Report on NTCIR-12: The Twelfth Round of NII Testbeds and Community for Information Access Research

Makoto P. Kato
Kyoto University, Japan
kato@dl.kuis.kyoto-u.ac.jp

Kazuaki Kishida
Keio University, Japan
kz_kishida@z8.keio.jp

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

Tetsuya Sakai
Waseda University, Japan
tetsuya@waseda.jp

Mark Sanderson
RMIT University, Australia
mark.sanderson@rmit.edu.au

Abstract

This is a report on the NTCIR-12 conference held in June 2016, in Tokyo, Japan. NTCIR-12 is the twelfth sesquiannual research project for evaluating information access technologies that organizes a diverse set of tasks related to information retrieval, question answering, and natural language processing. The NTCIR-12 conference is a venue in which task organizers and task participants presented their effort on their participating tasks, and attracted 236 participants from 21 countries/regions in this round. This report introduces the highlights of the conference, describes the scope and task designs of nine tasks organized at NTCIR-12, and provides a brief introduction to NTCIR-13, which started from June 2016 and will be closed in December 2017.

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.

The twelfth round of NTCIR, NTCIR-12, started in December 2014 and was concluded in June 2016, with the NTCIR-12 conference held in Tokyo, Japan[1]. The conference began with a satellite workshop on evaluating information access (EVIA 2016)[2] (see also an EVIA 2016 report at SIGIR Forum [7]). The main conference was initiated by an overview of NTCIR-12, and followed by a keynote given by Alistair Moffat, about desirable properties of evaluation metrics. Each task was then introduced by task organizers and further discussed at their own session, where task participants had oral presentations on their approaches. Poster sessions were arranged at every lunch during the conference and provided a place for task participants to exchange information and ideas on these tasks. The conference was wrapped up with invited talks about the *Todai* robot project, news from TREC and CLEF, and an introduction to NTCIR-13 tasks. The details of the conference are reported in Section 2.

There were nine tasks organized in NTCIR-12: six *core* tasks (IMine-2, MedNLPDoc, MobileClick-2, SpokenQuery&Doc-2, Temporalia-2, and MathIR) and three *pilot* tasks (Lifelog, QALab-2 and STC). The NTCIR-12 tasks cover a broad range of IA topics, and can be summarized as follows [6]: (1) advanced search techniques considering user context or intents, (2) IA techniques tailored to mobile computers, and (3) IR and QA techniques dependent on specific domains. A brief introduction to these tasks are provided in Section 3.

Nine tasks have been already accepted at NTCIR-13 and now call for task participation as of December 2016[3]. NTCIR-13 keeps its diversity and offers a wide range of IA tasks: Lifelog-2, MedWeb, OpenLiveQ, QALab-3, STC-2, AKG, ECA, NAILS, and WWW. These tasks are also introduced in Section 4.

For more details, please refer to the online proceedings of NTCIR-12 and EVIA 2016:

- NTCIR-12 proceedings [5]:
  `http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings12/NTCIR/toc_ntcir.html`
- EVIA 2016 proceedings [8]:
  `http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings12/EVIA/toc_evia.html`

2 NTCIR-12 Conference

The NTCIR-12 conference was held from June 7 to 10 in National Institute of Informatics (Tokyo, Japan), and attracted 236 participants from 21 countries/regions.

EVIA 2016 was organized by Charles L.A. Clarke and Emine Yilmaz, and held on the first day of the NTCIR-12 conference. The workshop consisted of a keynote presented by Yiqun Liu, seven paper presentations, and a panel discussion on the future of IA evaluation. Yiqun Liu from Tsinghua University talked about understanding and prediction of search satisfaction in a heterogeneous environment, and presented his recent work on the correlation between relevance/usefulness and satisfaction, search satisfaction with SERPs including heterogeneous contents such as news, images, and videos, and a satisfaction prediction framework based on mouse movement patterns. Seven papers presented in EVIA

[1] `http://research.nii.ac.jp/ntcir/ntcir-12/`
[2] `http://research.nii.ac.jp/ntcir/evia2016/`
[3] `http://research.nii.ac.jp/ntcir/ntcir-13/`

2016 were concerned with evaluation in diverse kinds of scenarios, e.g. methodology for building test collections, topic set size design, and new evaluation metrics. Four panelists (Gareth J.F. Jones, Yiqun Liu, Mark Sanderson, and Ian Soboroff) discussed the future of IA evaluation and EVIA at the end of the workshop.

