Overview of NTCIR-12 Pilot Task
Short Text Conversation (STC)

Lifeng Shang[a], Tetsuya Sakai[b], Zhengdong Lu[a], Hang Li[a]
Ryuichiro Higashinaka[c], Yusuke Miyao[d]

Huawei[a] Waseda[b] NTT[c] NII[d]

Outline

- Background: Short Text Conversation
  - Existing Methods
  - STC task@NTCIR-12
- Chinese Subtask
  - Dataset & evaluation methods
  - Evaluation results and Summaries of the methods
- Japanese Subtask
  - Dataset & evaluation methods
  - Evaluation results and Summaries of the methods
- Conclusion and Future work

Dialogue Systems

- Real-life Applications
  - Voice Assistant & Technical Support Service

- Big Challenges
  - Language representation & understanding
  - Context modeling
  - Reasoning with knowledge

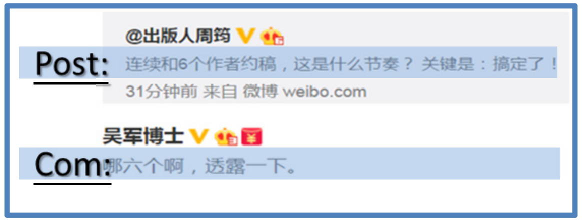
Dialogue Systems

- Task-oriented dialogue systems
  - E.g. ATIS, technical support services
  - Grammar-based, Frame-based (dialogue state tracking challenge), Information state-based methods
- Non-task-oriented systems
  - E.g. ELIZA, chatbot **Xiaoice** of MS
  - Loebner prize (Turing test)
  - Retrieval-based(**STC-1**) & Generation-based(**STC-2**)

STC belongs to non-task-oriented dialogue systems

Short-Text Conversation

- Observation
  - Post-Comment forms one round of conversation



Post:

@出版人周筠

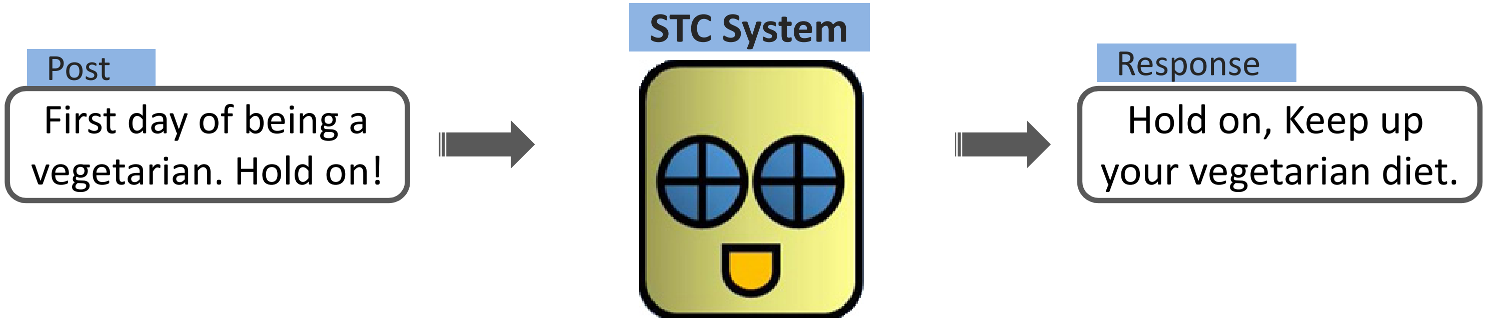
连续和6个作者约稿，这是什么节奏？ 关键是：搞定了！

31分钟前 来自 微博 weibo.com

吴军博士

Com: 哪六个啊，透露一下。

- STC Task Definition
  - Considers one round of conversation
  - Response should be coherent & useful to the post



Post

First day of being a vegetarian. Hold on!

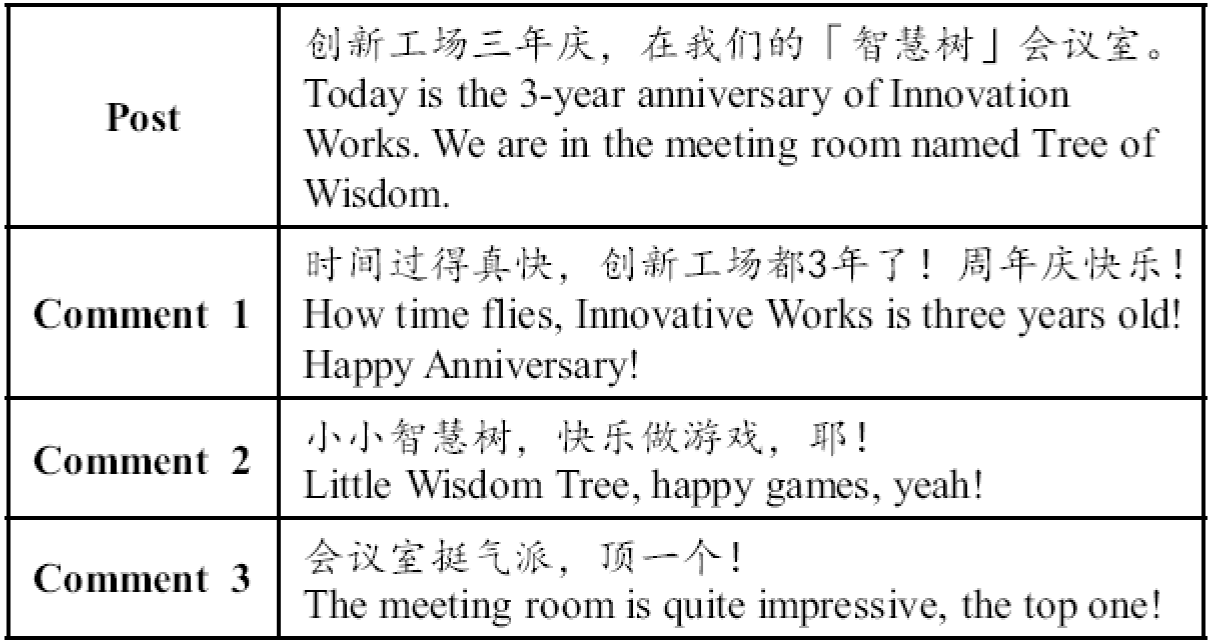
STC System

Response

Hold on, Keep up your vegetarian diet.

