Overview of the NTCIR-13 Short Text Conversation Task

Lifeng Shang
Noah's Ark Lab of Huawei, Hong Kong
Shang.Lifeng@huawei.com

Tetsuya Sakai
Waseda University, Tokyo, Japan
tetsuyasakai@acm.org

Hang Li
Noah's Ark Lab of Huawei, Hong Kong
Hang.Li@huawei.com

Ryuichiro Higashinaka
Nippon Telegraph and Telephone Corporation, Japan
higashinaka.ryuichiro@lab.ntt.co.jp

Yusuke Miyao
National Institute of Informatics, Japan
yusuke@nii.ac.jp

Yuki Arase
Osaka University, Japan
arase@ist.osaka-u.ac.jp

Masako Nomoto
Yahoo Japan Corporation, Japan
mnomoto@yahoo-corp.jp

ABSTRACT

We give an overview of the NII Testbeds and Community for Information access Research (NTCIR)-13 Short Text Conversation (STC) task, which was a core task of NTCIR-13. At NTCIR-12, STC was taken as an IR problem by maintaining a large repository of post-comment pairs then finding a clever method of reusing these existing comments to respond to new posts. At NTCIR-13, besides the *retrieval-based method*, we focused on a new method called *generation-based method* to generate "new" comments. The generation-based method has gained a great deal of attention in recent years, even though there the problem still remains of whether the retrieval-based method should be wholly replaced with or combined with the generation-based method for the STC task. By organizing this task at NTCIR-13, we provided a transparent platform to compare the two aforementioned methods by conducting comprehensive evaluations. For the Chinese subtask, there were a total of 34 registrations, and 22 teams finally submitted 120 runs. For the Japanese subtask, there were a total of 9 registrations, and 5 teams submitted 15 runs. In this paper, we review the task definition, evaluation measures, test collections, and evaluation results of all teams.

Keywords

artificial intelligence, dialogue systems, evaluation, information retrieval, deep learning, natural language processing, social media; test collections

1. INTRODUCTION

With the emergence of social media and the spread of mobile devices, conversation via short texts has become an important method of communication. This is why we proposed to organize a pilot task on conversation at NTCIR-12 to bring together researchers interested in natural language conversation. At NTCIR-12, STC consisted of two subtasks: one was a Chinese subtask by using the post-comment pairs crawled from Weibo, and the other was a Japanese subtask by providing the IDs of such pairs from Twitter [5]. At NTCIR-13, we had the same two subtasks, the main difference was the consideration of the *generation-based method*.

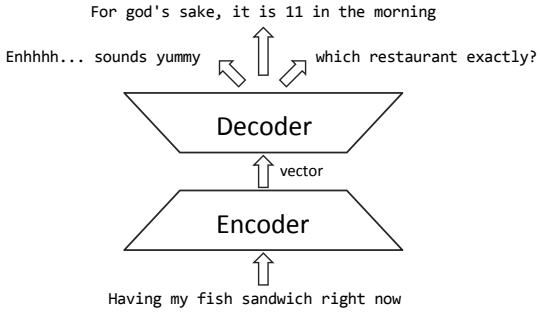

Figure 1: Concept of generation-based method involving RNN-based models

We still define short text conversation (STC) as a simplified version of natural language conversation: one round of conversation formed by two short texts, with the former being a message from a human and the latter being a comment to the message given by a computer. For the *retrieval-based method*, the basic idea is maintaining a large repository of STC data (i.e. post-comment pairs) and finding a clever method of retrieving related comments from the repository and return the most appropriate comment. A typical method of finding appropriate comments is designing various matching features (e.g. the workhorse BM25 and the recently proposed deep-matching models) then using machine leaning models to learn to combine these features. With this method, we can reuse the existing comment of repository as response to the current post.

There are two types of generation-based method for STC, 1) statistical machine translation (SMT) [1], and 2) recurrent neural network (RNN)-based models [4]. The SMT-based method treats comment generation as a translation problem, in which the model is trained on a parallel corpus of post-comment pairs. Most attention has recently been focused on the generation-based method that involves RNN-based models. Figure 1 graphically shows the basic idea of an RNN-based model. It first adaptively encodes

the input post into a fixed-length vector then feeds this representation to a decoder to generate comments word-by-word. The encoder mimics the language-understanding process of humans and the decoder acts as a language model that can sequentially generate words by taking into account the meaning of the post from the encoder. The building blocks of the encoder and decoder can be various NNs (e.g. RNN, convolutional NNs (CNNs), or recursive NNs); however, how to effectively design the structure of the encoder or decoder with such building blocks or invent new effective blocks is still being investigated. To advance the research on this topic, it is necessary to build a transparent platform to attract researchers with diverse research backgrounds (e.g. information retrieval (IR), natural language processing (NLP), and machine learning) to easily test their ideas. Other widely used traditional natural-language-generation (NLG) methods, such as template-filling-based, rule-based, and linguistic-based generators, are also encouraged.

The goal of the STC task is 1) to clarify the effectiveness and limitations of *retrieval-based* and *generation-based* methods used in this task, 2) to find an effective method of combining the two aforementioned methods, 3) to advance the research on automatic evaluation of natural language conversation, and 4) to stimulate research on more advanced methods for IR, NLP, and machine learning, especially on new neural models for conversation.

Thirty-seven teams registered to take part in the STC task, and we ultimately received 120 runs from 22 teams in the Chinese subtask and 15 runs from 5 teams in the Japanese subtask. The group name, organization, and number of runs submitted to the Chinese and Japanese subtasks are listed in Tables 1 and 11, respectively.

The remainder of this paper is organized as follows. In Section 2, we describe the Chinese subtask from the aspects of task definition, evaluation measures, dataset collection, and evaluation results. In Section 3, we describe the details of the Japanese subtask. In Section 4, we conclude the paper and mention future work.

Table 1: Organization and number of submitted runs of participating teams in STC Chinese subtask

Team ID	Organization	#runs
BUPTTeam	Beijing University of Posts and Telecommunications	2
Beihang	Beihang University	5
CIAL	Institute of Information Science, Academia Sinica	5
CYIII	Chaoyang University of Technology	6
DeepIntell	DeepIntell Co., Ltd	5
Gbot	Institute of Computing Technology, Chinese Academy of Sciences	6
ITNLP	Harbin Institute of Technology	3
MSRSC	Microsoft Research/University of Science and Technology of China	10
Nders	NetDragon Websoft, Inc/Minjiang University	5
PolyU	The Hong Kong Polytechnic University	6
SG01	Sogou, Inc/Tsinghua University	8
SLSTC	Waseda University	1
SMIPG	South China University of Technology	1
TUA1	Tokushima University	9
UB	University at Buffalo	5
WIDM	National Central University	4
WUST	Wuhan University of Science and Technology	2
ckip	Academia Sinica	4
iNLP	Alibaba Group/Onehome (Beijing) Network Technology Co. Ltd.	10
rucir	Renmin University of China	8
splab	Shanghai Jiao Tong University	5
srcb	Ricoh Software Research Center (Beijing) Co., Ltd.	10

2. CHINESE SUBTASK

2.1 Task Definition

For the retrieval-based method, the task definition was the same as that for NTCIR-12. For the generation-base method, we also provided the same repository of post-comment pairs in advance to the participants. During the training period, generation models can be learned from this repository, and during the evaluation period, the results from all participating teams are pooled and labeled by humans. Graded relevance IR measures are used for evaluation. The main difference is in the design of criteria for assessing relevance; we need to consider extra facets for generation-based subtasks, e.g. fluency and grammatical correctness.

2.2 Evaluation Measures

Following the NTCIR-12 STC-1 Chinese subtask, we used three evaluation measures: *nG@1* (normalized gain at cutoff 1), *P+*, and *nERR@10* (normalized expected reciprocal rank at cutoff 10) [5].

As described in Section 2.3, we obtained three independent labels for each returned string (either a retrieved comment or generated string); each label is either 2 (*fluent*, *coherent*, *self-sufficient*, and *substantial*), 1 (*fluent* and *coherent* but not *self-sufficient* and/or not *substantial*), or 0 (not *self-sufficient* and/or not *substantial*). By summing the labels of the three assessors per returned string, we obtained our gold standard data with $0, 1, \ldots, 6$ as grades. Our *official* evaluation results treat the above grades as *gain values* for computing the three measures. The evaluation script `NTCIREVAL`[1] was used with the option `-g 1:2:3:4:5:6`. Note that in the NTCIR-12 STC-1 Chinese subtask, the final grades were $0, 1, and\ 2$, and that the corresponding gain values were $0, 1, and\ 3$ (exponential gain-value setting, with the `NTCIREVAL` option `-g 1:3`).

We also obtained an additional set of results using a different gain-value setting by applying the *unanimity-aware gain* approach of Sakai [3]. Instead of using the sum of labels as is, this method takes into account whether different assessors agreed with one another. To be more specific, let N be the number of independent assessors (3 in our case), D_{max} be the highest possible label on an interval scale (2 in our

[1]Available in the `NTCIREVAL` package: http://research.nii.ac.jp/ntcir/tools/ntcireval-en.html.

Table 2: Raw gain (for official results) vs. unanimity-aware gain ($p = 0.2, N = 3, D_{max} = 2$)

labels			*rawG*	*D*	$pN(D_{max} - D)$	*UnanG*
2	2	2	**6**	0	1.2	**7.2**
1	2	2	**5**	1	0.6	**5.6**
1	1	2	**4**	1	0.6	**4.6**
0	2	2	**4**	2	0	**4**
1	1	1	**3**	0	1.2	**4.2**
0	1	2	**3**	2	0	**3**
0	1	1	**2**	1	0.6	**2.6**
0	0	2	**2**	2	0	**2**
0	0	1	**1**	1	0.6	**1.6**

Table 3: Statistics of dataset for Chinese subtask

Repository	No. of posts	219,174
	No. of comments	4,305,706
	No. of original pairs	4,433,949
Labeled Data	No. of posts	769
	No. of comments	11,535
	No. of labeled pairs	11,535
Test Data	No. of query posts	100

case), D be the difference between the highest and lowest rating for a particular string, and *RawG* be the sum of the labels for that string. Then given a parameter $p (0 \leq p \leq 1)$, unanimity-aware gain is given by

$$UnanG = RawG + pN(D_{max} - D) \quad (1)$$

if $RawG > 0$; otherwise $UnanG = RawG = 0$. For example, if all N assessors are in complete agreement (i.e., $D = 0$), then unanimity-aware gain adds an extra pND_{max} to the raw gain. That is, we assume that pN "virtual" assessors gave the string the highest possible rating. We let $p = 0.2$, although this is an arbitrary choice. Table 2 shows what this means in our experimental setting. Unlike the raw gain, the unanimity-aware gain rates the labels $(1, 1, 2)$ higher than $(0, 2, 2)$; $(1, 1, 1)$ higher than $(0, 1, 2)$ and even $(0, 2, 2)$; and $(0, 1, 1)$ higher than $(0, 0, 2)$. See Sakai [3] for more details on unanimity-aware gain.

2.3 Chinese Test Collection

2.3.1 Weibo Corpus

We used post-comment pairs from Weibo for the Chinese subtask. To construct the million-scale repository for the Chinese subtask, we randomly selected half the post-comment pairs from the repository used at NTCIR-12 then strictly followed the method described in [6] to construct the other new half.

Table 3 lists the statistics of the retrieval repository, labeled data, and query posts that we provided in the task. We collected 219,174 Weibo posts and the 4,305,706 corresponding comments and finally obtained 4,433,949 post-comment pairs. Each post had 20 different comments on average, and one comment can be used to respond to multiple posts.

2.3.2 Training Data

We also manually labeled 769 query posts, each of which had about 15 candidate comments. Note that for each selected (query) post, the labeled comments were originally posted in response to posts other than the query post. Finally, we labeled the 11,535 comments as "suitable", "neutral", and "unsuitable". The details of the labeling criteria are given in the following section 2.3.4.

2.3.3 Test Data

We carefully selected the test query posts to make the task adequate, balanced, and sufficiently challenging. For each method (i.e. retrieval-based or generation-based), a participating team could submit up to five runs. In each run, a ranking list of ten comments for each test query was requested. The participants were also encouraged to rank their submitted runs by preference.

- For comparison, at least one compulsory run that did not use any external evidence was also requested. External evidence means evidence beyond the given dataset. For instance, this includes other data or information from Weibo, as well as other corpora, e.g., HowNet or the web.

- Beyond this, the participants were at liberty to submit manual, external runs, which could be useful to improve the quality of the test collections.

