-1 at NTCIR-12, which attempts to develop systems replying a short answer to the user in response to her/his short question. This is highly correlated with the recently hot topic of conversational system in both academic and industrial researches. 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 con-

sidered. 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 retrieval-based and generation-based methods by comprehensive evaluations.

3.6 AKG (Pilot Task) [1]

AKG task is a new pilot task in NTCIR-13. It aims to help users (especially search engine users) to gain actionable suggestions after submitting a query containing entities. The definition of "Actionable 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.

3.7 ECA (Pilot Task) [2]

ECA is a newly proposed pilot task which aims to locate the stimuli, or the cause of emotions besides just identifying the emotions. 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.

3.8 NAILS (Pilot Task) [4]

NAILS is a newly proposed pilot task which 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.

Description of the data set can be found at the task organizers' CHIIR workshop paper in NeuroIIR 2017. 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.

Table 1: Number of groups by country and region.

Country / Region	# Groups
Japan	27
China	22
Taiwan	6
USA	5
Australia	2
Portugal	1
Hong Kong	1
Singapore	1
Total	65

3.9 WWW (Pilot Task) [6]

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. This new pilot task named We Want Web (WWW) can be regarded as a recent forum in the continuous efforts to tackle basic Web search problems. Considering that deep learning techniques have brought back researchers' interests in solving basic ranking problems, it may be a nice timing to set up a standard benchmark for the community to share ideas and compare methods.

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.

4. PARTICIPANTS AND RESULTS

Table 2 shows the numbers of participants who submitted results. In this table, the numbers are given for all the tasks from NTCIR-1 to NTCIR-13. Task overview papers (see References) describe evaluation of the results submitted by the participants. At NTCIR-13, 71 research groups have participated in the tasks and the number of participants decreases from NTCIR-12 (*i.e.* 97 groups). Note that some research groups participated in two tasks, which were counted as different groups. The decrease from the previous round was anticipated as five out of nine tasks were newly proposed tasks at NTCIR-13. The 63 unique groups include over 200 members in total, from which some people would attend the conference. Table 1 shows geographical distribution of participants in NTCIR-13. Japan and China are dominant countries, but some groups have participated from USA, Australia and other areas in the world. In total, 8 countries or regions appears in Table 1[1], which is a big drop

[1]The total number of unique groups in this table is larger than that of unique groups, since some groups consist of members from two different countries and are counted twice.

Table 2: Number of participants (from NTCIR-1 to NTCIR-13)

Year	1999	2001	2002	2004	2005	2007	2008	2010	2011	2013	2014	2016	2017
Task/NTCIR round	1	2	3	4	5	6	7	8	9	10	11	12	13
Total number	37	39	61	74	79	81	80	66	102	108	93	97	71
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)											11	12	11
Temporal Information Access (Temporalia) (11, 12)											8	14	
Cooking Recipe Search (RecipeSearch) (11)											4		
Personal Lifelog Organisation & Retrieval (Lifelog) (12, 13)												8	4
Short Text Conversation (STC) (12, 13)												22	27
Open Live Test for Question Retrieval (OpenLiveQ) (13)													7
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)													5

from the previous round (*i.e.* 20 countries at NTCIR-12).

5. CONCLUSIONS

This paper presented the overview of the 13th cycle of NTCIR activity carried out from July 2016 to December 2017. NTCIR-13 has nine evaluation tasks, which suggests the great diversity of IA challenges addressed by this project. Most parts of the test collections developed by NTCIR-13 evaluation tasks will be released to non-participating research groups in the near future.

We faced the decrease of participants compared to the previous rounds possibly due to the drastic changes in the organized tasks. In NTCIR-14, we are planning to bring big changes in the publication, which are expected to attract more participants and enhance the quality of NTCIR papers.

6. ACKNOWLEDGMENTS

We would like to thank the organizers of all NTCIR-13 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 feedback on 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.

7. REFERENCES

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

[2] Q. Gao, H. Jiannan, X. Ruifeng, G. Lin, Y. He, K.-F.

Wong, and Q. Lu. Overview of ntcir-13 eca task. In *NTCIR-13 Conference*, 2017.

[3] 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 *NTCIR-13 Conference*, 2017.

[4] G. Healy, T. Ward, C. Gurrin, and A. Smeaton. Overview of ntcir-13 nails task. In *NTCIR-13 Conference*, 2017.

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

[6] C. Luo, T. Sakai, Y. Liu, Z. Dou, C. Xiong, and J. Xu. Overview of the ntcir-13 we want web task. In *NTCIR-13 Conference*, 2017.

[7] 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 *NTCIR-13 Conference*, 2017.

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

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