Example of Weibo

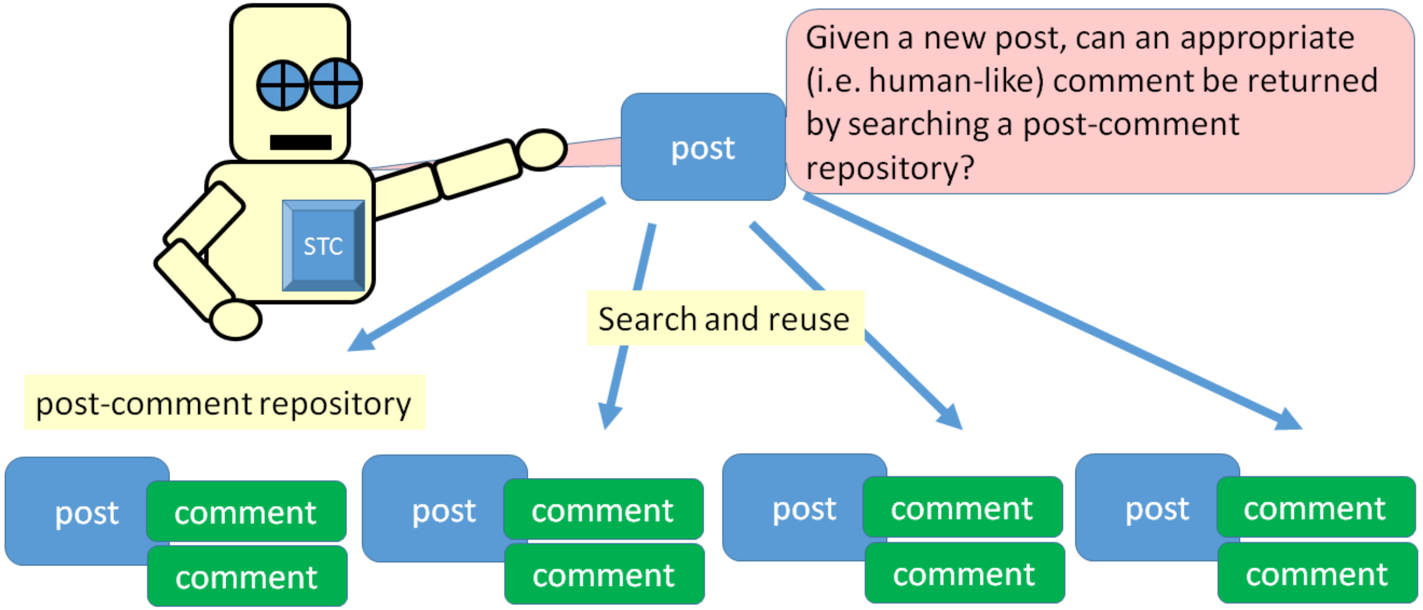
- One post multiple comments



Post	创新工场三年庆，在我们的「智慧树」会议室。 Today is the 3-year anniversary of Innovation Works. We are in the meeting room named Tree of Wisdom.
Comment 1	时间过得真快，创新工场都3年了！周年庆快乐！ How time flies, Innovative Works is three years old! Happy Anniversary!
Comment 2	小小智慧树，快乐做游戏，耶！ Little Wisdom Tree, happy games, yeah!
Comment 3	会议室挺气派，顶一个！ The meeting room is quite impressive, the top one!

STC Research Question

- The First Step at NTCIR-12
  - Take it as an IR problem
  - Build a useful dialogue system that can interact naturally with humans

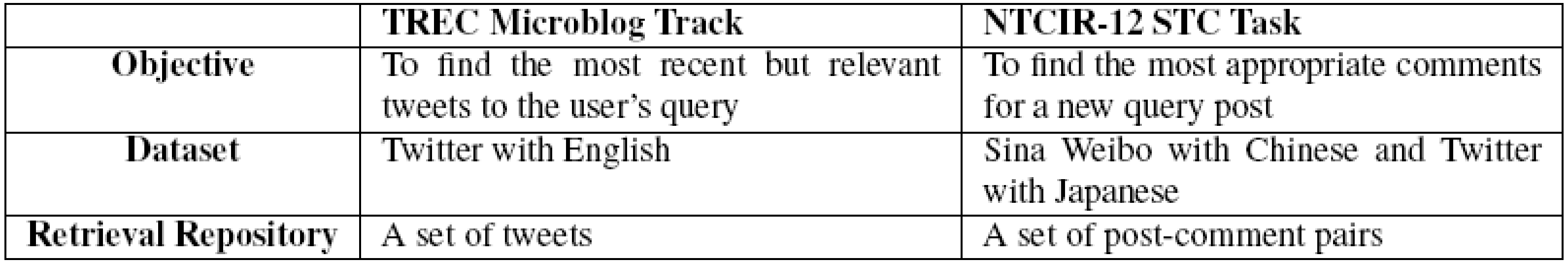


Given a new post, can an appropriate (i.e. human-like) comment be returned by searching a post-comment repository?

Related Tasks

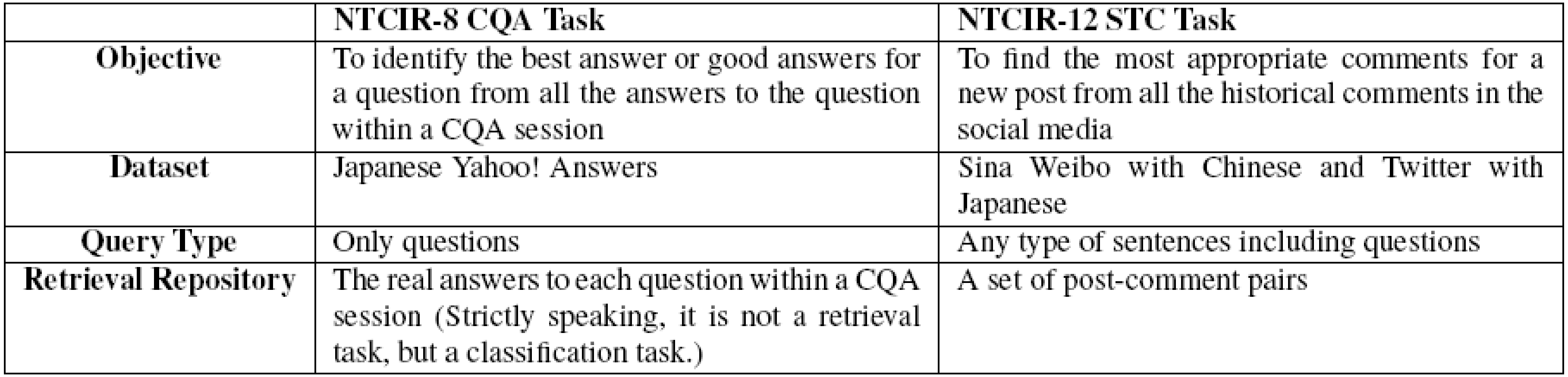
- Difference to the other tasks

Table 1: Difference between TREC Microblog Track and STC Task



	TREC Microblog Track	NTCIR-12 STC Task
Objective	To find the most recent but relevant tweets to the user's query	To find the most appropriate comments for a new query post
Dataset	Twitter with English	Sina Weibo with Chinese and Twitter with Japanese
Retrieval Repository	A set of tweets	A set of post-comment pairs