2.3.4 Relevance Assessments

We used conventional IR-evaluation methodology. All the results (either retrieved or generated) from participants were pooled using the `NTCIRPOOL` tool[2], and the returned comments were judged manually. Three assessors were instructed to imagine that they were the authors of the original posts and to judge whether a comment is appropriate for an input post. The assessors had to choose from three relevance levels L0, L1, and L2, as defined below.

To make the annotation task operable, the appropriateness of retrieved or generated comments is judged from the following four criteria:

(1) **Fluent**: the comment is acceptable as a natural language text;

(2) **Coherent**: the comment should be logically connected and topically relevant to the original post (i.e. the comment makes sense in the eye of the originator of the post);

(3) **Self-sufficient**: the assessor can judge that the comment is appropriate by reading nothing other than the post-comment pair;

(4) **Substantial**: the comment provides new information in the eye of the originator of the post;

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". Our labeling procedure can also be concisely described by the pseudocode shown in Figure 3.

Figure 2 shows an example of the labeling results of a post and its comments. The first two comments are labeled "L0" because of the logic consistency and semantic relevance errors (i.e. coherent criterion). Comment 3 just repeats the same opinion as presented in the post, but it was still a comment that the author of the post wanted to see. Comment 4 depends on the scenario (i.e., the current score is

[2]http://research.nii.ac.jp/ntcir/tools/ntcirpool-en.html

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
Comment 1	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	(2) Coherent	L0
Comment 2	意大利的食物太美味了 Italian food is absolutely delicious.	(2) Coherent	L0
Comment 3	太夸张了吧！ Unbelievable!	(4) Substantial	L1
Comment 4	哈哈哈仍然是0：0 。还没看到进球。 Haha, it is still 0:0, no goal so far.	(3) Self-sufficient	L1
Comment 5	这正是意大利式防守足球。 This is exactly the Italian defending style football game	———	L2

Figure 2: Example post and its five candidate comments with human annotation. Content of post implies that football match had already started, while author of Comment 1 was still waiting for the match to start. Comment 2 talked about food of Italy. Comment 3 was a widely used response, but was appropriate for this post. Comment 4 stated that current score was still 0:0 and was appropriate comment only for this specific scenario.

0:0) or lacked enough context information, and was therefore labeled as "(+1)". Comment 5 is coherent to the post and provided some new useful information to the author of the post, so it is labeled "(+2)".

```
IF (fluent AND coherent)
    IF (self-sufficient AND substantial)
        assign L2
    ELSE
        assign L1
ELSE
    assign L0.
```

Figure 3: Pseudocode of labeling procedure for Chinese subtask of STC-2

Compared to the evaluation method at STC@NTCIR-12, the main difference is in the four criteria: (a) we merged the two criteria "(1) Coherent" and "(2) Topically relevant" at NTCIR-12 into one criterion "(2) Coherent" at NTCIR-13, since topical relevance is already a necessary condition for coherence, (b) we added a new fluency criterion, because the generation-based method may have fluency and grammar problems. As at NTCIR-12, all the submitted comments (no matter generated or retrieved) from all the participants were pooled to perform manual evaluation.

2.4 Chinese Run Results

Table 4 shows the run statistics of the STC2 Chinese subtask: we received a total of 64 retrieval-based runs (R-runs) and 56 generation-based runs (G-runs). Brief descriptions of the R-runs and G-runs are respectively given in Tables 18 and 19 in the Appendix.

Tables 7 and 8 show the mean official/unanimity-aware nG@1, P+, and nERR@10 results. Only the top 90 runs according to each evaluation measure are shown.

Tables 9 and 10 summarize the statistical significance test results. One best run was selected from each team based on a particular evaluation measure, then a randomized Tukey HSD test [2] with $B = 10,000$ trials using the `Discpower` toolkit[3] was conducted to compare every pair of teams at the significance criterion $\alpha = 0.05$. The differences across the two tables are indicated in bold.

From the official results with nG@1, it can be observed that:

Table 4: STC-2 Chinese run statistics (R-runs: retrieval-based runs; G-runs: generation-based runs).

Team	R-runs	G-runs	total
BUPTTeam	0	2	2
Beihang	5	0	5
CIAL	4	1	5
CYIII	1	5	6
DeepIntell	5	0	5
Gbot	1	5	6
ITNLP	3	0	3
MSRSC	5	5	10
Nders	5	0	5
PolyU	4	2	6
SG01	3	5	8
SLSTC	1	0	1
SMIPG	0	1	1
TUA1	4	5	9
UB	5	0	5
WIDM	3	1	4
WUST	2	0	2
ckip	0	4	4
iNLP	5	5	10
rucir	3	5	8
splab	0	5	5
srcb	5	5	10
	64	56	120

[3] http://research.nii.ac.jp/ntcir/tools/discpower-en.html

Table 5: Randomized Tukey HSD test p-values: differences between official and unanimity-aware results

(a) nG@1		
Run pair	Official	Unanimity
rucir-C-R2 > MSRSC-C-R4	0.0738	0.0491
SMIPG-C-G1 > WUST-C-R2	0.0705	0.0272
BUPTTeam-C-G1 > ckip-C-G3	0.0619	0.0419
(b) P+		
MSRSC-C-R4 > WUST-C-R2	0.0607	0.0468

Table 6: Kendall's τ values with 95% confidence intervals (120 STC-2 Chinese runs).

(a) Official results	
Mean nG@1 vs. P+	0.903 [0.879, 0.930]
Mean nG@1 vs. nERR@10	0.898 [0.875, 0.924]
P+ vs nERR@10	0.955 [0.937, 0.973]
(b) Unanimity-Aware results	
Mean nG@1 vs. P+	0.901 [0.877, 0.928]
Mean nG@1 vs. nERR@10	0.894 [0.869, 0.922]
P+ vs nERR@10	0.956 [0.937, 0.976]
(c) Official vs. Unanimity	
Mean nG@1	0.985 [0.977, 0.993]
P+	0.990 [0.985, 0.997]
nERR@10	0.987 [0.980, 0.995]

- SG01 was the top performing team, in that it statistically significantly outperformed 13 other teams.
- The second best teams were sblab, Beihang, Nders, and srcb, which statistically significantly outperformed 9 other teams.
- The third best teams were DeepIntell, iNLP, CYIII, TUA1, UB, WIDM, and Gbot, which statistically significantly outperformed 8 other teams.

From the official results with P+, it can be observed that:

- SG01 was the top performing team, in that it statistically significantly outperformed 13 other teams;
- The second best teams were splab, Beihang, DeepIntell, Nder, srcb, iNLP, and CYIII, which statistically significantly outperformed 9 other teams;
- The third best teams were UB, TUA, WIDM, and rucir, which statistically significantly outperformed 8 other teams.

Similarly, from the official results with nERR@10, it can be observed that:

- SG01 was the top performing team, in that it statistically significantly outperformed 12 other teams;
- The second best teams were splab, Beihang, DeepIntell, Nders, srcb, iNLP, CYIII, TUA1, and UB, which statistically significantly outperformed 9 other teams. The third best teams were WIDM, rucir, and Gbot, which statistically significantly outperformed 8 other teams.

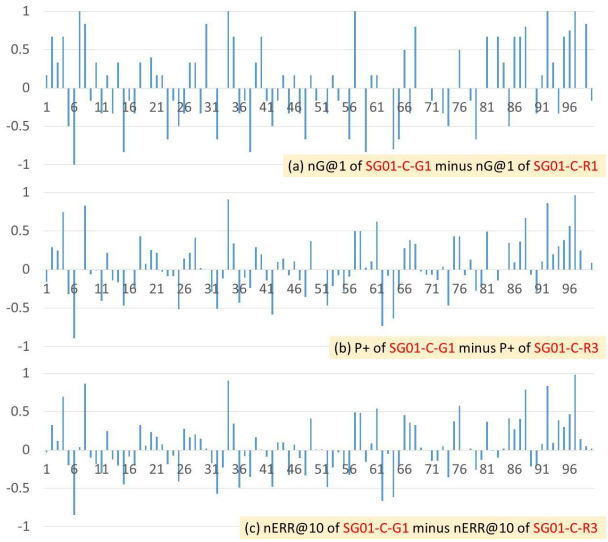

Figure 4: Per-topic comparison of best G-run and best R-run (official results)

Table 5 shows the details of the disagreements between the official and unanimity-aware results in terms of the randomized Tukey HSD test, which are indicated in bold in Tables 9 and 10. For example, while rucir-C-R2 outperformed MSRSC-C-R4 in terms of mean nG@1 (for both official and unanimity-aware gain values), the difference is not statistically significant with the official results ($p = 0.0738$), while statistically significant with the unanimity-aware results ($p = 0.0491$). These results indicate that the unanimity-aware gains can affect research conclusions to some extent.

Table 6 compares different run rankings in terms of Kendall's τ: all 120 runs are considered. It can be observed that P+ and nERR@10 produce very similar (but not identical) rankings, and that official and unanimity-aware gain values produce very similar (but not identical) rankings.

One notable result in Table 7 is that the best G-run `SG01-C-G1` outperformed the best R-runs (`SG01-C-R1` for Mean nG@1 and `SG01-C-R3` for Mean P+ and Mean nERR@10) on average. We conducted a randomization test, again using `Discpower` with $B = 10,000$ trials to investigate whether the differences are "real." The p-values for the difference between the best G-run and best R-run in terms of the three measures were 0.3051, 0.2138, *and*0.2448: thus, the differences are not statistically significant. The corresponding effect sizes (i.e., standardized mean differences) [2] were 0.1039, 0.1319, *and*0.1239, indicating small effects.

Figure 4 illustrates the per-topic-score differences between the best G-run and best R-run for each evaluation measure. The bars above the horizontal axis represent topics for which the G-run outperformed the R-run; those below the horizontal axis represent topics for which the R-run outperformed the G-run. This figure and aforementioned statistical-significance test results suggest that it is too early to conclude that "generation-based runs are now better than retrieval-based runs."

3. JAPANESE SUBTASK

Table 7: STC-2 Chinese official results (top 90 runs only)