The main conference of NTCIR-12 started with the overview presentation from general chairs and program committee co-chairs. It was reported that 47 task organizers and 91 research groups were involved in NTCIR-12. Alistair Moffat from the University of Melbourne gave a keynote presentation on "What Would We Like IR Metrics to Measure?" He showed high variance in the expected number of documents and queries users need for completing search tasks, and introduced evaluation metrics that can take into account such variance. After the keynote presentation, the overview of each task was presented by task organizers. There were nine tasks organized in NTCIR-12, and their data, task design, and experimental results were presented for familiarizing each task for all the participants in the NTCIR-12 conference.

Task organizers then hosted their own task sessions in parallel. Some participating groups were selected by task organizers and asked to have oral presentations in those 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. Participants mainly presented their approaches and results, with their own error analysis, which triggered questions from the other participants who tackled the same problem and resulted in deep discussion that may not be seen in ordinary conferences. All the participants had a chance to present their work at poster sessions during lunch. Moreover, some task organizers hosted *break-out sessions*, where organizers and participants discussed the current task and planned for the next round.

An interesting trial in this round was a remote Lifelog satellite session – a group of Lifelog task participants got together in Glasgow for ESF Science Meeting on "Evaluating Personal Lifelog" on June 9, and it was followed by the satellite session connected with NII via Skype. Although these two events were not well synchronized due to the time difference between Glasgow and Tokyo, participants could listen to presentations given by the other side.

At the last day of the NTCIR-12 conferences, three invited speakers presented their projects. Noriko Arai from National Institute of Informatics introduced the *Todai* robot project (Can a robot get into the University of Tokyo?) that aims to develop a question-answering system that can pass the exam of the University of Tokyo. Ian Soboroff from the National Institute of Standards and Technology talked about each track in TREC 2015, and announced the 25th anniversary event of TREC. Gareth J.F. Jones from Dublin City University introduced tasks in MediaEval 2016 and labs in CLEF 2015. At the end of the conference, NTCIR-13 tasks were presented by task organizers.

3 NTCIR-12 Tasks

NTCIR-12 included six *core* tasks (IMine-2, MedNLPDoc, MobileClick-2, SpokenQuery&Doc-2, Temporalia-2, and MathIR) and three *pilot* tasks (Lifelog, QALab-2 and STC). The former targeted relatively well-known IA problems, while the latter targeted novel problems. There were 157 teams registered in NTCIR-12, of which 91 teams from 20 countries/regions submitted their runs in time. Each NTCIR-12 task is briefly explained below (please refer to the NTCIR-12 overview paper [6] and overview paper of each task [12, 2, 9, 1, 4, 13, 3, 11, 10]

for details).

Search Intent and Task Mining (IMine-2)

IMine-2 task [12] was designed to develop techniques or methods of automatically identifying users' intents behind their search queries. There were two subtasks in IMine-2: query understanding and vertical incorporating. In the query understanding subtask, participants were asked to identify relevant topic types inherent in a given search query. The types are specially called *verticals* (e.g. *Web*, *Image*, *News*, *QA*, *Download*, *Encyclopedia*, and *Shopping*). The vertical incorporating subtask is to generate a diversified ranked list that can cover major search intents for a given search topic.
Distinctive data: query log data from Sogou, Inc. and Yahoo Japan Corporation.

Medical Natural Language Processing for Clinical Document (MedNLP-Doc)

MedNLPDoc task [2] focused on topics of medical information retrieval, since there are some unique difficulties in text processing of medical documents. MedNLPDoc provided two subtasks: (1) phenotyping where participants were asked to allocate ICD (international codes for diseases) to a given medical record, and (2) the creative subtask where participants were able to devise an original research problem and test it.
Distinctive data: annotated, quasi electric health records.