Table 2: Difference between NTCIR-8 CQA Task and STC Task

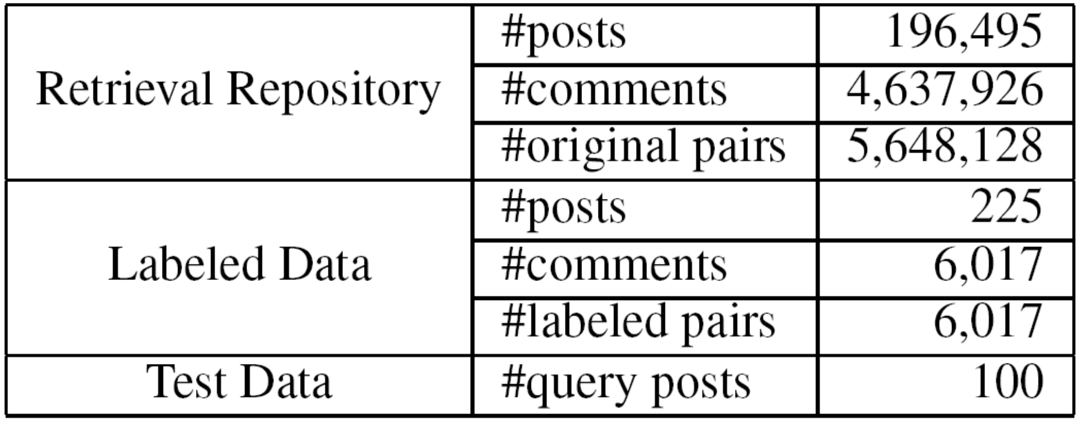


	NTCIR-8 CQA Task	NTCIR-12 STC Task
Objective	To identify the best answer or good answers for a question from all the answers to the question within a CQA session	To find the most appropriate comments for a new post from all the historical comments in the social media
Dataset	Japanese Yahoo! Answers	Sina Weibo with Chinese and Twitter with Japanese
Query Type	Only questions	Any type of sentences including questions
Retrieval Repository	The real answers to each question within a CQA session (Strictly speaking, it is not a retrieval task, but a classification task.)	A set of post-comment pairs

The Chinese Subtask

- The Construction of Dataset
  - Constructed based on our past work
  - Crawled raw data in hundreds of million scale
  - Each post has 28 different responses
  - Using Sakai's **topic set size design tool**

Statistics of dataset for Chinese subtask



Retrieval Repository	#posts	196,495
	#comments	4,637,926
	#original pairs	5,648,128
Labeled Data	#posts	225
	#comments	6,017
	#labeled pairs	6,017
Test Data	#query posts	100

The Chinese Subtask

- Evaluation Methods
  - Results are pooled to perform manual annotation.
  - Relevance is assessed from **four** criteria

    (1) *Coherent* (2) *Topically relevant*

    (3) *Context-independent* (4) *Non-repetitive*
  - Relevance labels L0, L1 and L2
  - Evaluation Measures