Run	Mean nG@1	Run	Mean P+	Run	Mean nERR@10
SG01-C-G1	0.5867	SG01-C-G1	0.6670	SG01-C-G1	0.7095
SG01-C-G3	0.5633	SG01-C-G3	0.6567	SG01-C-G3	0.6947
SG01-C-G2	0.5483	SG01-C-G2	0.6335	SG01-C-G2	0.6783
SG01-C-R1	0.5355	SG01-C-R3	0.6200	SG01-C-R3	0.6663
SG01-C-R2	0.5168	SG01-C-R1	0.6084	SG01-C-R1	0.6579
splab-C-G4	0.5080	splab-C-G4	0.6080	splab-C-G4	0.6492
SG01-C-R3	0.5048	SG01-C-R2	0.5944	SG01-C-R2	0.6461
Beihang-C-R4	0.4980	Beihang-C-R4	0.5818	splab-C-G1	0.6282
splab-C-G1	0.4848	splab-C-G1	0.5768	splab-C-G5	0.6175
Nders-C-R4	0.4780	splab-C-G5	0.5657	SG01-C-G4	0.6129
Nders-C-R2	0.4743	DeepIntell-C-R1	0.5564	Beihang-C-R4	0.6105
Nders-C-R3	0.4647	SG01-C-G4	0.5545	DeepIntell-C-R1	0.5994
Nders-C-R1	0.4593	Beihang-C-R2	0.5510	splab-C-G3	0.5966
Nders-C-R5	0.4550	Nders-C-R2	0.5497	Nders-C-R2	0.5882
Beihang-C-R2	0.4510	Nders-C-R5	0.5495	Nders-C-R5	0.5868
srcb-C-R5	0.4500	splab-C-G3	0.5451	Nders-C-R4	0.5809
SG01-C-G4	0.4483	Beihang-C-R1	0.5441	Nders-C-R1	0.5805
splab-C-G5	0.4472	srcb-C-R1	0.5395	srcb-C-G2	0.5781
splab-C-G3	0.4420	Nders-C-R1	0.5394	DeepIntell-C-R4	0.5774
srcb-C-R1	0.4343	iNLP-C-R1	0.5375	Nders-C-R3	0.5768
Beihang-C-R1	0.4343	srcb-C-R5	0.5367	srcb-C-G3	0.5737
DeepIntell-C-R1	0.4323	Nders-C-R4	0.5338	srcb-C-R1	0.5736
CYIII-C-R1	0.4262	CYIII-C-R1	0.5332	Beihang-C-R2	0.5716
TUA1-C-R4	0.4210	iNLP-C-R2	0.5324	DeepIntell-C-R2	0.5678
srcb-C-G2	0.4138	Nders-C-R3	0.5317	iNLP-C-R1	0.5674
iNLP-C-R1	0.4132	DeepIntell-C-R4	0.5270	CYIII-C-R1	0.5668
srcb-C-G3	0.4103	srcb-C-G3	0.5269	iNLP-C-R2	0.5667
UB-C-R1	0.4103	Beihang-C-R3	0.5268	srcb-C-R5	0.5644
splab-C-G2	0.4080	DeepIntell-C-R2	0.5258	Beihang-C-R1	0.5643
Beihang-C-R3	0.4080	Beihang-C-R5	0.5215	Beihang-C-R3	0.5623
DeepIntell-C-R4	0.4077	srcb-C-G2	0.5188	SG01-C-G5	0.5596
UB-C-R2	0.4060	UB-C-R4	0.5106	Beihang-C-R5	0.5544
iNLP-C-R2	0.4055	UB-C-R2	0.5105	TUA1-C-R4	0.5524
UB-C-R4	0.3978	UB-C-R1	0.5104	UB-C-R2	0.5484
srcb-C-R2	0.3972	DeepIntell-C-R3	0.5082	DeepIntell-C-R3	0.5484
Beihang-C-R5	0.3937	SG01-C-G5	0.5068	UB-C-R4	0.5473
DeepIntell-C-R2	0.3923	srcb-C-R2	0.5030	UB-C-R1	0.5445
TUA1-C-G4	0.3893	iNLP-C-R4	0.5025	iNLP-C-R4	0.5408
UB-C-R5	0.3858	DeepIntell-C-R5	0.5023	srcb-C-R2	0.5368
srcb-C-R3	0.3852	UB-C-R3	0.4980	DeepIntell-C-R5	0.5360
SG01-C-G5	0.3820	srcb-C-R3	0.4964	UB-C-R5	0.5334
UB-C-R3	0.3792	TUA1-C-R4	0.4952	UB-C-R3	0.5314
iNLP-C-R4	0.3790	WIDM-C-R1	0.4950	TUA1-C-R2	0.5298
DeepIntell-C-R3	0.3773	UB-C-R5	0.4932	srcb-C-R3	0.5272
TUA1-C-R2	0.3697	TUA1-C-R2	0.4913	iNLP-C-R3	0.5264
iNLP-C-R3	0.3695	TUA1-C-G4	0.4909	srcb-C-G4	0.5241
srcb-C-G4	0.3657	iNLP-C-R3	0.4899	WIDM-C-R1	0.5238
WIDM-C-R1	0.3620	splab-C-G2	0.4844	TUA1-C-G4	0.5227
Gbot-C-G4	0.3610	srcb-C-G4	0.4838	splab-C-G2	0.5147
DeepIntell-C-R5	0.3523	rucir-C-R2	0.4675	rucir-C-R2	0.5064
rucir-C-R2	0.3453	TUA1-C-R3	0.4662	srcb-C-G1	0.4997
TUA1-C-G1	0.3353	srcb-C-G1	0.4582	TUA1-C-R3	0.4963
WIDM-C-R2	0.3290	Gbot-C-G4	0.4512	WIDM-C-R2	0.4863
TUA1-C-G3	0.3187	WIDM-C-R2	0.4440	Gbot-C-G4	0.4840
TUA1-C-R3	0.3177	srcb-C-G5	0.4376	srcb-C-G5	0.4735
srcb-C-G1	0.3160	TUA1-C-G1	0.4326	TUA1-C-G1	0.4725
srcb-C-G5	0.3052	srcb-C-R4	0.4306	WIDM-C-R3	0.4707
Gbot-C-G3	0.3033	WIDM-C-R3	0.4304	srcb-C-R4	0.4688
TUA1-C-G5	0.3023	TUA1-C-G5	0.4239	TUA1-C-G5	0.4634
srcb-C-R4	0.2983	TUA1-C-G3	0.4216	TUA1-C-G3	0.4546
WIDM-C-R3	0.2983	Gbot-C-G3	0.4127	TUA1-C-R1	0.4479
Gbot-C-G2	0.2783	TUA1-C-R1	0.4060	Gbot-C-G3	0.4455
TUA1-C-R1	0.2653	rucir-C-R3	0.3924	rucir-C-R3	0.4272
SMIPG-C-G1	0.2637	Gbot-C-G2	0.3874	TUA1-C-G2	0.4125
rucir-C-R3	0.2617	rucir-C-R1	0.3802	rucir-C-R1	0.4079
rucir-C-R1	0.2567	TUA1-C-G2	0.3687	Gbot-C-G2	0.4064
rucir-C-G3	0.2543	iNLP-C-G2	0.3579	iNLP-C-G2	0.3911
iNLP-C-G4	0.2477	SMIPG-C-G1	0.3568	iNLP-C-G4	0.3839
TUA1-C-G2	0.2443	iNLP-C-G4	0.3490	rucir-C-G3	0.3828
iNLP-C-G2	0.2323	rucir-C-G3	0.3461	iNLP-C-G1	0.3732
iNLP-C-G1	0.2320	iNLP-C-G5	0.3414	SMIPG-C-G1	0.3721
iNLP-C-G5	0.2257	iNLP-C-G1	0.3411	iNLP-C-G3	0.3672
iNLP-C-G3	0.2227	iNLP-C-G3	0.3344	iNLP-C-G5	0.3654
iNLP-C-R5	0.2187	iNLP-C-R5	0.3142	iNLP-C-R5	0.3291
Gbot-C-G1	0.2073	Gbot-C-G1	0.3017	MSRSC-C-R4	0.3104
BUPTTeam-C-G1	0.1823	MSRSC-C-R4	0.2982	Gbot-C-G1	0.3052
MSRSC-C-R4	0.1767	BUPTTeam-C-G1	0.2755	BUPTTeam-C-G1	0.2746
MSRSC-C-R5	0.1517	MSRSC-C-R2	0.2498	MSRSC-C-R2	0.2611
WIDM-C-G1	0.1437	WIDM-C-G1	0.2311	PolyU-C-R2	0.2387
PolyU-C-G1	0.1342	MSRSC-C-R3	0.2274	MSRSC-C-R3	0.2378
MSRSC-C-R2	0.1300	MSRSC-C-R5	0.2263	MSRSC-C-R1	0.2208
CYIII-C-G1	0.1213	PolyU-C-R2	0.2253	MSRSC-C-R5	0.2202
PolyU-C-R3	0.1190	MSRSC-C-R1	0.2207	MSRSC-C-G4	0.2174
ITNLP-C-R3	0.1167	MSRSC-C-G4	0.2168	PolyU-C-R3	0.2164
CYIII-C-G2	0.1163	PolyU-C-R3	0.2117	WIDM-C-G1	0.2034
MSRSC-C-R1	0.1140	BUPTTeam-C-G2	0.1985	BUPTTeam-C-G2	0.2001
MSRSC-C-G2	0.1133	MSRSC-C-G2	0.1736	PolyU-C-R1	0.1776
MSRSC-C-G1	0.1133	MSRSC-C-G1	0.1720	CYIII-C-G1	0.1771
MSRSC-C-R3	0.1087	rucir-C-G1	0.1712	CYIII-C-G2	0.1662
PolyU-C-R2	0.1077	MSRSC-C-G5	0.1693	MSRSC-C-G2	0.1659

Table 8: STC-2 Chinese unanimity-aware ($p = 0.2$) results (top 90 runs only)

Run	Mean nG@1	Run	Mean P+	Run	Mean nERR@10
SG01-C-G1	0.5841	SG01-C-G1	0.6580	SG01-C-G1	0.7130
SG01-C-G3	0.5630	SG01-C-G3	0.6495	SG01-C-G3	0.6993
SG01-C-G2	0.5472	SG01-C-G2	0.6253	SG01-C-G2	0.6839
SG01-C-R1	0.5316	SG01-C-R3	0.6158	SG01-C-R3	0.6741
SG01-C-R2	0.5157	SG01-C-R1	0.6019	SG01-C-R1	0.6623
SG01-C-R3	0.5074	splab-C-G4	0.6016	splab-C-G4	0.6579
splab-C-G4	0.5062	SG01-C-R2	0.5895	SG01-C-R2	0.6528
Beihang-C-R4	0.4993	splab-C-G1	0.5763	splab-C-G1	0.6416
splab-C-G1	0.4903	Beihang-C-R4	0.5761	splab-C-G5	0.6278
Nders-C-R4	0.4810	splab-C-G5	0.5634	SG01-C-G4	0.6265
Nders-C-R2	0.4772	DeepIntell-C-R1	0.5568	Beihang-C-R4	0.6188
Nders-C-R3	0.4673	SG01-C-G4	0.5549	DeepIntell-C-R1	0.6108
Nders-C-R1	0.4613	Beihang-C-R2	0.5470	splab-C-G3	0.6084
Nders-C-R5	0.4596	Nders-C-R5	0.5469	Nders-C-R5	0.5978
SG01-C-G4	0.4588	splab-C-G3	0.5461	Nders-C-R2	0.5969
srcb-C-R5	0.4536	Nders-C-R2	0.5446	DeepIntell-C-R4	0.5907
splab-C-G5	0.4536	iNLP-C-R1	0.5405	srcb-C-G2	0.5902
Beihang-C-R2	0.4514	Beihang-C-R1	0.5405	Nders-C-R4	0.5900
splab-C-G3	0.4481	iNLP-C-R2	0.5370	Nders-C-R1	0.5897
srcb-C-R1	0.4375	srcb-C-R5	0.5358	srcb-C-G3	0.5876
DeepIntell-C-R1	0.4368	Nders-C-R1	0.5351	iNLP-C-R1	0.5862
Beihang-C-R1	0.4359	srcb-C-R1	0.5345	iNLP-C-R2	0.5859
iNLP-C-R1	0.4283	Nders-C-R4	0.5304	Nders-C-R3	0.5857
CYIII-C-R1	0.4263	Nders-C-R3	0.5287	srcb-C-R1	0.5824
srcb-C-G2	0.4224	CYIII-C-R1	0.5284	Beihang-C-R2	0.5793
iNLP-C-R2	0.4207	srcb-C-G3	0.5282	DeepIntell-C-R2	0.5773
srcb-C-G3	0.4194	DeepIntell-C-R4	0.5272	srcb-C-R5	0.5750
TUA1-C-R4	0.4192	DeepIntell-C-R2	0.5247	CYIII-C-R1	0.5750
splab-C-G2	0.4181	Beihang-C-R3	0.5234	Beihang-C-R1	0.5727
DeepIntell-C-R4	0.4156	srcb-C-G2	0.5193	Beihang-C-R3	0.5705
Beihang-C-R3	0.4124	Beihang-C-R5	0.5193	SG01-C-G5	0.5687
UB-C-R1	0.4090	UB-C-R4	0.5090	Beihang-C-R5	0.5646
UB-C-R2	0.4075	UB-C-R2	0.5065	TUA1-C-R4	0.5582
UB-C-R4	0.4041	DeepIntell-C-R3	0.5061	DeepIntell-C-R3	0.5571
TUA1-C-G4	0.3994	UB-C-R1	0.5050	UB-C-R4	0.5567
srcb-C-R2	0.3992	SG01-C-G5	0.5042	UB-C-R2	0.5553
Beihang-C-R5	0.3992	iNLP-C-R4	0.5038	iNLP-C-R4	0.5526
DeepIntell-C-R2	0.3952	DeepIntell-C-R5	0.5018	UB-C-R1	0.5498
srcb-C-R3	0.3913	srcb-C-R2	0.4993	DeepIntell-C-R5	0.5494
UB-C-R5	0.3904	WIDM-C-R1	0.4960	srcb-C-R2	0.5453
SG01-C-G5	0.3873	srcb-C-R3	0.4955	UB-C-R5	0.5424
iNLP-C-R4	0.3865	UB-C-R3	0.4952	TUA1-C-R2	0.5420
UB-C-R3	0.3840	TUA1-C-R4	0.4919	UB-C-R3	0.5403
DeepIntell-C-R3	0.3785	TUA1-C-R2	0.4916	srcb-C-R3	0.5402
iNLP-C-R3	0.3768	TUA1-C-G4	0.4914	TUA1-C-G4	0.5395
srcb-C-G4	0.3746	UB-C-R5	0.4911	iNLP-C-R3	0.5384
TUA1-C-R2	0.3744	iNLP-C-R3	0.4892	srcb-C-G4	0.5379
WIDM-C-R1	0.3704	splab-C-G2	0.4877	WIDM-C-R1	0.5355
Gbot-C-G4	0.3683	srcb-C-G4	0.4859	splab-C-G2	0.5335
DeepIntell-C-R5	0.3608	TUA1-C-R3	0.4666	rucir-C-R2	0.5188
rucir-C-R2	0.3531	rucir-C-R2	0.4663	srcb-C-G1	0.5102
TUA1-C-G1	0.3440	srcb-C-G1	0.4575	TUA1-C-R3	0.5078
WIDM-C-R2	0.3355	Gbot-C-G4	0.4520	Gbot-C-G4	0.5014
TUA1-C-G3	0.3276	WIDM-C-R2	0.4444	WIDM-C-R2	0.4987
TUA1-C-R3	0.3247	srcb-C-G5	0.4400	srcb-C-G5	0.4905
srcb-C-G1	0.3202	TUA1-C-G1	0.4339	TUA1-C-G1	0.4882
srcb-C-G5	0.3159	WIDM-C-R3	0.4306	WIDM-C-R3	0.4826
Gbot-C-G3	0.3135	srcb-C-R4	0.4287	TUA1-C-G5	0.4809
TUA1-C-G5	0.3110	TUA1-C-G5	0.4253	srcb-C-R4	0.4772
WIDM-C-R3	0.3041	TUA1-C-G3	0.4232	TUA1-C-G3	0.4711
srcb-C-R4	0.3018	Gbot-C-G3	0.4156	Gbot-C-G3	0.4637
Gbot-C-G2	0.2843	TUA1-C-R1	0.4058	TUA1-C-R1	0.4585
SMIPG-C-G1	0.2753	rucir-C-R3	0.3944	rucir-C-R3	0.4416
rucir-C-R3	0.2700	Gbot-C-G2	0.3860	TUA1-C-G2	0.4313
TUA1-C-R1	0.2690	rucir-C-R1	0.3823	rucir-C-R1	0.4227
rucir-C-G3	0.2680	TUA1-C-G2	0.3731	Gbot-C-G2	0.4199
rucir-C-R1	0.2650	iNLP-C-G2	0.3625	iNLP-C-G2	0.4089
iNLP-C-G4	0.2631	SMIPG-C-G1	0.3624	iNLP-C-G4	0.4042
TUA1-C-G2	0.2553	iNLP-C-G4	0.3545	rucir-C-G3	0.4040
iNLP-C-G1	0.2454	rucir-C-G3	0.3528	SMIPG-C-G1	0.3933
iNLP-C-G2	0.2443	iNLP-C-G1	0.3459	iNLP-C-G1	0.3930
iNLP-C-G3	0.2339	iNLP-C-G5	0.3444	iNLP-C-G3	0.3851
iNLP-C-G5	0.2294	iNLP-C-G3	0.3386	iNLP-C-G5	0.3806
iNLP-C-R5	0.2242	iNLP-C-R5	0.3172	iNLP-C-R5	0.3441
Gbot-C-G1	0.2193	Gbot-C-G1	0.3081	Gbot-C-G1	0.3277
BUPTTeam-C-G1	0.1866	MSRSC-C-R4	0.3002	MSRSC-C-R4	0.3241
MSRSC-C-R4	0.1822	BUPTTeam-C-G1	0.2757	BUPTTeam-C-G1	0.2848
MSRSC-C-R5	0.1547	MSRSC-C-R2	0.2540	MSRSC-C-R2	0.2737
WIDM-C-G1	0.1526	WIDM-C-G1	0.2376	PolyU-C-R2	0.2516
PolyU-C-G1	0.1407	MSRSC-C-R3	0.2316	MSRSC-C-R3	0.2506
MSRSC-C-R2	0.1358	PolyU-C-R2	0.2303	MSRSC-C-R1	0.2324
CYIII-C-G1	0.1293	MSRSC-C-R5	0.2290	MSRSC-C-G4	0.2314
PolyU-C-R3	0.1255	MSRSC-C-R1	0.2231	MSRSC-C-R5	0.2311
CYIII-C-G2	0.1235	MSRSC-C-G4	0.2208	PolyU-C-R3	0.2289
MSRSC-C-G2	0.1203	PolyU-C-R3	0.2153	WIDM-C-G1	0.2205
MSRSC-C-G1	0.1203	BUPTTeam-C-G2	0.2024	BUPTTeam-C-G2	0.2131
MSRSC-C-R1	0.1174	MSRSC-C-G2	0.1788	PolyU-C-R1	0.1921
ITNLP-C-R3	0.1172	MSRSC-C-G1	0.1764	CYIII-C-G1	0.1909
MSRSC-C-R3	0.1167	rucir-C-G1	0.1758	MSRSC-C-G2	0.1796
CYIII-C-G4	0.1162	MSRSC-C-G5	0.1748	CYIII-C-G2	0.1778