Mobile Information Access (MobileClick-2)

MobileClick-2 task [9] was designed to develop techniques or methods for allowing users to easily access information on the small screen of mobile devices. There were two subtasks in MobileClick-2: iUnit ranking and summarization. The iUnit denotes an *information unit*, which is defined as a piece of information that is smaller than a document. In the iUnit ranking subtask, participants were asked to rank a set of iUnits according to a given search query. The summarization subtask set the task of generating a structured textual output given a query. The output consisted of a set of iUnits and a set of intents.
Distinctive data: query log data from Yahoo Japan Corporation.

Spoken Query and Spoken Document Retrieval (SpokenQuery&Doc-2)

SpokenQuery&Doc-2 [1] mainly focused on spoken document retrieval (SDR). As speech operations for mobile devices have become popular, the importance of SDR has been increasing recently. At SpokenQuery&Doc-2, two subtasks were devised by the task organizers: spoken term detection (STD) and spoken content retrieval (SCR). STD tried to detect positions where search terms appear in spoken documents. In SCR, participants were asked to find spoken segments containing information relevant to a search topic, which can be considered as an ad-hoc retrieval task for spoken documents.
Distinctive data: lecture speech data, recordings of the annual Spoken Document Processing Workshop (SDPWS), and manual and automatic transcriptions.

Temporal Information Access (Temporalia-2)

Temporalia-2 [4] aimed at enhancing document retrieval for information needs in which time matters. The subtasks of Temporalia-2 are temporal intent disambiguation (TID) and temporally diversified retrieval (TDR). In TID, participants were asked to measure to what

extent a given search query was relevant to four classes of temporal intents, namely, *past*, *recency*, *future*, and *atemporal*. In TDR, participants tried to retrieve relevant documents for each class of temporal intents as well as *temporally diversified* search results.
Distinctive data: corpora from Sogou, Inc. and LivingKnowledge.

Mathematical Information Retrieval (MathIR)

MathIR task [13] aimed to develop techniques of mathematical information retrieval that enable users to access mathematical formulas and to understand mathematical concepts or objects in documents. Two corpora, an arXiv dataset and a set of Wikipedia articles, were used in MathIR. Participants were asked to produce a ranked list of paragraphs for each query that mainly consists of mathematical formulas and keywords. Optionally, participants could tackle a formula similarity subtask and a formula browsing subtask.
Distinctive data: corpora including mathematical formulas from arXiv and Wikipedia.

Lifelog Task (Lifelog)

Lifelog [3] was a new task at NTCIR-12. It aimed to achieve effective and efficient access to lifelog data that consist of records of individuals' experiences such as pictures and sensory data taken by wearable devices. The organizers of Lifelog set up two subtasks: a lifelog semantic access task (LSAT) and a lifelog insight task (LIT). The purpose of LSAT is to develop ad hoc retrieval systems in an interactive or an automatic manner. LIT explored knowledge mining and visualization of lifelog data without requiring any specific output from participants, in order to provide an opportunity for gaining insights into lifelog data uses.
Distinctive data: lifelog data including photos, locations, and user activities.

QA Lab for Entrance Exam (QALab-2)

QALab-2 [11] tackled development of advanced question-answering (QA) systems that can solve entrance exam questions created for Japanese universities. Questions on the subject of *world history* were selected from the National Center Test for University Admissions (multiple choice-type questions) and secondary exams at five universities in Japan (complex questions including essays). QALab-2 covered several types of question formats including complex essay, simple essay, factoid, slot-filling, and true-or-false, and provided several phases of different difficulties with question formats revealed or unrevealed.
Distinctive data: National Center Test questions and answers, university entrance examinations, textbook corpora on the subject of world history, and event ontology.