    Graded-relevance IR evaluation measures

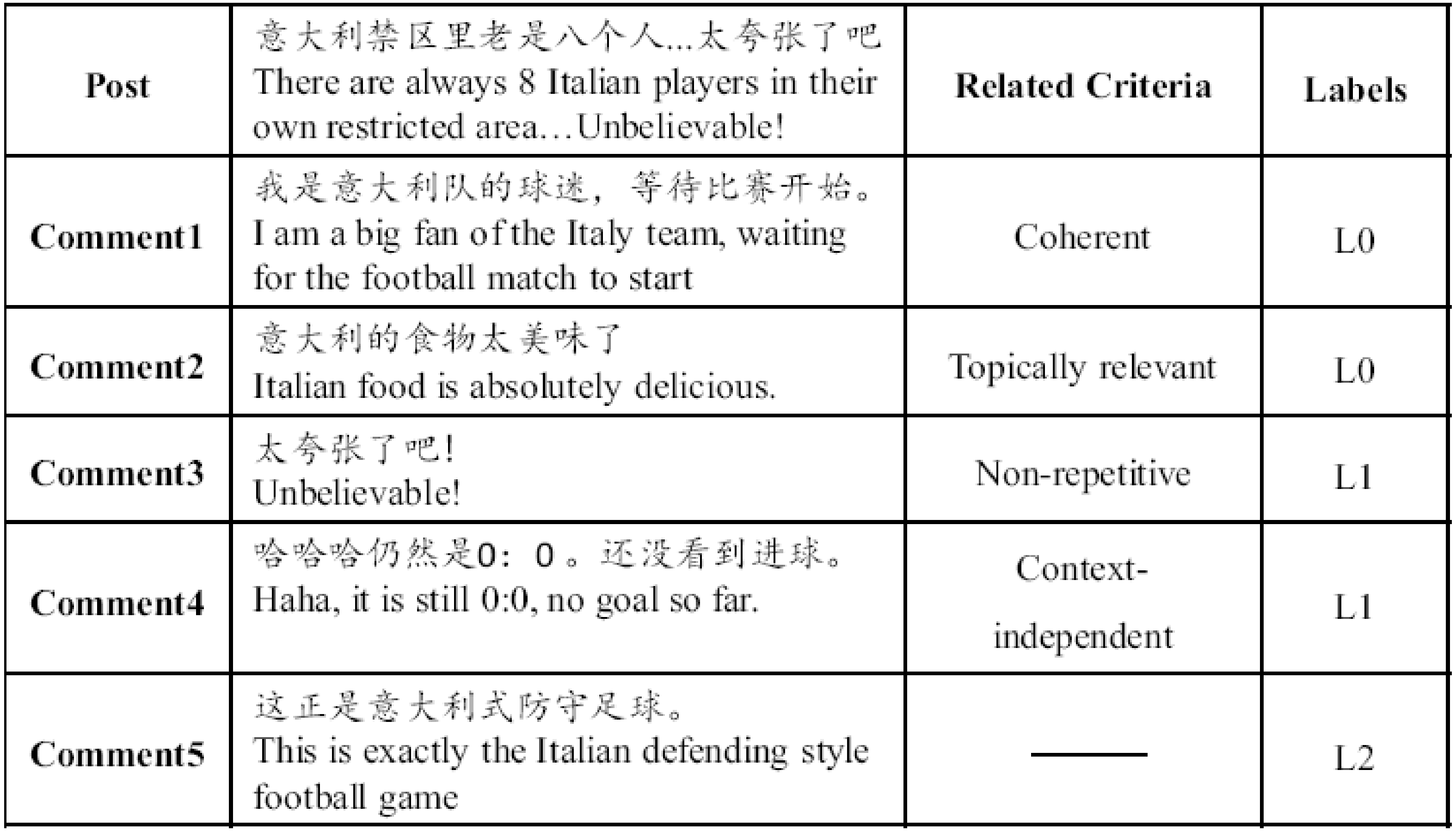
    (1) *nG@1* (2) *nERR@10* (3) P^{+}

    computed by **NTCIREVAL** tool

If either (1) or (2) is untrue, the retrieved comment should be labeled “L0”; if either (3) or (4) is untrue, the label should be “L1”; otherwise, the label is “L2”.

The Chinese Subtask

- Example of Relevance Assessment



Post	意大利禁区里老是八个人...太夸张了吧 There are always 8 Italian players in their own restricted area…Unbelievable!	Related Criteria	Labels
Comment1	我是意大利队的球迷，等待比赛开始。 I am a big fan of the Italy team, waiting for the football match to start	Coherent	L0
Comment2	意大利的食物太美味了 Italian food is absolutely delicious.	Topically relevant	L0
Comment3	太夸张了吧! Unbelievable!	Non-repetitive	L1
Comment4	哈哈哈仍然是0：0。还没看到进球。 Haha, it is still 0:0, no goal so far.	Context-independent	L1
Comment5	这正是意大利式防守足球。 This is exactly the Italian defending style football game	——	L2

The Chinese Subtask

- Evaluation Measures
  - $nG@1$: normalized gain at rank 1

$$nG@r = \frac{g(r)}{g^*(r)}$$

  - $g(r)$ denote the gain of a comment at rank r
  - let $g(r) = 2^2 - 1 = 3$ if the comment is L2-relevant
  - This is a crude measure, in our setting, it takes values $0, \frac{1}{3}$ or 1

The Chinese Subtask

- Evaluation Measures
  - $nERR@10$: expected reciprocal rank
  - Suitable for navigational intents
  - The probability the user is satisfied at rank r

$$p(r) = \frac{g(r)}{2^H}$$

  - The probability that the user reaches as far as rank r and the stops scanning the list

$$P_{r_{ERR}}(r) = p(r)\prod_{k=1}^{r-1}(1 - p(k))$$

  - the normalized one is:

$$nERR@l = \frac{\sum_{r=1}^{r} P_{r_{ERR}}(r)}{\sum_{r=1}^{r} P_{r^*_{ERR}}(r)}$$

The Chinese Subtask

- Evaluation Measures
  - P^+: similar to Q-measure, for navigational intents
  - r_p preferred rank
  - Assumption: *the distribution of users who will stop scanning the ranked list at a particular rank is uniform over all relevant documents at or above*

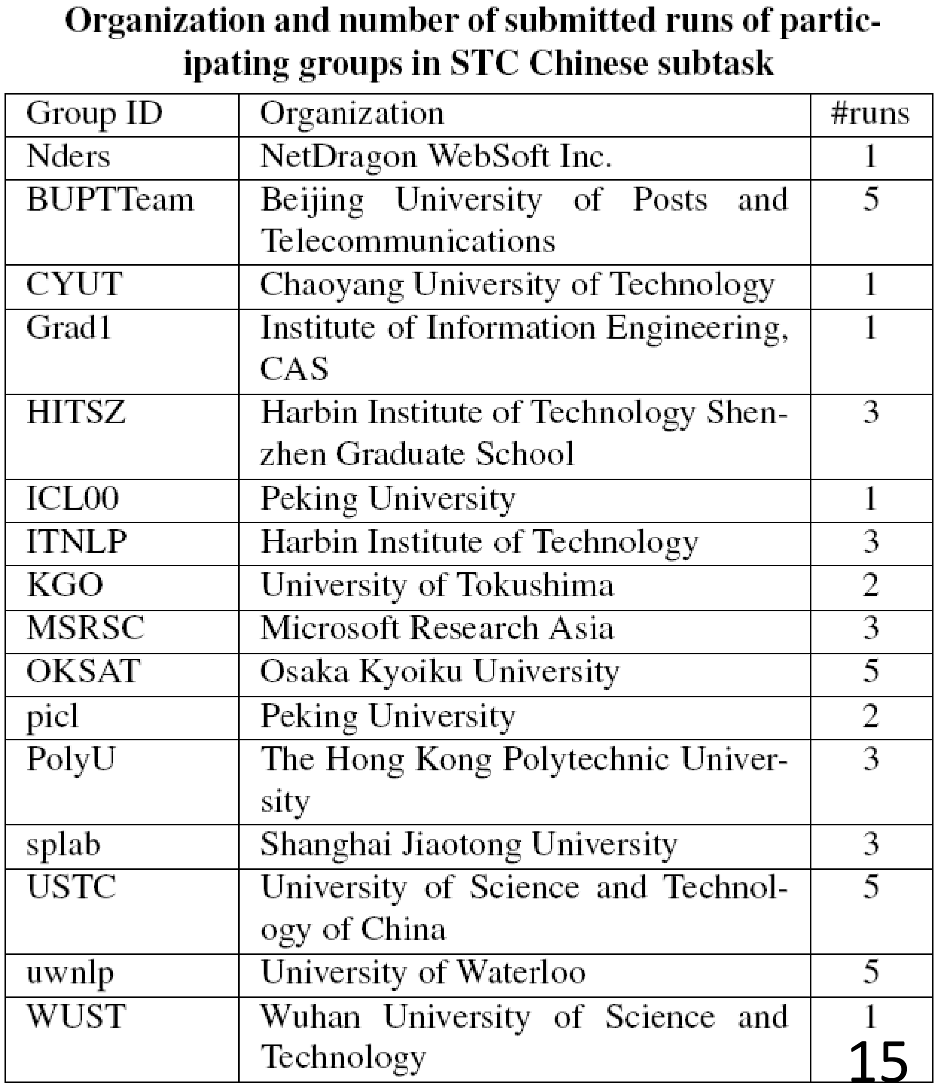
$$P^+ = \sum_r Pr_+(r)BR(r) = \frac{1}{\sum_{k=1}^{r_p} I(k)} \sum_{r=1}^{r_p} I(r)BR(r)$$