Table 9: Statistical significance with best run from each team according to official STC-2 Chinese performances (randomized Tukey HSD test, $B = 10,000, \alpha = 0.05$).

These runs are	significantly better than these runs in terms of mean official nG@1
SG01-C-G1	UB-C-R1,WIDM-C-R1,Gbot-C-G4,rucir-C-R2,SMIPG-C-G1,BUPTTeam-C-G1,MSRSC-C-R4,PolyU-C-G1,ITNLP-C-R3, WUST-C-R2,SLSTC-C-R1,CIAL-C-G1,ckip-C-G3
splab-C-G4	SMIPG-C-G1,BUPTTeam-C-G1,MSRSC-C-R4,PolyU-C-G1,ITNLP-C-R3,WUST-C-R2,SLSTC-C-R1,CIAL-C-G1,ckip-C-G3
Beihang-C-R4	SMIPG-C-G1,BUPTTeam-C-G1,MSRSC-C-R4,PolyU-C-G1,ITNLP-C-R3,WUST-C-R2,SLSTC-C-R1,CIAL-C-G1,ckip-C-G3
Nders-C-R4	SMIPG-C-G1,BUPTTeam-C-G1,MSRSC-C-R4,PolyU-C-G1,ITNLP-C-R3,WUST-C-R2,SLSTC-C-R1,CIAL-C-G1,ckip-C-G3
srcb-C-R5	SMIPG-C-G1,BUPTTeam-C-G1,MSRSC-C-R4,PolyU-C-G1,ITNLP-C-R3,WUST-C-R2,SLSTC-C-R1,CIAL-C-G1,ckip-C-G3
DeepIntell-C-R1	BUPTTeam-C-G1,MSRSC-C-R4,PolyU-C-G1,ITNLP-C-R3,WUST-C-R2,SLSTC-C-R1,CIAL-C-G1,ckip-C-G3
CYIII-C-R1	BUPTTeam-C-G1,MSRSC-C-R4,PolyU-C-G1,ITNLP-C-R3,WUST-C-R2,SLSTC-C-R1,CIAL-C-G1,ckip-C-G3
TUA1-C-R4	BUPTTeam-C-G1,MSRSC-C-R4,PolyU-C-G1,ITNLP-C-R3,WUST-C-R2,SLSTC-C-R1,CIAL-C-G1,ckip-C-G3
iNLP-C-R1	BUPTTeam-C-G1,MSRSC-C-R4,PolyU-C-G1,ITNLP-C-R3,WUST-C-R2,SLSTC-C-R1,CIAL-C-G1,ckip-C-G3
UB-C-R1	BUPTTeam-C-G1,MSRSC-C-R4,PolyU-C-G1,ITNLP-C-R3,WUST-C-R2,SLSTC-C-R1,CIAL-C-G1,ckip-C-G3
WIDM-C-R1	BUPTTeam-C-G1,MSRSC-C-R4,PolyU-C-G1,ITNLP-C-R3,WUST-C-R2,SLSTC-C-R1,CIAL-C-G1,ckip-C-G3
Gbot-C-G4	BUPTTeam-C-G1,MSRSC-C-R4,PolyU-C-G1,ITNLP-C-R3,WUST-C-R2,SLSTC-C-R1,CIAL-C-G1,ckip-C-G3
rucir-C-R2	PolyU-C-G1,ITNLP-C-R3,WUST-C-R2,SLSTC-C-R1,CIAL-C-G1,ckip-C-G3
SMIPG-C-G1	SLSTC-C-R1,CIAL-C-G1,ckip-C-G3
These runs are	significantly better than these runs in terms of mean official P+
SG01-C-G1	TUA1-C-R4,WIDM-C-R1,rucir-C-R2,Gbot-C-G4,SMIPG-C-G1,MSRSC-C-R4,BUPTTeam-C-G1,PolyU-C-R2,ITNLP-C-R3, WUST-C-R2,CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
splab-C-G4	SMIPG-C-G1,MSRSC-C-R4,BUPTTeam-C-G1,PolyU-C-R2,ITNLP-C-R3,WUST-C-R2,CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
Beihang-C-R4	SMIPG-C-G1,MSRSC-C-R4,BUPTTeam-C-G1,PolyU-C-R2,ITNLP-C-R3,WUST-C-R2,CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
DeepIntell-C-R1	SMIPG-C-G1,MSRSC-C-R4,BUPTTeam-C-G1,PolyU-C-R2,ITNLP-C-R3,WUST-C-R2,CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
Nders-C-R2	SMIPG-C-G1,MSRSC-C-R4,BUPTTeam-C-G1,PolyU-C-R2,ITNLP-C-R3,WUST-C-R2,CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
srcb-C-R1	SMIPG-C-G1,MSRSC-C-R4,BUPTTeam-C-G1,PolyU-C-R2,ITNLP-C-R3,WUST-C-R2,CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
iNLP-C-R1	SMIPG-C-G1,MSRSC-C-R4,BUPTTeam-C-G1,PolyU-C-R2,ITNLP-C-R3,WUST-C-R2,CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
CYIII-C-R1	SMIPG-C-G1,MSRSC-C-R4,BUPTTeam-C-G1,PolyU-C-R2,ITNLP-C-R3,WUST-C-R2,CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
UB-C-R4	MSRSC-C-R4,BUPTTeam-C-G1,PolyU-C-R2,ITNLP-C-R3,WUST-C-R2,CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
TUA1-C-R4	MSRSC-C-R4,BUPTTeam-C-G1,PolyU-C-R2,ITNLP-C-R3,WUST-C-R2,CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
WIDM-C-R1	MSRSC-C-R4,BUPTTeam-C-G1,PolyU-C-R2,ITNLP-C-R3,WUST-C-R2,CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
rucir-C-R2	MSRSC-C-R4,BUPTTeam-C-G1,PolyU-C-R2,ITNLP-C-R3,WUST-C-R2,CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
Gbot-C-G4	BUPTTeam-C-G1,PolyU-C-R2,ITNLP-C-R3,WUST-C-R2,CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
SMIPG-C-G1	ITNLP-C-R3,WUST-C-R2,CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
MSRSC-C-R4	CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
BUPTTeam-C-G1	ckip-C-G3
PolyU-C-R2	ckip-C-G3
These runs are	significantly better than these runs in terms of mean official nERR@10
SG01-C-G1	WIDM-C-R1,rucir-C-R2,Gbot-C-G4,SMIPG-C-G1,MSRSC-C-R4,BUPTTeam-C-G1,PolyU-C-R2,ITNLP-C-R3,WUST-C-R2, CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
splab-C-G4	SMIPG-C-G1,MSRSC-C-R4,BUPTTeam-C-G1,PolyU-C-R2,ITNLP-C-R3,WUST-C-R2,CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
Beihang-C-R4	SMIPG-C-G1,MSRSC-C-R4,BUPTTeam-C-G1,PolyU-C-R2,ITNLP-C-R3,WUST-C-R2,CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
DeepIntell-C-R1	SMIPG-C-G1,MSRSC-C-R4,BUPTTeam-C-G1,PolyU-C-R2,ITNLP-C-R3,WUST-C-R2,CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
Nders-C-R2	SMIPG-C-G1,MSRSC-C-R4,BUPTTeam-C-G1,PolyU-C-R2,ITNLP-C-R3,WUST-C-R2,CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
srcb-C-G2	SMIPG-C-G1,MSRSC-C-R4,BUPTTeam-C-G1,PolyU-C-R2,ITNLP-C-R3,WUST-C-R2,CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
iNLP-C-R1	SMIPG-C-G1,MSRSC-C-R4,BUPTTeam-C-G1,PolyU-C-R2,ITNLP-C-R3,WUST-C-R2,CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
CYIII-C-R1	SMIPG-C-G1,MSRSC-C-R4,BUPTTeam-C-G1,PolyU-C-R2,ITNLP-C-R3,WUST-C-R2,CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
TUA1-C-R4	SMIPG-C-G1,MSRSC-C-R4,BUPTTeam-C-G1,PolyU-C-R2,ITNLP-C-R3,WUST-C-R2,CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
UB-C-R2	SMIPG-C-G1,MSRSC-C-R4,BUPTTeam-C-G1,PolyU-C-R2,ITNLP-C-R3,WUST-C-R2,CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
WIDM-C-R1	MSRSC-C-R4,BUPTTeam-C-G1,PolyU-C-R2,ITNLP-C-R3,WUST-C-R2,CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
rucir-C-R2	MSRSC-C-R4,BUPTTeam-C-G1,PolyU-C-R2,ITNLP-C-R3,WUST-C-R2,CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
Gbot-C-G4	MSRSC-C-R4,BUPTTeam-C-G1,PolyU-C-R2,ITNLP-C-R3,WUST-C-R2,CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
SMIPG-C-G1	ITNLP-C-R3,WUST-C-R2,CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
MSRSC-C-R4	WUST-C-R2,CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
BUPTTeam-C-G1	ckip-C-G3
PolyU-C-R2	ckip-C-G3

For the Japanese subtask, we had five participating teams. Each team was allowed to submit up to five runs. We received 15 runs in total (see Table 11).