Short Text Conversation Task (STC)

STC task [10] was a new challenge, which attempted to develop systems returning a short reply in response to a short message from a human. STC at NTCIR-12 can be viewed as a task to realize computer-human conversation systems based on IR techniques: messages used as input were selected from microblogs, and replies were also retrieved from microblogs by participants' systems. STC targeted both Chinese and Japanese and turned out to be the largest task of NTCIR-12, with 22 research groups participating.
Distinctive data: microblog posts from Weibo and Twitter.

Lastly, languages of each task are shown in Table 1 for highlighting the linguistic diversity of the NTCIR-12 tasks. English and Japanese were used in many of the tasks, and Chinese was used in three tasks. It can be seen that NTCIR-12 provided opportunities to address

Table 1: Languages used in each NTCIR-12 task.

Task	English	Japanese	Chinese
IMine-2	✓	✓	✓
MedNLPDoc		✓	
MobileClick-2	✓	✓	
SpokenQuery&Doc-2		✓	
Temporalia-2	✓		✓
MathIR	✓		
Lifelog	✓		
QALab-2	✓	✓	
STC		✓	✓
	6	6	3

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 NTCIR-13 Tasks

Proposals for NTCIR-13 tasks were submitted in April 2016, and reviewed by the NTCIR-13 program committee members. Nine accepted tasks were announced at the NTCIR-12 conference, and are looking for participants as of December 2016. NTCIR-13 offers a wide range of IA tasks as NTCIR-12 did, which include five core tasks (Lifelog-2, MedWeb, Open-LiveQ, QALab-3, and STC-2), and four pilot tasks (AKG, ECA, NAILS, and WWW). The objectives of NTCIR-13 can be summarized as follows: (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 illustrates these three objectives with a central focus on the user.

Below, we briefly introduce the NTCIR-13 tasks based on the current task description (see `http://research.nii.ac.jp/ntcir/ntcir-13/` for details).

Personal Lifelog Organisation & Retrieval Task (Lifelog-2)

Lifelog-2 develops a more semantically rich test collection than that at the previous round, and proposes the continuation of two sub-tasks from NTCIR-12 (LSAT and LIT) and include one new task called lifelog insight task (LET), which focuses on exploring approaches to semantic enrichment of raw lifelog data.
Distinctive data: two lifeloggers' data including photos, music listening history, biometrics, locations, physical activities, and computer usage for 45 days.
Website: `http://ntcir-lifelog.computing.dcu.ie/`

Medical Natural Language Processing for Web Document (MedWeb)

MedWeb extends the scope of MedNLPDoc to social media for analyzing the trend of diseases