  - where blended ratio

$$BR(r) = \frac{\sum_{k=1}^{r} I(k) + \sum_{k=1}^{r} g(k)}{r + \sum_{k=1}^{r} g^*(k)}$$

The Chinese Subtask

- Participants Info.
  - – There were a total of 38 registrations, and 16 of them finally submitted 44 runs.



Organization and number of submitted runs of participating groups in STC Chinese subtask

Group ID	Organization	#runs
Nders	NetDragon WebSoft Inc.	1
BUPTTeam	Beijing University of Posts and Telecommunications	5
CYUT	Chaoyang University of Technology	1
Grad1	Institute of Information Engineering, CAS	1
HITSZ	Harbin Institute of Technology Shenzhen Graduate School	3
ICL00	Peking University	1
ITNLP	Harbin Institute of Technology	3
KGO	University of Tokushima	2
MSRSC	Microsoft Research Asia	3
OKSAT	Osaka Kyoiku University	5
picl	Peking University	2
PolyU	The Hong Kong Polytechnic University	3
splab	Shanghai Jiaotong University	3
USTC	University of Science and Technology of China	5
uwnlp	University of Waterloo	5
WUST	Wuhan University of Science and Technology	1

The Chinese Subtask

- Brief analysis: matching features
  - the similarity between two short texts

Feature Name	#(Teams)	The Teams
vector space model*: TF-IDF, word2vec	10	BUPTTeam, MSRSC, OKSAT, USTC, UWNLP, ICL00,Nders, CYUT,PolyU,WUST
lexical features (LCS, co-occurring)	4	Splab, USTC, UWNLP, ICL00,
syntactic features	1	ICL00
semantic features (CNN, seq2seq)	3	Splab, USTC,ITNLP
learning from some raw features by NN	1	ITNLP

– * the vector can be TF-IDF, word2vec, topic mode, etc.

The Chinese Subtask

- Brief analysis : re-ranking model
  - Learn to combine matching features

Ranking models	#(Teams)	The Teams
ranking SVM	2	Splab, USTC,
random walk	1	BUPTTeam
empirically determined	3	MSRSC,UWNLP, Nders
random forest	1	UWNLP
NULL	4	OKSAT, CYUT, PolyU, WUST

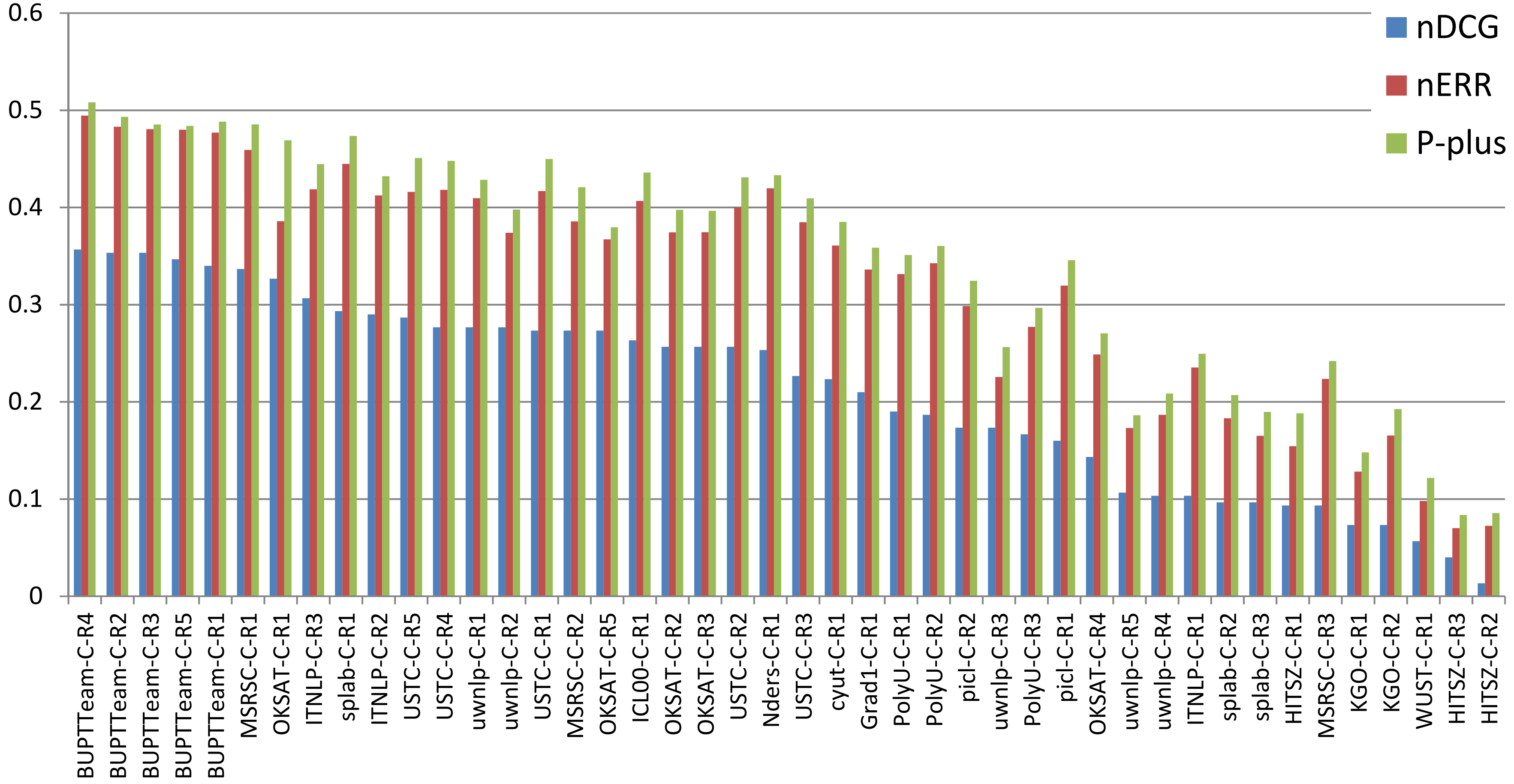
Classification models	#(Teams)	The Teams
Logistic regression	1	ITNLP
MLP	1	ITNLP
SVM	1	ICL00