3.1 Task Definition

In the Japanese subtask of STC-2, we used Yahoo! News comments data instead of Twitter data we used in STC-1. We changed the dataset because of the problem that the data frequently disappear in Twitter because users sometimes protect their accounts or remove their tweets, which causes the problem of reproducibility in follow-on experiments. Yahoo! News comments data are composed of approximately one million comment-response pairs. Unlike STC1, both retrieval- and generative-based methods are allowed for response generation.

The task definition is summarized below.

- In the **development** phrase, participants are provided with development data (comment-response pairs) with fluency, coherence, context-dependence, and informativeness labels (see Section 3.3.2). Participants develop their own models to retrieve or generate responses for a given comment.
- In the **test** phase, participants are given a set of test

Table 10: Statistical significance with best run from each team according to unanimity-aware ($p = 0.2$) STC-2 Chinese performances (randomized Tukey HSD test, $B = 10,000, \alpha = 0.05$). Results that differ from official ones (Table 9) are indicated in bold.

These runs are	significantly better than these runs in terms of mean unanimity-aware ($p = 0.2$) nG@1
SG01-C-G1	UB-C-R1,WIDM-C-R1,Gbot-C-G4,rucir-C-R2,SMIPG-C-G1,BUPTTeam-C-G1,MSRSC-C-R4,PolyU-C-G1,ITNLP-C-R3, WUST-C-R2,CIAL-C-G1,SLSTC-C-R1,ckip-C-G3
splab-C-G4	SMIPG-C-G1,BUPTTeam-C-G1,MSRSC-C-R4,PolyU-C-G1,ITNLP-C-R3,WUST-C-R2,CIAL-C-G1,SLSTC-C-R1,ckip-C-G3
Beihang-C-R4	SMIPG-C-G1,BUPTTeam-C-G1,MSRSC-C-R4,PolyU-C-G1,ITNLP-C-R3,WUST-C-R2,CIAL-C-G1,SLSTC-C-R1,ckip-C-G3
Nders-C-R4	SMIPG-C-G1,BUPTTeam-C-G1,MSRSC-C-R4,PolyU-C-G1,ITNLP-C-R3,WUST-C-R2,CIAL-C-G1,SLSTC-C-R1,ckip-C-G3
srcb-C-R5	SMIPG-C-G1,BUPTTeam-C-G1,MSRSC-C-R4,PolyU-C-G1,ITNLP-C-R3,WUST-C-R2,CIAL-C-G1,SLSTC-C-R1,ckip-C-G3
DeepIntell-C-R1	BUPTTeam-C-G1,MSRSC-C-R4,PolyU-C-G1,ITNLP-C-R3,WUST-C-R2,CIAL-C-G1,SLSTC-C-R1,ckip-C-G3
iNLP-C-R1	BUPTTeam-C-G1,MSRSC-C-R4,PolyU-C-G1,ITNLP-C-R3,WUST-C-R2,CIAL-C-G1,SLSTC-C-R1,ckip-C-G3
CYIII-C-R1	BUPTTeam-C-G1,MSRSC-C-R4,PolyU-C-G1,ITNLP-C-R3,WUST-C-R2,CIAL-C-G1,SLSTC-C-R1,ckip-C-G3
TUA1-C-R4	BUPTTeam-C-G1,MSRSC-C-R4,PolyU-C-G1,ITNLP-C-R3,WUST-C-R2,CIAL-C-G1,SLSTC-C-R1,ckip-C-G3
UB-C-R1	BUPTTeam-C-G1,MSRSC-C-R4,PolyU-C-G1,ITNLP-C-R3,WUST-C-R2,CIAL-C-G1,SLSTC-C-R1,ckip-C-G3
WIDM-C-R1	BUPTTeam-C-G1,MSRSC-C-R4,PolyU-C-G1,ITNLP-C-R3,WUST-C-R2,CIAL-C-G1,SLSTC-C-R1,ckip-C-G3
Gbot-C-G4	BUPTTeam-C-G1,MSRSC-C-R4,PolyU-C-G1,ITNLP-C-R3,WUST-C-R2,CIAL-C-G1,SLSTC-C-R1,ckip-C-G3
rucir-C-R2	**MSRSC-C-R4**,PolyU-C-G1,ITNLP-C-R3,WUST-C-R2,CIAL-C-G1,SLSTC-C-R1,ckip-C-G3
SMIPG-C-G1	**WUST-C-R2**,CIAL-C-G1,SLSTC-C-R1,ckip-C-G3
BUPTTeam-C-G1	**ckip-C-G3**
These runs are	significantly better than these runs in terms of mean unanimity-aware ($p = 0.2$) P+
SG01-C-G1	WIDM-C-R1,TUA1-C-R4,rucir-C-R2,Gbot-C-G4,SMIPG-C-G1,MSRSC-C-R4,BUPTTeam-C-G1,PolyU-C-R2,ITNLP-C-R3, WUST-C-R2,CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
splab-C-G4	SMIPG-C-G1,MSRSC-C-R4,BUPTTeam-C-G1,PolyU-C-R2,ITNLP-C-R3,WUST-C-R2,CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
Beihang-C-R4	SMIPG-C-G1,MSRSC-C-R4,BUPTTeam-C-G1,PolyU-C-R2,ITNLP-C-R3,WUST-C-R2,CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
DeepIntell-C-R1	SMIPG-C-G1,MSRSC-C-R4,BUPTTeam-C-G1,PolyU-C-R2,ITNLP-C-R3,WUST-C-R2,CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
Nders-C-R5	SMIPG-C-G1,MSRSC-C-R4,BUPTTeam-C-G1,PolyU-C-R2,ITNLP-C-R3,WUST-C-R2,CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
iNLP-C-R1	SMIPG-C-G1,MSRSC-C-R4,BUPTTeam-C-G1,PolyU-C-R2,ITNLP-C-R3,WUST-C-R2,CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
srcb-C-R5	SMIPG-C-G1,MSRSC-C-R4,BUPTTeam-C-G1,PolyU-C-R2,ITNLP-C-R3,WUST-C-R2,CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
CYIII-C-R1	SMIPG-C-G1,MSRSC-C-R4,BUPTTeam-C-G1,PolyU-C-R2,ITNLP-C-R3,WUST-C-R2,CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
UB-C-R4	MSRSC-C-R4,BUPTTeam-C-G1,PolyU-C-R2,ITNLP-C-R3,WUST-C-R2,CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
WIDM-C-R1	MSRSC-C-R4,BUPTTeam-C-G1,PolyU-C-R2,ITNLP-C-R3,WUST-C-R2,CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
TUA1-C-R4	MSRSC-C-R4,BUPTTeam-C-G1,PolyU-C-R2,ITNLP-C-R3,WUST-C-R2,CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
rucir-C-R2	MSRSC-C-R4,BUPTTeam-C-G1,PolyU-C-R2,ITNLP-C-R3,WUST-C-R2,CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
Gbot-C-G4	BUPTTeam-C-G1,PolyU-C-R2,ITNLP-C-R3,WUST-C-R2,CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
SMIPG-C-G1	ITNLP-C-R3,WUST-C-R2,CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
MSRSC-C-R4	**WUST-C-R2**,CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
BUPTTeam-C-G1	ckip-C-G3
PolyU-C-R2	ckip-C-G3
These runs are	significantly better than these runs in terms of mean unanimity-aware ($p = 0.2$) nERR@10
SG01-C-G1	WIDM-C-R1,rucir-C-R2,Gbot-C-G4,SMIPG-C-G1,MSRSC-C-R4,BUPTTeam-C-G1,PolyU-C-R2,ITNLP-C-R3,WUST-C-R2, CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
splab-C-G4	SMIPG-C-G1,MSRSC-C-R4,BUPTTeam-C-G1,PolyU-C-R2,ITNLP-C-R3,WUST-C-R2,CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
Beihang-C-R4	SMIPG-C-G1,MSRSC-C-R4,BUPTTeam-C-G1,PolyU-C-R2,ITNLP-C-R3,WUST-C-R2,CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
DeepIntell-C-R1	SMIPG-C-G1,MSRSC-C-R4,BUPTTeam-C-G1,PolyU-C-R2,ITNLP-C-R3,WUST-C-R2,CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
Nders-C-R5	SMIPG-C-G1,MSRSC-C-R4,BUPTTeam-C-G1,PolyU-C-R2,ITNLP-C-R3,WUST-C-R2,CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
srcb-C-G2	SMIPG-C-G1,MSRSC-C-R4,BUPTTeam-C-G1,PolyU-C-R2,ITNLP-C-R3,WUST-C-R2,CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
iNLP-C-R1	SMIPG-C-G1,MSRSC-C-R4,BUPTTeam-C-G1,PolyU-C-R2,ITNLP-C-R3,WUST-C-R2,CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
CYIII-C-R1	SMIPG-C-G1,MSRSC-C-R4,BUPTTeam-C-G1,PolyU-C-R2,ITNLP-C-R3,WUST-C-R2,CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
TUA1-C-R4	SMIPG-C-G1,MSRSC-C-R4,BUPTTeam-C-G1,PolyU-C-R2,ITNLP-C-R3,WUST-C-R2,CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
UB-C-R4	SMIPG-C-G1,MSRSC-C-R4,BUPTTeam-C-G1,PolyU-C-R2,ITNLP-C-R3,WUST-C-R2,CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
WIDM-C-R1	MSRSC-C-R4,BUPTTeam-C-G1,PolyU-C-R2,ITNLP-C-R3,WUST-C-R2,CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
rucir-C-R2	MSRSC-C-R4,BUPTTeam-C-G1,PolyU-C-R2,ITNLP-C-R3,WUST-C-R2,CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
Gbot-C-G4	MSRSC-C-R4,BUPTTeam-C-G1,PolyU-C-R2,ITNLP-C-R3,WUST-C-R2,CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
SMIPG-C-G1	ITNLP-C-R3,WUST-C-R2,CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
MSRSC-C-R4	WUST-C-R2,CIAL-C-R2,SLSTC-C-R1,ckip-C-G3
BUPTTeam-C-G1	ckip-C-G3
PolyU-C-R2	ckip-C-G3

comments. Each system outputs a ranked list of up to ten responses to a given comment.

- In the **evaluation** phase, all the results are pooled and labeled by humans. Each retrieved/generated response is labeled with relevance labels by multiple assessors to derive the values of evaluation measures described in Section 3.2.

3.2 Evaluation Measures

We used the same evaluation measures as in STC-1 (see the overview paper of STC-1 [5]).

We used $nG@1$ and $nERR@2$ (only the top two comments were evaluated due to budgetary reasons) as evaluation measures. For these two evaluation measures, since we used multiple assessors for a response, we used the following definition for $g(r)$ as *averaged gain*:

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

where n is the number of labels given to each comment (in our setting, $n = 5$), and $g_i(r)$ is the i-th relevance label for

Table 11: Organization and number of submitted runs of participating teams in STC Japanese subtask

Group ID	Organization	No. of runs
AITOK	Tokushima University	1
KIT16	Kyoto Institute of Technology	4
KSU	Kyoto Sangyo University	3
mnmlb	The University of Electro-Communications	5
YJTI	Yahoo Japan Corporation	2

the comment at rank r. With this averaged gain, we can use the same definition of $nG@1$ and $nERR@2$ as in the Chinese task. The P^+ was not used in the Japanese task.