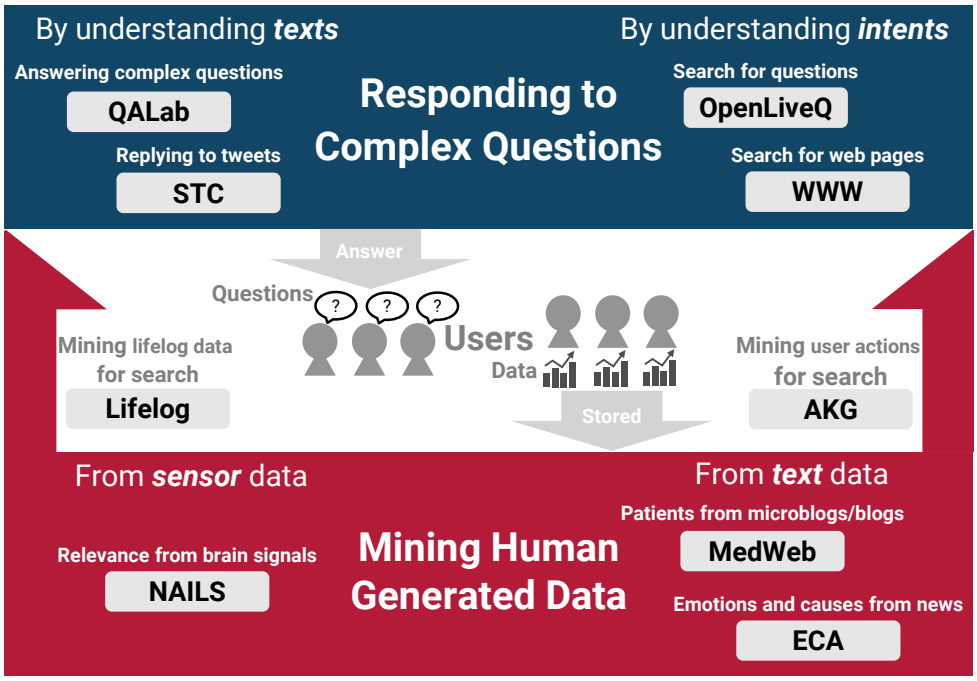

Figure 1: Overview of the NTCIR-13 tasks. Sensor and text data generated from user activities are mined (MedWeb, ECA, and NAILS) and applied to various IA tasks (Lifelog-2 and AKG). Understanding texts and user intents is also addressed for achieving complex question answering (QALab-3 and STC-2) as well as Web and question search (OpenLiveQ and WWW).

and tracking patient statuses, and requires participants to extract disease information from Japanese, English, and Chinese Twitter posts and disease journal texts.
Distinctive data: annotated Twitter posts and blog articles written by cancer patients.
Website: `http://mednlp.jp/medweb/NTCIR-13/`

Open Live Test for Question Retrieval (OpenLiveQ)

OpenLiveQ provides an open live test environment of Yahoo Japan Corporations community QA service for question retrieval systems, and aims to provide an opportunity for a more realistic evaluation to address problems specific to a production environment.
Distinctive data: clickthrough data collected at Yahoo Japan Corporations community QA search.
Website: `http://www.openliveq.net/`

QA Lab for Entrance Exam (QALab-3)

The goal of QALab-3 is to investigate the real-world complex QA technologies using Japanese university entrance exams on the subject of *world history*. Questions were selected from two different stages: the National Center Test for University Admissions (multiple choice-type questions) and from secondary exams at universities (complex questions including essays).

Distinctive data: National Center Test questions and answers, university entrance examinations, textbook corpora on the subject of world history, and event ontology.
Website: `http://research.nii.ac.jp/qalab/`

Short Text Conversation (STC-2)

STC-2 adds a new subtask that allows participants to generate a reply from scratch for a given microblog post, unlike the retrieval-based method at NTCIR-12 in which replies are retrieved from a large repository of existing microblog posts.
Distinctive data: microblog posts from Weibo and Twitter.
Website: `http://ntcirstc.noahlab.com.hk/STC2/stc-cn.htm`
Website: `http://ntcirstc.noahlab.com.hk/STC2/stc-jp.htm`

Actionable Knowledge Graph (AKG)

AKG aims to generate *actionable* (or transactional) knowledge graphs that show entity properties useful for users to take some actions included in search queries (e.g. "visiting a temple in Kyoto" and "staying at York University").
Website: `http://ntcirakg.github.io/index.html`

Emotion Cause Analysis (ECA)

ECA proposes two subtasks on emotion and cause extraction/detection from Chinese and English news articles: (1) detecting clauses that contain emotion causes, and (2) detecting the exact boundary of the emotion cause.
Distinctive data: annotated English and Chinese news articles.
Website: `http://hlt.hitsz.edu.cn/ECA.html`

Neurally Augmented Image Labelling Strategies (NAILS)

NAILS introduces IA researchers to EEG (Electroencephalography) data, and provides a opportunity for exploring the application of EEG data via an image labeling task, where participants are expected to predict relevant images based on viewers' EEG data.
Distinctive data: EEG data in raw and numerous pre-processed formats.
Website: `http://ntcir-nails.computing.dcu.ie/`

We Want Web (WWW)

WWW runs an ad hoc Web task for at least three rounds at NTCIR for quantifying the progress, and evaluate systems not only with traditional measures but also with more advanced measures that better reflect user experiences.
Distinctive data: user behavior data, new corpus and query log from Sogou, Inc.
Website: `http://www.thuir.cn/ntcirwww`

5 Summary

We reported the NTCIR-12 conference held in June 2016, which attracted 236 participants from 21 countries/regions. We also briefly introduced NTCIR-12 tasks (IMine-2, MedNLP-Doc, MobileClick-2, SpokenQuery&Doc-2, Temporalia-2, MathIR, Lifelog, QALab-2 and STC), and on-going NTCIR-13 tasks (Lifelog-2, MedWeb, OpenLiveQ, QALab-3, STC-2,

AKG, ECA, NAILS, and WWW). We hope that readers are interested in NTCIR and participate in the NTCIR-13 tasks.