The Chinese Subtask

- Brief analysis : using rules
  - – using heuristic rules to perform filtering

Heuristic Rules	#(Teams)	The Teams
considering the popularity of comments	2	UWNLP, MSRSC
considering comments by their lengths	2	PolyU, UWNLP
gave priority to short comments	1	OKSAT
filtering comments by characteristic words	1	OKSAT
adding new attributes to post & comment	1	OKSAT
building a general comments database	1	PolyU

The Chinese Subtask



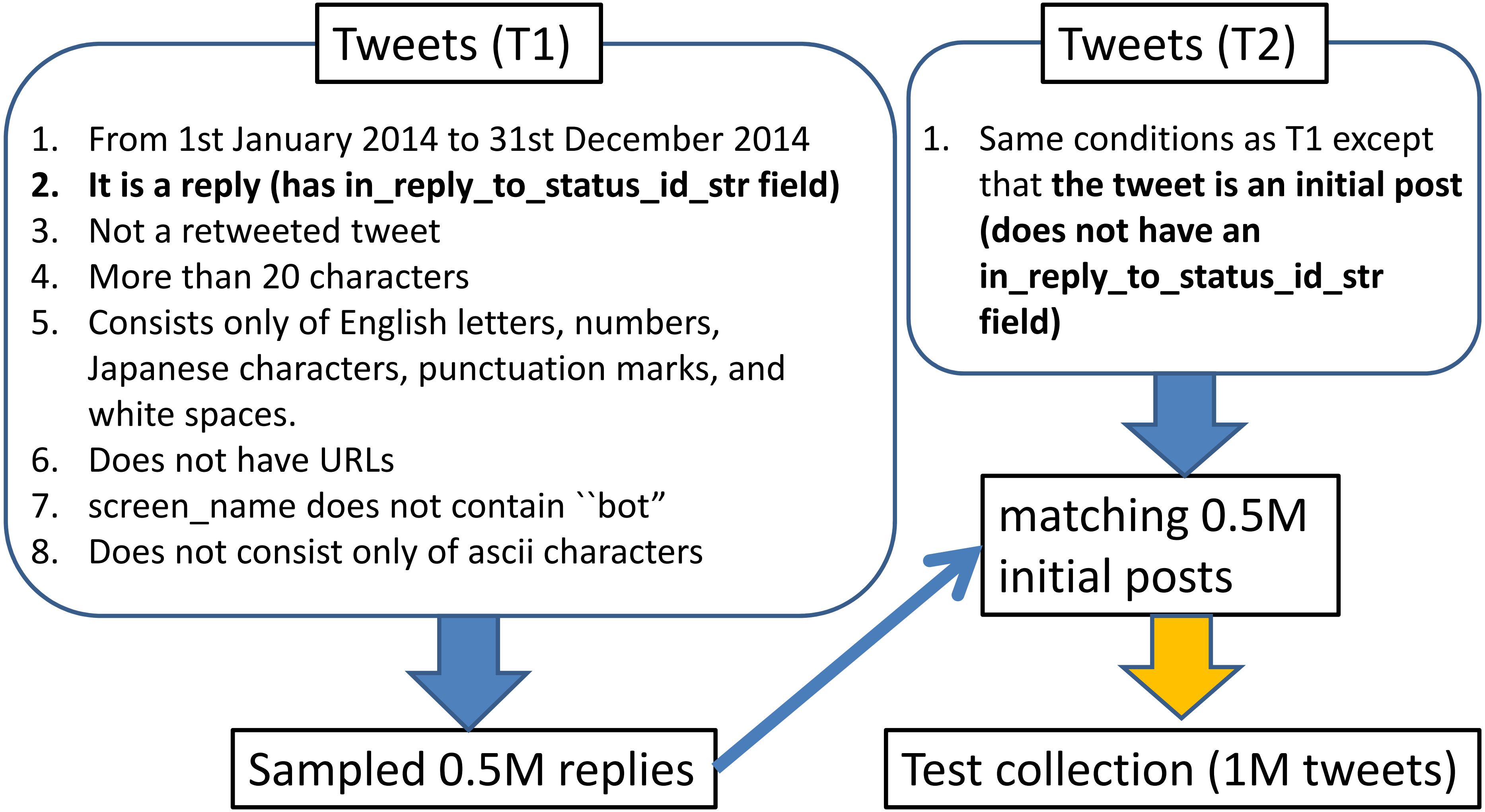
The Japanese Subtask

- Main differences from the Chinese subtask
  - Data
    - Twitter was used instead of Weibo
    - Test collection is composed of tweet-reply pairs
  - Evaluation method
    - We used multiple annotators to evaluate each retrieved tweet to cope with the subjective nature of the task
  - Evaluation measure
    - In addition to nDCG and nERR, we used accuracy
    - We did not use P+

The Japanese Subtask

- The construction of dataset
  - Test collection created by crawling Twitter
    - Due to a license issue, we provided only tweet IDs instead of raw text
  - The training data contain 1M tweets
    - 0.5M tweets (initial posts) and their replies
  - The test data contain 202 tweets
- Since tweets are deleted on a daily basis, only tweets that existed at the time of the formal run were used for evaluation

The Japanese Subtask



Tweets (T1)

1. From 1st January 2014 to 31st December 2014
2. **It is a reply (has in_reply_to_status_id_str field)**
3. Not a retweeted tweet
4. More than 20 characters
5. Consists only of English letters, numbers, Japanese characters, punctuation marks, and white spaces.
6. Does not have URLs
7. screen_name does not contain ``bot”
8. Does not consist only of ascii characters

Sampled 0.5M replies

Tweets (T2)