In addition to $nG@1$ and $nERR@2$, we used *accuracy* $Acc_G@k$:

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

where $l_i(r)$ is the i-th relevance label. The term G specifies relevance labels regarded as "correct". This measure computes the average number of labels judged as correct $(l_i(r) \in G)$. In this task, we evaluated the results with $G = \{L2\}$ and $G = \{L1, L2\}$ for $k = 1$ and $k = 2$.

Note that we did not use the unanimity-aware gain in the Japanese subtask.

3.3 Japanese Test Collection

We created the Japanese test collection by using Yahoo! News comments data.

3.3.1 Yahoo! News comments data

Yahoo! News comments data are composed of approximately one million pairs of comments and replies to the articles and related information on the comment-reply pairs. Comments-replies were retrieved from the comment-reply pairs that were posted by the users on the articles published in Yahoo! News in approximately two months. The information included in Yahoo! News comments data is as follows:

Textual data Comment text, Reply text

Information on comment-reply ID of comment-reply pair

Information on comments Date and time of posting, Number (hereafter No.) of replies, No. of agrees, No. of disagrees.

Information on replies Date and time of posting, No. of replies, No. of agrees, No. of disagrees.

Information on articles Article ID, Category, Title in Yahoo! Topics*, Genre*, Theme*. Asterisk indicates optional pieces of information that are provided to some of the news articles.

Figure 5 shows an example comment and its responses for a news article in the development data. Replies 1 and 3 are gold replies (that is, original replies) and reply 2 is the response retrieved by the baseline.

We split the data into two sets; one is a repository to be distributed to participants and the other as a held-out set for creating the development and test data. Statistical information on the repository is provided in Table 12.

Table 12: Statistics of dataset for Japanese subtask

Repository	No. of news articles	43,729
	No. of comments	293,457
	No. of replies	894,998
Development data	No. of news articles	147
	No. of comments	147
	No. of labeled replies	1470
Test data	No. of news articles	100
	No. of comments	100

3.3.2 Development data

We created our development data in the following manner. First, we randomly sampled 147 comments from the held-out set. Then, for each sampled comment, we retrieved responses from the repository.

For the retrieval, we used the same procedure as STC-1. First, we indexed the repository with Lucene (version 5.2.1) using the built-in JapaneseAnalyzer. A comment and its response pair was treated as a document to be added to the index.

Given an input comment, the index is searched to find a document whose comment matches the input comment, then its response is returned. We retrieved the top five documents in this manner. We also searched for a document whose response matches the input comment, and used the matched responses as retrieval results. In this way, we additionally retrieved five more documents, resulting in obtaining ten documents for each comment. We used default search parameters when using Lucene.

Gold responses (original responses that were given to input comments) were added to the retrieval results, and the data were annotated for relevance assessment. The development data consisted of 1470 comment-response pairs for 147 news articles. We used our annotators (not crowdsourcing) for relevance assessment. They were all undergraduate and graduate students majoring in computer science, consisting of 4 males and 1 female in early twenties. They were recruited at the organizer's university as part-timers for this annotation task. Each retrieved response or gold response was annotated by these five annotators on the basis of the following perspectives and labels. The annotators were requested to conduct web search when they were unfamiliar with the topics in the comment-reply pairs. Note that since we are using news data as well as comment-response pairs, the perspectives slightly differ from those of the Chinese subtask.

- (1) Fluent: The response is fluent and understandable from a grammatical point of view. (L1: fluent, L0: not fluent and hard to understand)
- (2) Coherent: The response maintains coherence with the news topic and comment. (L1: coherent, L0: not coherent)
- (3) Context-dependent: The response depends on and is related to the comment. (L2: context-dependent, L1: context-dependent to some extent, L0: not context-dependent)
- (4) Informative: The response is informative and influences the author of the comment. (L2: informative

News-related information	
Category	base
Title in Yahoo! Topics	M・ラミレス 高知上陸は3月 M. Ramirez to arrive in Kochi in March
Genre	スポーツ (sports)
Theme	四国アイランドリーグplus|ベースボール・チャレンジ・リーグ（BCリーグ）|群馬ダイヤモンドペガサス|福井ミラクルエレファンツ Shikoku Island League Plus|Baseball Challenge League (BC League)|Gunma Diamond Pegasus|Fukui Miracle Elephants
Comment and replies	
Comment	日本でプレーしたい理由が気になります Curious why he wants to play in Japan.
Reply 1	2013年3月12日、台湾の義大ライノズ契約したが、同6月19日、家族と離れて長くプレーすることはできないとの理由で退団を表明した。らしい。 今度は3ヶ月もつか。 Signed with EDA Rhinos on March 12, 2013 but left the team on June 19, saying that he cannot play away from his family for a long time. Will he last 3 months this time?
	A1 (L1,L1,L2,L2) [L2,L2], A2 (L1,L1,L2,L2) [L2,L2], A3 (L1,L1,L2,L2) [L2,L2], A4 (L1,L1,L1,L2) [L1,L1], A5 (L1,L1,L1,L2) [L1,L1]
Reply 2	理由が気になる Wonder why he wants to play in Japan.
	A1 (L1,L1,L2,L1) [L1,L1], A2 (L1,L1,L2,L0) [L1,L0], A3 (L1,L1,L0,L0) [L1,L0], A4 (L1,L1,L2,L0) [L1,L0], A5 (L1,L1,L1,L0) [L1,L0]
Reply 3	理由はなんでもいい。 日本でプレーし続けてくれるのなら。 Reason does not matter as long as he keeps playing in Japan.
	A1 (L1,L1,L2,L2) [L2,L2], A2 (L1,L1,L2,L0) [L1,L0], A3 (L1,L1,L2,L1) [L1,L1], A4 (L1,L1,L2,L1) [L1,L1], A5 (L1,L1,L1,L1) [L1,L1]

Figure 5: Example comment and its three comments for news article in development data. For each reply, relevance assessment results are shown for each reply; A1 to A5 stand for the five annotators, and the parentheses contain the relevance labels for fluency, coherence, context-dependence, and informativeness, respectively. Brackets contain final relevance assessment labels for Rule-1 and Rule-2, respectively.

enough to continue and extend the dialogue to discuss a new topic; L1: informative to some extent but not enough to continue and extend the dialogue, including agreement and disagreement; L0: not informative, including counter-questions.)

3.3.3 Test Data

To create test data, we randomly sampled 100 comments from the held-out set. The news related to the sampled comments do not overlap those of the development data.

3.4 Relevance Assessments

For each comment in the formal run, up to ten results were allowed. However, for budgetary reasons, we used only the top two retrieved/generated replies for relevance assessment. All the retrieved responses from the participating teams were labeled L0, L1, or L2 (for context-dependent and informative). We used the following two rules to decide the final relevance label on the basis of fluency, coherence, context-dependence, and informativeness labels.

Rule-1 is similar to that used in the Chinese subtask. Rule-2 differs from the first in that it penalizes context-independent or uninformative responses. In a way, Rule-1 is not strict regarding the content of a response as long as the conversation can be continued.

```
RULE-1:
IF fluent & coherent = L1
  IF context-dependent & informative = L2
    THEN L2
  ELSE L1
ELSE
  L0

RULE-2:
IF fluent & coherent = L1
  IF context-dependent & informative = L2
    THEN L2
  ELSE IF context-dependent or informative = L0
    THEN L0
  ELSE L1
ELSE
  L0
```

Since the labeling task can be quite subjective, we used five annotators (the same ones who annotated the development set) for evaluating each response.

3.5 Japanese Run Results

Tables 13 and 14 list the official STC results for the 15 Japanese runs from 5 teams when Rule-1 and Rule-2 were used, respectively. Brief descriptions of the runs are given in Table 17 in the Appendix. The runs were sorted by the mean values of the evaluation measures. GOLD indicates the original responses given to the test comments, and BASELINE indicates a simple Lucene-based baseline, which we used to create the development data.

KIT16 and YJTI seemed to have achieved good results for both Rule-1 and Rule-2 with AITOK performing well for Rule-1. Although thorough examination is needed, from the results of Acc_{L2}@1 for AITOK for Rule-1, we can see that it is possible to continue the conversation by using pattern-based responses. However, when we look at the results of Rule-2, there is still some gap between GOLD and proposed methods, indicating that context-dependent or informative responses are difficult.

We also used a randomized Tukey HSD test with $B = 1000$ trials for each evaluation measure.

When we used Rule-1, of the $17 * 16/2 = 136$ run pairs (including GOLD and BASELINE as runs), we obtained the

Table 13: Official STC results for 15 Japanese runs from 5 teams (Rule-1)

Run	Mean nG@1	Run	Mean $nERR$@2	Run	Mean Acc_{L2}@1
GOLD-J-R1	0.7753	GOLD-J-R1	0.7757	GOLD-J-R1	0.4720
KIT16-J-R1	0.5014	KIT16-J-R1	0.5580	YJTI-J-R2	0.2040
YJTI-J-R2	0.4893	YJTI-J-R2	0.5468	YJTI-J-R1	0.1860
KIT16-J-R4	0.4804	KIT16-J-R4	0.5372	KIT16-J-R1	0.1800
AITOK-J-R1	0.4468	AITOK-J-R1	0.4838	BASELINE-J-R1	0.1680
YJTI-J-R1	0.4322	YJTI-J-R1	0.4731	KIT16-J-R4	0.1660
KSU-J-R1	0.4150	KSU-J-R1	0.4538	KSU-J-R1	0.1560
mnmlb-J-R2	0.3690	mnmlb-J-R2	0.4410	KSU-J-R3	0.1220
BASELINE-J-R1	0.3518	BASELINE-J-R1	0.4330	mnmlb-J-R2	0.1040
KIT16-J-R2	0.3484	KSU-J-R3	0.3737	KIT16-J-R2	0.0960
KSU-J-R3	0.3303	mnmlb-J-R1	0.3463	mnmlb-J-R4	0.0940
mnmlb-J-R1	0.2949	mnmlb-J-R5	0.3066	KIT16-J-R3	0.0860
KIT16-J-R3	0.2744	KIT16-J-R2	0.2952	mnmlb-J-R1	0.0700
mnmlb-J-R4	0.2584	KSU-J-R2	0.2858	mnmlb-J-R5	0.0680
mnmlb-J-R5	0.2544	mnmlb-J-R4	0.2799	mnmlb-J-R3	0.0560
KSU-J-R2	0.2541	mnmlb-J-R3	0.2538	AITOK-J-R1	0.0280
mnmlb-J-R3	0.2230	KIT16-J-R3	0.2312	KSU-J-R2	0.0020
Run	Mean Acc_{L2}@2	Run	Mean $Acc_{L1,L2}$@1	Run	Mean $Acc_{L1,L2}$@2
GOLD-J-R1	0.4430	AITOK-J-R1	0.9840	AITOK-J-R1	0.9710
YJTI-J-R2	0.2030	GOLD-J-R1	0.8980	GOLD-J-R1	0.8840
BASELINE-J-R1	0.1900	KIT16-J-R1	0.8240	KIT16-J-R1	0.7980
KIT16-J-R1	0.1690	KIT16-J-R4	0.8000	KIT16-J-R4	0.7700
KIT16-J-R4	0.1610	YJTI-J-R2	0.7620	YJTI-J-R2	0.7310
YJTI-J-R1	0.1490	KSU-J-R1	0.6680	mnmlb-J-R2	0.6600
KSU-J-R1	0.1350	mnmlb-J-R2	0.6540	KSU-J-R1	0.6320
mnmlb-J-R2	0.1210	YJTI-J-R1	0.6480	KIT16-J-R2	0.6320
KSU-J-R3	0.1130	KIT16-J-R2	0.6320	YJTI-J-R1	0.6210
KIT16-J-R2	0.0960	KSU-J-R2	0.5840	BASELINE-J-R1	0.5900
KIT16-J-R3	0.0860	mnmlb-J-R1	0.5400	mnmlb-J-R1	0.5360
mnmlb-J-R4	0.0750	KSU-J-R3	0.5300	KSU-J-R2	0.5290
mnmlb-J-R1	0.0710	BASELINE-J-R1	0.5200	KSU-J-R3	0.5210
mnmlb-J-R5	0.0690	KIT16-J-R3	0.4660	KIT16-J-R3	0.4660
AITOK-J-R1	0.0660	mnmlb-J-R5	0.4520	mnmlb-J-R5	0.4640
mnmlb-J-R3	0.0450	mnmlb-J-R4	0.4200	mnmlb-J-R3	0.3930
KSU-J-R2	0.0030	mnmlb-J-R3	0.4020	mnmlb-J-R4	0.3800

following significant differences; "$X > Y$" means "X statistically significantly outperformed Y at $\alpha = 0.05$".