6 Acknowledgements

Contributions of the NTCIR-12 program committee and organizing committee members, task organizers, data providers, student volunteers, and NTCIR office staffs were vital for the success of NTCIR-12 and the NTCIR-12 conference. We would like to express our gratitude to the following organizations for sponsoring the NTCIR-12 conference: Yahoo! JAPAN Corporation, Fuji Xerox Co. Ltd., Rakuten Institute of Technology, Japan Patent Information Organization, IR-Advanced Linguistic Technologies Inc., IBM Japan, Ltd., and kizasi Company, Inc.

References

[1] T. Akiba, H. Nishizaki, H. Nanjo, and G. J. F. Jones. Overview of the NTCIR-12 SpokenQuery&Doc-2 task. In *Proceedings of the NTCIR-12 Conference*, 2016.

[2] E. Aramaki, M. Morita, Y. Kano, and T. Ohkuma. Overview of the NTCIR-12 MedNLP-Doc task. In *Proceedings of the NTCIR-12 Conference*, 2016.

[3] C. Gurrin, H. Joho, F. Hopfgartner, L. Zhou, and R. Albatal. Overview of NTCIR-12 Lifelog task. In *Proceedings of the NTCIR-12 Conference*, 2016.

[4] H. Joho, A. Jatowt, R. Blanco, H. Yu, and S. Yamamoto. Overview of NTCIR-12 Temporal Information Access (Temporalia-2) task. In *Proceedings of the NTCIR-12 Conference*, 2016.

[5] N. Kando, K. Kishida, M. P. Kato, and S. Yamamoto, editors. *Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies*. National Institute of Informatics, 2016. `http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings12/NTCIR/toc_ntcir.html`.

[6] K. Kishida and M. P. Kato. Overview of NTCIR-12. In *Proceedings of the NTCIR-12 Conference*, 2016.

[7] C. L. A. Clarke and E. Yilmaz. EVIA 2016: The seventh international workshop on evaluating information access. In *ACM SIGIR Forum*, volume 50, December 2016.

[8] C. L. A. Clarke and E. Yilmaz, editors. *Proceedings of the Seventh International Workshop on Evaluating Information Access (EVIA 2016), a Satellite Workshop of the NTCIR-12 Conference*. National Institute of Informatics, 2016. `http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings12/EVIA/toc_evia.html`.

[9] M. P. Kato, T. Sakai, T. Yamamoto, V. Pavlu, H. Morita, and S. Fujita. Overview of the NTCIR-12 MobileClick-2 task. In *Proceedings of the NTCIR-12 Conference.*, 2016.

[10] L. Shang, T. Sakai, Z. Lu, H. Li, R. Higashinaka, and Y. Miyao. Overview of the NTCIR-12 Short Text Conversation task. In *Proceedings of the NTCIR-12 Conference*, 2016.

[11] H. Shibuki, K. Samamoto, M. Ishioroshi, A. Fujita, Y. Kano, T. Mitamura, T. Mori, and N. Kando. Overview of the NTCIR-12 QA Lab-2 task. In *Proceedings of the NTCIR-12 Conference*, 2016.

[12] T. Yamamoto, Y. Liu, M. Zhang, Z. Dou, K. Zhou, I. Markov, M. P. Kato, H. Ohshima, and S. Fujita. Overview of the NTCIR-12 IMine-2 task. In *Proceedings of the NTCIR-12 Conference*, 2016.

[13] R. Zanibbi, A. Aizawa, M. Kohlhase, I. Ounis, G. Topi, and K. Davila. NTCIR-12 MathIR task overview. In *Proceedings of the NTCIR-12 Conference*, 2016.