1. Same conditions as T1 except that **the tweet is an initial post (does not have an in_reply_to_status_id_str field)**

matching 0.5M initial posts

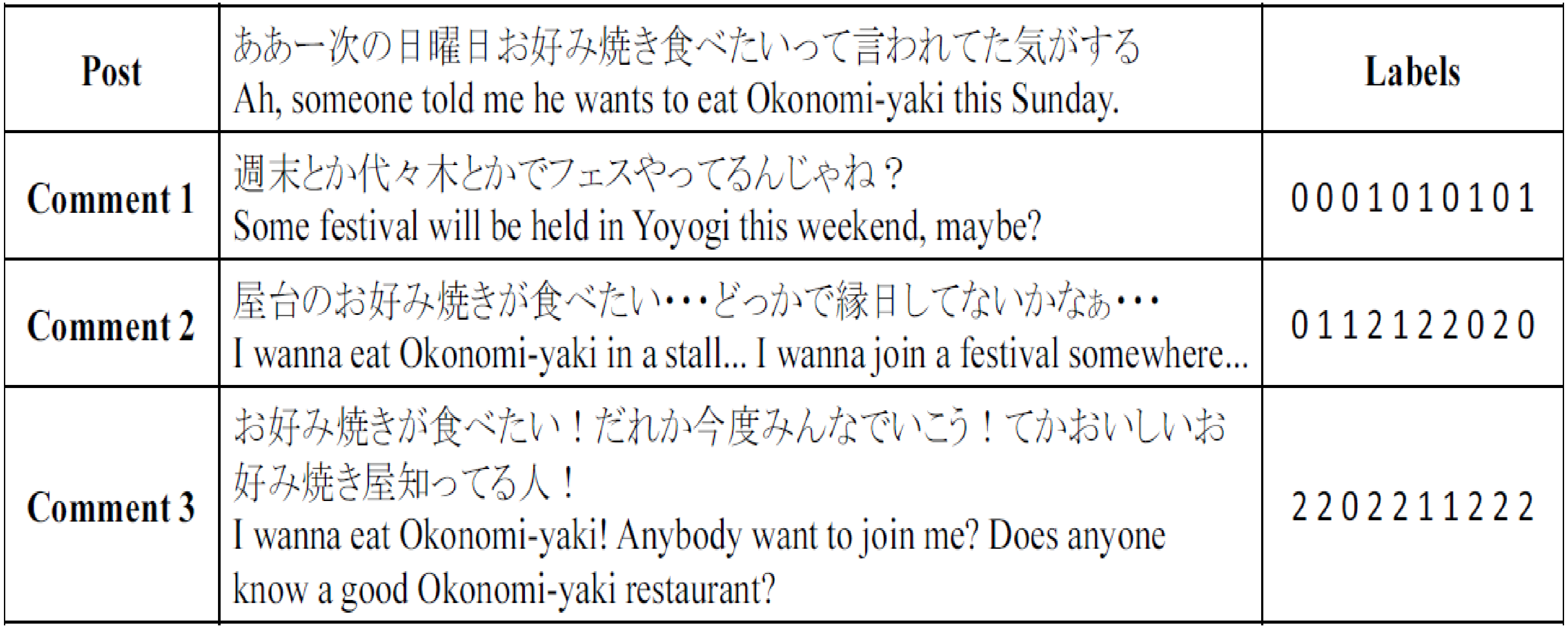
Test collection (1M tweets)

The Japanese Subtask

- Evaluation methods
  - Results are pooled to perform manual annotation
  - Retrieved tweets were annotated by ten annotators with L0, L1, and L2 labels
    - Same criterion as the Chinese subtask was used for labeling
  - Inter-annotator agreement in Fleiss' kappa
    - L0, L1, L2 => 0.317, L0, {L1, L2} => 0.421
    - confirms the subjective nature of the task

The Japanese Subtask

- Example



Post	ああー次の日曜日お好み焼き食べたいって言われてた気がする Ah, someone told me he wants to eat Okonomi-yaki this Sunday.	Labels
Comment 1	週末とか代々木とかでフェスやってるんじゃね？ Some festival will be held in Yoyogi this weekend, maybe?	0 0 0 1 0 1 0 1 0 1
Comment 2	屋台のお好み焼きが食べたい･･･どっかで縁日してないかなぁ･･･ I wanna eat Okonomi-yaki in a stall... I wanna join a festival somewhere...	0 1 1 2 1 2 2 0 2 0
Comment 3	お好み焼きが食べたい！だれか今度みんなでいこう！てかおいしいお好み焼き屋知ってる人！ I wanna eat Okonomi-yaki! Anybody want to join me? Does anyone know a good Okonomi-yaki restaurant?	2 2 0 2 2 1 1 2 2 2

The Japanese Subtask

- Evaluation measures
  - nDCG@1 and nERR@5 calculated with averaged gain:

$$g(r) = \frac{\sum_{i=1}^{n} g_i(r)}{n}$$

  - AccG@k: the ratio of correct labels (G) within top-k

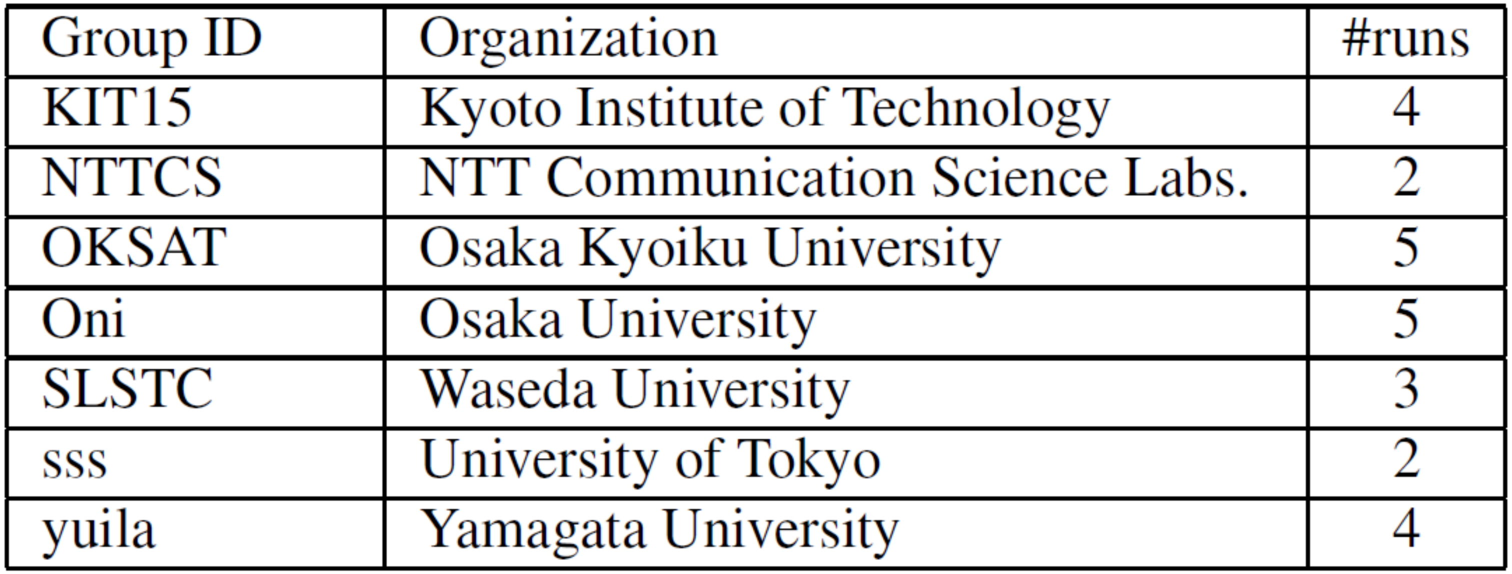
$$Acc_G@k = \frac{1}{nk}\sum_{r=1}^{k}\sum_{i=1}^{n}\delta(l_i(r) \in G)$$