- In terms of Mean $Acc_{L1,L2}$@1,
  AITOK-J-R1 > YJTI-J-R2, KSU-J-R1, mnmlb-J-R2, BASELINE-J-R1;
  GOLD-J-R1 > KSU-J-R1, mnmlb-J-R2, BASELINE-J-R1;
  KIT16-J-R1 > BASELINE-J-R1;
  YJTI-J-R2 > BASELINE-J-R1;

- In terms of Mean $Acc_{L1,L2}$@2,
  AITOK-J-R1 > KIT16-J-R1, YJTI-J-R2, mnmlb-J-R2, KSU-J-R1, BASELINE-J-R1;
  GOLD-J-R1 > mnmlb-J-R2, KSU-J-R1, BASELINE-J-R1;
  KIT16-J-R1 > KSU-J-R1, BASELINE-J-R1;

- In terms of Mean Acc_{L2}@1,
  GOLD-J-R1 > YJTI-J-R2, KIT16-J-R1, BASELINE-J-R1, KSU-J-R1, mnmlb-J-R2, AITOK-J-R1;
  YJTI-J-R2 > AITOK-J-R1;
  KIT16-J-R1 > AITOK-J-R1;
  BASELINE-J-R1 > AITOK-J-R1;
  KSU-J-R1 > AITOK-J-R1;

- In terms of Mean Acc_{L2}@2,
  GOLD-J-R1 > YJTI-J-R2, BASELINE-J-R1, KIT16-J-R1, KSU-J-R1, mnmlb-J-R2, AITOK-J-R1;
  YJTI-J-R2 > AITOK-J-R1;
  BASELINE-J-R1 > AITOK-J-R1;
  KIT16-J-R1 > AITOK-J-R1;

- In terms of Mean nG@1,
  GOLD-J-R1 > KIT16-J-R1, YJTI-J-R2, AITOK-J-R1, KSU-J-R1, mnmlb-J-R2, BASELINE-J-R1;
  KIT16-J-R1 > BASELINE-J-R1;

- In terms of Mean $nERR$@2,
  GOLD-J-R1 > KIT16-J-R1, YJTI-J-R2, AITOK-J-R1, KSU-J-R1, mnmlb-J-R2, BASELINE-J-R1;

When we used Rule-2, we obtained the following significant differences at the significance level of $\alpha = 0.05$.

- In terms of Mean $Acc_{L1,L2}$@1,
  GOLD-J-R1 > KIT16-J-R1, KSU-J-R1, mnmlb-J-R2, BASELINE-J-R1, AITOK-J-R1;
  YJTI-J-R2 > mnmlb-J-R2, BASELINE-J-R1, AITOK-J-R1;
  KIT16-J-R1 > AITOK-J-R1;

Table 14: Official STC results for 15 Japanese runs from 5 teams (Rule-2)

Run	Mean nG@1	Run	Mean $nERR$@2	Run	Mean Acc_{L2}@1
GOLD-J-R1	0.7646	GOLD-J-R1	0.7639	GOLD-J-R1	0.4720
YJTI-J-R2	0.4726	YJTI-J-R2	0.5288	YJTI-J-R2	0.2040
KIT16-J-R1	0.4173	KIT16-J-R1	0.4676	YJTI-J-R1	0.1860
YJTI-J-R1	0.4171	KIT16-J-R4	0.4549	KIT16-J-R1	0.1800
KIT16-J-R4	0.4014	YJTI-J-R1	0.4544	BASELINE-J-R1	0.1680
KSU-J-R1	0.3762	KSU-J-R1	0.4101	KIT16-J-R4	0.1660
BASELINE-J-R1	0.3320	BASELINE-J-R1	0.4094	KSU-J-R1	0.1560
mnmlb-J-R2	0.3144	mnmlb-J-R2	0.3804	KSU-J-R3	0.1220
KSU-J-R3	0.2912	KSU-J-R3	0.3317	mnmlb-J-R2	0.1040
KIT16-J-R2	0.2748	mnmlb-J-R1	0.2829	KIT16-J-R2	0.0960
mnmlb-J-R1	0.2518	mnmlb-J-R5	0.2573	mnmlb-J-R4	0.0940
KIT16-J-R3	0.2385	mnmlb-J-R4	0.2415	KIT16-J-R3	0.0860
mnmlb-J-R4	0.2212	KIT16-J-R2	0.2338	mnmlb-J-R1	0.0700
mnmlb-J-R5	0.2144	mnmlb-J-R3	0.2018	mnmlb-J-R5	0.0680
mnmlb-J-R3	0.1792	KIT16-J-R3	0.2012	mnmlb-J-R3	0.0560
AITOK-J-R1	0.0816	AITOK-J-R1	0.1758	AITOK-J-R1	0.0280
KSU-J-R2	0.0177	KSU-J-R2	0.0230	KSU-J-R2	0.0020
Run	Mean Acc_{L2}@2	Run	Mean $Acc_{L1,L2}$@1	Run	Mean $Acc_{L1,L2}$@2
GOLD-J-R1	0.4430	GOLD-J-R1	0.8660	GOLD-J-R1	0.8430
YJTI-J-R2	0.2030	YJTI-J-R2	0.7200	YJTI-J-R2	0.6900
BASELINE-J-R1	0.1900	KIT16-J-R1	0.6320	KIT16-J-R1	0.6050
KIT16-J-R1	0.1690	KIT16-J-R4	0.6200	KIT16-J-R4	0.5900
KIT16-J-R4	0.1610	YJTI-J-R1	0.6100	YJTI-J-R1	0.5750
YJTI-J-R1	0.1490	KSU-J-R1	0.5760	mnmlb-J-R2	0.5360
KSU-J-R1	0.1350	mnmlb-J-R2	0.5300	KSU-J-R1	0.5360
mnmlb-J-R2	0.1210	BASELINE-J-R1	0.4740	BASELINE-J-R1	0.5360
KSU-J-R3	0.1130	KIT16-J-R2	0.4680	KIT16-J-R2	0.4680
KIT16-J-R2	0.0960	KSU-J-R3	0.4420	KSU-J-R3	0.4330
KIT16-J-R3	0.0860	mnmlb-J-R1	0.4380	mnmlb-J-R1	0.3880
mnmlb-J-R4	0.0750	KIT16-J-R3	0.3840	KIT16-J-R3	0.3840
mnmlb-J-R1	0.0710	mnmlb-J-R5	0.3600	mnmlb-J-R5	0.3550
mnmlb-J-R5	0.0690	mnmlb-J-R4	0.3380	AITOK-J-R1	0.3100
AITOK-J-R1	0.0660	mnmlb-J-R3	0.3020	mnmlb-J-R4	0.3050
mnmlb-J-R3	0.0450	AITOK-J-R1	0.1400	mnmlb-J-R3	0.2910
KSU-J-R2	0.0030	KSU-J-R2	0.0360	KSU-J-R2	0.0370

KSU-J-R1 > AITOK-J-R1;
mnmlb-J-R2 > AITOK-J-R1;
BASELINE-J-R1 > AITOK-J-R1;

- In terms of Mean $Acc_{L1,L2}$@2,
GOLD-J-R1 > KIT16-J-R1, BASELINE-J-R1, mnmlb-J-R2, KSU-J-R1, AITOK-J-R1;
YJTI-J-R2 > AITOK-J-R1;
KIT16-J-R1 > AITOK-J-R1;
BASELINE-J-R1 > AITOK-J-R1;
mnmlb-J-R2 > AITOK-J-R1;
KSU-J-R1 > AITOK-J-R1;

- In terms of Mean Acc_{L2}@1,
GOLD-J-R1 > YJTI-J-R2, KIT16-J-R1, BASELINE-J-R1, KSU-J-R1, mnmlb-J-R2, AITOK-J-R1;
YJTI-J-R2 > AITOK-J-R1;
KIT16-J-R1 > AITOK-J-R1;
BASELINE-J-R1 > AITOK-J-R1;
KSU-J-R1 > AITOK-J-R1;

- In terms of Mean Acc_{L2}@2,
GOLD-J-R1 > YJTI-J-R2, BASELINE-J-R1, KIT16-J-R1, KSU-J-R1, mnmlb-J-R2, AITOK-J-R1;
YJTI-J-R2 > AITOK-J-R1;
BASELINE-J-R1 > AITOK-J-R1;
KIT16-J-R1 > AITOK-J-R1;

- In terms of Mean nG@1,
GOLD-J-R1 > YJTI-J-R2, KIT16-J-R1, KSU-J-R1, BASELINE-J-R1, mnmlb-J-R2, AITOK-J-R1;
YJTI-J-R2 > mnmlb-J-R2, AITOK-J-R1;
KIT16-J-R1 > AITOK-J-R1;
KSU-J-R1 > AITOK-J-R1;
BASELINE-J-R1 > AITOK-J-R1;
mnmlb-J-R2 > AITOK-J-R1;

- In terms of Mean $nERR$@2,
GOLD-J-R1 > YJTI-J-R2, KIT16-J-R1, KSU-J-R1, BASELINE-J-R1, mnmlb-J-R2, AITOK-J-R1;
YJTI-J-R2 > mnmlb-J-R2, AITOK-J-R1;
KIT16-J-R1 > AITOK-J-R1;
KSU-J-R1 > AITOK-J-R1;
BASELINE-J-R1 > AITOK-J-R1;
mnmlb-J-R2 > AITOK-J-R1;

Tables 15 and 16 compare the rankings according to the six evaluation measures in terms of Kendall's τ, with 95% confidence intervals, for Rule-1 or Rule-2, respectively.

Table 15: Run ranking similarity across six measures: Kendall's τ values with 95% CIs (Rule-1)

	nG@1	$nERR$@2	Acc_{L2}@1	Acc_{L2}@2	$Acc_{L1,L2}$@1	$Acc_{L1,L2}$@2
nG@1	-	.868 [.691, 1.044]	.662 [.369, .954]	.706 [.418, .993]	.735 [.533, .937]	.745 [.586, .905]
$nERR$@2	.868 [.691, 1.044]	-	.588 [.294, .883]	.603 [.295, .911]	.691 [.499, .883]	.701 [.527, .876]
Acc_{L2}@1	.662 [.369, .954]	.588 [.294, .883]	-	.897 [.774, 1.02]	.397 [.009, .785]	.406 [.045, .766]
Acc_{L2}@2	.706 [.418, .993]	.603 [.295, .911]	.897 [.774, 1.02]	-	.441 [.045, .837]	.450 [.101, .799]
$Acc_{L1,L2}$@1	.735 [.533, .937]	.691 [.499, .883]	.397 [.009, .785]	.441 [.045, .837]	-	.893 [.766, 1.02]
$Acc_{L1,L2}$@2	.745 [.586, .905]	.701 [.527, .876]	.406 [.045, .766]	.450 [.101, .799]	.893 [.766, 1.02]	-

Table 16: Run ranking similarity across six measures: Kendall's τ values with 95% CIs (Rule-2)

	nG@1	$nERR$@2	Acc_{L2}@1	Acc_{L2}@2	$Acc_{L1,L2}$@1	$Acc_{L1,L2}$@2
nG@1	-	.882 [.722, 1.043]	.897 [.78, 1.014]	.882 [.737, 1.027]	.941 [.86, 1.022]	.915 [.814, 1.015]
$nERR$@2	.882 [.722, 1.043]	-	.809 [.644, .973]	.794 [.594, .994]	.882 [.727, 1.038]	.855 [.69, 1.02]
Acc_{L2}@1	.897 [.78, 1.014]	.809 [.644, .973]	-	.897 [.774, 1.02]	.838 [.69, .987]	.825 [.678, .972]
Acc_{L2}@2	.882 [.737, 1.027]	.794 [.594, .994]	.897 [.774, 1.02]	-	.853 [.699, 1.007]	.870 [.726, 1.014]
$Acc_{L1,L2}$@1	.941 [.86, 1.022]	.882 [.727, 1.038]	.838 [.69, .987]	.853 [.699, 1.007]	-	.959 [.875, 1.043]
$Acc_{L1,L2}$@2	.915 [.814, 1.015]	.855 [.69, 1.02]	.825 [.678, .972]	.870 [.726, 1.014]	.959 [.875, 1.043]	-

4. CONCLUSIONS AND FUTURE WORK

The main conclusions from the Chinese subtask are as follows.