G = {L2} or {L1,L2}
k = 1 or 5

  - We did not use P+ because it was not trivial to calculate the value with multiple annotators

The Japanese Subtask

- Participants INFO
  - We had a total of 12 registrations, and 7 of them finally submitted 25 runs.



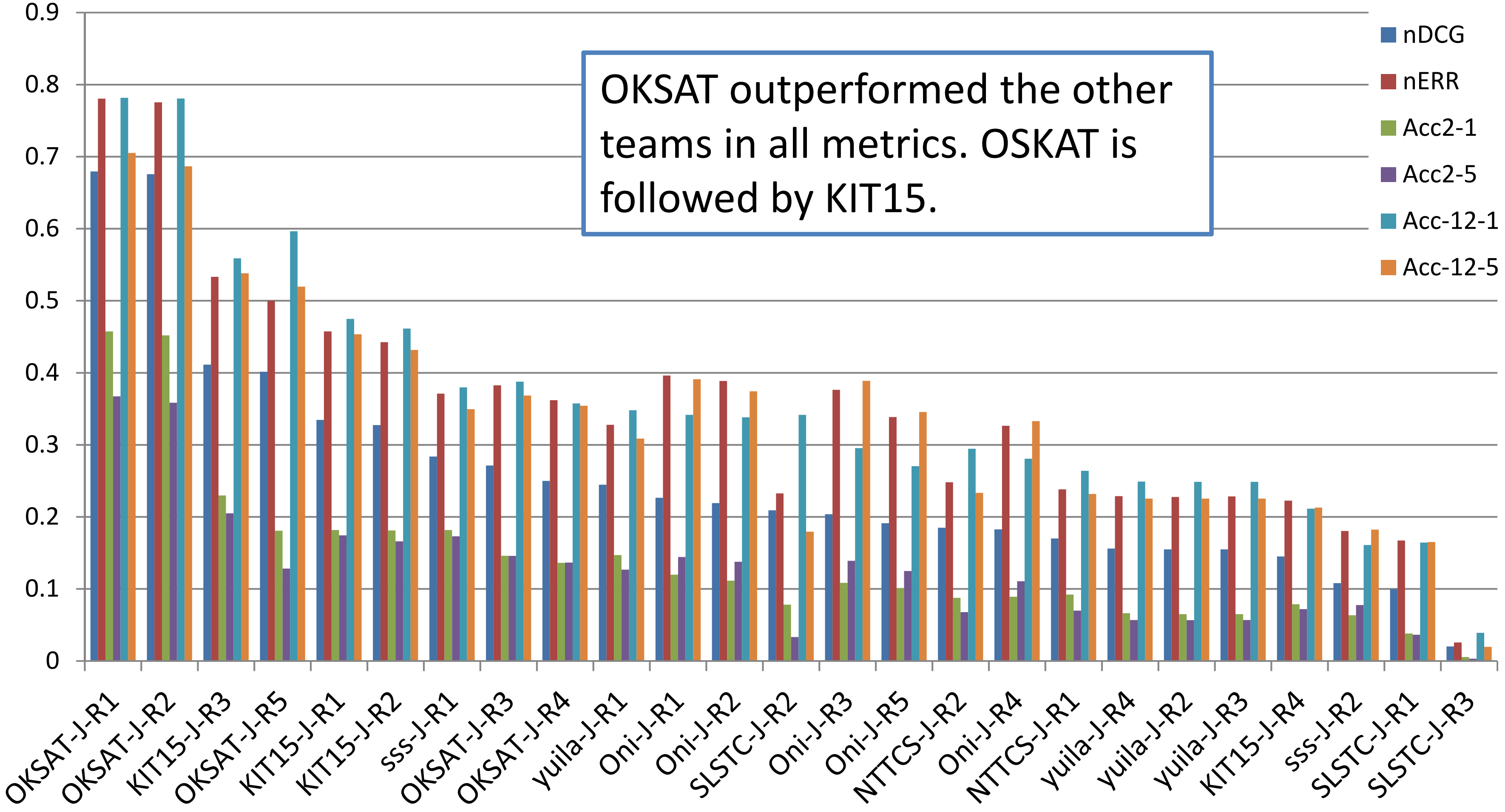
Group ID	Organization	#runs
KIT15	Kyoto Institute of Technology	4
NTTCS	NTT Communication Science Labs.	2
OKSAT	Osaka Kyoiku University	5
Oni	Osaka University	5
SLSTC	Waseda University	3
sss	University of Tokyo	2
yuila	Yamagata University	4

The Japanese Subtask

- Brief summary of the methods

Team	Methods
KIT	Semantic similarity using LDA
NTTCS	word2vec-based similarity with machine learning (DNN)
OKSAT	Rule-based method
Oni	Similarity (TFIDF, word2vec), machine learning (random forest), Weighted Text Matrix Factorization model
SLSTC	Learning using Error Back Propagation, graph-based model
sss	Machine learning (LSTM, kernel-based classifier)
yuila	Similarity (TFIDF)

The Japanese Subtask



OKSAT outperformed the other teams in all metrics. OSKAT is followed by KIT15.

Summary and Future work

- Filtering comments by using some manually designed rules was simple but effective.
- Representing a post (or comment) by the word2vec/topic models was helpful to perform semantic-level matching.
- Perform more analysis on the properties of post-comment pairs from the aspects of comment length, popularity, dialogue act, and sentiment to obtain more effective methods