- SG01 statistically significantly outperformed 13 other teams in terms of all three evaluation measures.
- splab, Beihang, Nders, and srcb statistically significantly outperformed 9 other teams in terms of all three evaluation measures.
- The best G-run from SG01 outperformed the best R-runs from the same team *on average*, but the differences are not statistically significant, and the effects are small. It is too early to conclude that "generation-runs are now better than rule-based runs."
- The additional unanimity-aware results were very similar to the official results, but a few extra statistically significant differences were found. Hence, this approach may deserve further investigation.

The main conclusions from the Japanese subtask are as follows.

- KIT16 and YJTI achieved good results for both Rule-1 and Rule-2 with AITOK performing well for Rule-1.
- KIT16 and YJTI statistically significantly outperformed the baseline in some metrics for Rule-1, and only YJTI statistically significantly outperformed the baseline in $Acc_{L1,L2}$@1 for Rule-2.
- From the results of AITOK for Rule-1, it seems possible to continue the conversation by using pattern-based responses, but from the results of AITOK for Rule-2, it is also evident that it is difficult to achieve context-dependent and informative responses.
- There is still a large gap between the proposed methods and the upper bound (GOLD).
- There are not many generation-based runs in the Japanese subtask, making it difficult to compare retrieval-based and generation-based methods for the Japanese subtask.

Short text conversation is the largest task of NTCIR-13, so we plan to continue to run this task at NTCIR-14 and look forward to seeing new improvements at the next round.

5. ACKNOWLEDGMENTS

We would like to thank all the STC task participants for their effort in exploring new techniques and submitting their runs and reports. We also thank the general chairs and program co-chairs of NTCIR-13 for their encouragement and support.

6. ADDITIONAL AUTHORS

7. REFERENCES

[1] A. Ritter, C. Cherry, and W. B. Dolan. Data-driven response generation in social media. In *Proceedings of EMNLP 2011*, pages 583–593, 2011.

[2] T. Sakai. Statistical reform in information retrieval? *SIGIR Forum*, 48(1):3–12, 2014.

[3] T. Sakai. Unanimity-aware gain for highly subjective assessments. In *Proceedings of EVIA 2017*, 2017.

[4] L. Shang, Z. Lu, and H. Li. Neural responding machine for short-text conversation. In *Proceedings of ACL 2015*, pages 1577–1586, 2015.

[5] 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 NTCIR-12*, pages 473–484, 2016.

[6] H. Wang, Z. Lu, H. Li, and E. Chen. A dataset for research on short-text conversations. In *Proceedings of EMNLP 2013*, pages 935–945, 2013.

Appendix

Table 17: Descriptions of 15 Japanese runs

AITOK-J-R1	Pattern-based response generation depending on whether the comment has ambiguity, understanding result is unreliable, and there is a lack of knowledge.
KIT16-J-R1	Retrieval-based method with TF-IDF and Word2Vec
KIT16-J-R2	Generation-based method with a seq2seq model
KIT16-J-R3	Retrieval-based method with topic-modeling using Chinese restaurant process
KIT16-J-R4	Retrieval-based method with TF-IDF
KSU-J-R1	Retrieval-based method that uses the similarity based on a title and appropriate theme
KSU-J-R2	Generation-based method that uses language and vision information
KSU-J-R3	Retrieval-based method that uses the similarity based on a title and theme
YJTI-J-R1	Retrieval-based method based on a LSTM-RNN model trained over a large dialogue corpus
YJTI-J-R2	Retrieval-based method based on a LSTM-RNN model trained over a large question-answering corpus
mnmlb-J-R1	Retrieval-based method that uses bi-directional LSTM with attention for ranking. Training data are selected using such information as N-grams.
mnmlb-J-R2	Same as R1 without data selection
mnmlb-J-R3	Same as R1 but with a vanilla LSTM
mnmlb-J-R4	Same as R3 without data selection
mnmlb-J-R5	Retrieval-based method that uses CNN ranking.

Table 18: SYSDESC fields of 64 Chinese retrieval-based runs. Note that not all are informative.

Beihang-C-R1.txt	[Naive Solr]
Beihang-C-R2.txt	[Solr qc qp Sim]
Beihang-C-R3.txt	[Annoy+Solr Q-P Q-C Sim]
Beihang-C-R4.txt	[solr+ner + sim + rerank]
Beihang-C-R5.txt	[insert a short description in English here]
CIAL-C-R1.txt	[search original posts and retrieve qualified comments]
CIAL-C-R2.txt	[search original posts and retrieve qualified comments with extension]
CIAL-C-R3.txt	[search extended posts and retrieve qualified comments]
CIAL-C-R4.txt	[search extended posts and retrieve qualified comments with extension]
CYIII-C-R1.txt	[Our system use Lucene to do it. We pick the Noun and Verb to search. And then use TF-IDF to rerank.]
DeepIntell-C-R1.txt	ranking with word-level-feature and DM_penalty2 feature on V1-retrieval results
DeepIntell-C-R2.txt	ranking with word-level-feature and DM_penalty2 feature on V3-retrieval results
DeepIntell-C-R3.txt	ranking with DM_penalty2 feature on V3-retrieval results
DeepIntell-C-R4.txt	ranking with word-level-feature and DM feature on V1-retrieval results
DeepIntell-C-R5.txt	ranking with DM_penalty feature on V1-retrieval results
Gbot-C-R5.txt	Retrieval method using pairwise learning to rank based on CNN
iNLP-C-R1.txt	reranking method with multiple match features
iNLP-C-R2.txt	reranking method with multiple match features (without Elasticsearch score)
iNLP-C-R3.txt	rank candidate comments via post-comt word2vec based similarity with query expansion
iNLP-C-R4.txt	rank candidate comments with Elasticsearch relevance score
iNLP-C-R5.txt	rank candidate comments via post-comt word2vec based similarity without query expansion
ITNLP-C-R1.txt	use a shallow pattern method
ITNLP-C-R2.txt	use shallow pattern and deep pattern combination method
ITNLP-C-R3.txt	use a deep pattern method
MSRSC-C-R1.txt	tfidf_weighted_image_feature,v2,avg_avgtfidf
MSRSC-C-R2.txt	fastrank_training,image_char_feature,w3c2
MSRSC-C-R3.txt	fastrank_training,char_feature
MSRSC-C-R4.txt	baseline,cmp,rerank20
MSRSC-C-R5.txt	fastrank_training,image_feature,v2
Nders-C-R1.txt	Using both Pattern_idf and RandomWalk for Ranking
Nders-C-R2.txt	Added Pattern_idf for Ranking
Nders-C-R3.txt	Added RandomWalk for Ranking(R4+Ranking)
Nders-C-R4.txt	Using LSI model as the component of topic similarity in our system
Nders-C-R5.txt	Using LDA model as the component of topic similarity in our system
PolyU-C-R1.txt	[retrieval with method1]
PolyU-C-R2.txt	[retrieval with method2]
PolyU-C-R3.txt	[retrieval with method3]
PolyU-C-R4.txt	[retrieval with method4]
rucir-C-R1.txt	using word2vec, IDF and Euclidean distance
rucir-C-R2.txt	using word2vec and cosine similarity
rucir-C-R3.txt	using word2vec, cosine similarity and IDF
SG01-C-R1.txt	deep sentence match, LTR, v1
SG01-C-R2.txt	deep sentence match, LTR, v2
SG01-C-R3.txt	deep sentence match, LTR, v3
SLSTC-C-R1.txt	retrieval
srcb-C-R1.txt	Search by Solr and rank by features.
srcb-C-R2.txt	Search by Solr and rank by features.
srcb-C-R3.txt	Search by Solr and rank by features.
srcb-C-R4.txt	Search by Solr and rank by features.
srcb-C-R5.txt	Search by common words and rank by multi-features.
TUA1-C-R1.txt	retrieval with doc2vec method
TUA1-C-R2.txt	retrieval method
TUA1-C-R3.txt	retrieval with RNN method
TUA1-C-R4.txt	retrieval with LSI method
UB-C-R1.txt	Baseline run of searching comments with BM25 ranking
UB-C-R2.txt	Reranking UB-C-R1 by applying rules based on sentimental words
UB-C-R3.txt	Reranking UB-C-R1 by utilizing information of post-comment pairs of those relevant posts retrieved with BM25
UB-C-R4.txt	Reranking UB-C-R1 by combining rules of UB-C-R2 and UB-C-R3 in one way
UB-C-R5.txt	Reranking UB-C-R1 by combining rules of UB-C-R2 and UB-C-R3 in another way
WIDM-C-R1.txt	[Use cosine Similarity to sort]
WIDM-C-R2.txt	[query with noun, verb, adjective of post and order by cosine Similarity]
WIDM-C-R3.txt	[rerank with SVMRank]
WUST-C-R1.txt	word2vecSim*VSM
WUST-C-R2.txt	word2vecSim+lcs+keyovelap+cluster

Table 19: SYSDESC fields of 56 Chinese generation-based runs. Note that not all are informative.

BUPTTeam-C-G1.txt	BUPTTeam run1
BUPTTeam-C-G2.txt	BUPTTeam run2
CIAL-C-G1.txt	[generation using seq to seq 2 layer LSTM plus attn, input pretrained embedding from all posts and comments]
ckip-C-G1.txt	test-out-embed-general
ckip-C-G2.txt	test-out-w2w-general
ckip-C-G3.txt	test-out-w2w-trigram
ckip-C-G4.txt	test-out-ps2cw
CYIII-C-G1.txt	Using 200k training data, and use part of speech (NVA)
CYIII-C-G2.txt	Using 200k training data, and use part of speech (NV)
CYIII-C-G3.txt	Using 200k training data, and use part of speech (N)
CYIII-C-G4.txt	Using 200k training data, and use part of speech (V)
CYIII-C-G5.txt	Using 200k training data, and use part of speech (NVA), use lstm cell
Gbot-C-G1.txt	Seq2seq-based method using dual learning
Gbot-C-G2.txt	Seq2seq-based method using reinforcement learning, the reward is pre-trained matching model score
Gbot-C-G3.txt	Seq2seq-based method using reinforcement learning, the reward is sentence similarity
Gbot-C-G4.txt	Standard seq2seq model with attention
Gbot-C-G5.txt	Generation-based method using Conditional Wasserstein GAN
iNLP-C-G1.txt	An RNN Model with Attention and VAE Decoder + 1st post topic, $Z_mean = Norm(0.0, 0.9)$
iNLP-C-G2.txt	An RNN Model with Attention and VAE Decoder + 1st cmnt topic, $Z_mean = Norm(0.0, 0.9)$
iNLP-C-G3.txt	VAE with $Z_mean = N(0.1, 0.8)$
iNLP-C-G4.txt	VAE with $Z_mean = N(0.0, 0.8)$
iNLP-C-G5.txt	An RNN Model with Attention and VAE Decoder + 2nd cmnt topic, $Z_mean = Norm(0.0, 0.9)$
MSRSC-C-G1.txt	[attention+emotion+rnnlm+filter_name_ads+diversity]
MSRSC-C-G2.txt	[attention+emotion+rnnlm+filter_name_ads]
MSRSC-C-G3.txt	[attention+emotion+rnnlm+filter_ads+diversity]
MSRSC-C-G4.txt	[attention+emotion+filter_ads+diversity]
MSRSC-C-G5.txt	[attention+filter_ads+diversity]
PolyU-C-G1.txt	[generation with vae]
PolyU-C-G2.txt	[generation with seq2seq and attention]
rucir-C-G1.txt	rank by post and post similarity, our model
rucir-C-G2.txt	rank by post and comt similarity, our model
rucir-C-G3.txt	no rank, our model
rucir-C-G4.txt	rank by post and comt similarity, pmi words only
rucir-C-G5.txt	rank by post and comt similarity, nrm model only local encoder
SG01-C-G1.txt	rerank-base-[vae-predata,seq2seq-2dataset]
SG01-C-G2.txt	rerank-base-[vae-predata]
SG01-C-G3.txt	rerank-base-[seq2seq-2dataset]
SG01-C-G4.txt	origin-seq2seq-2dataset
SG01-C-G5.txt	origin-vae-pre-data
SMIPG-C-G1.txt	We use a very simple model with a single GRU as encoder and a single GRU with attention as decoder, and rerank candidates use beam search.
splab-C-G1.txt	[NVA_long.result]
splab-C-G2.txt	[wlmm.txt]
splab-C-G3.txt	[sys_merge_rescore.txt]
splab-C-G4.txt	[attn̊result]
splab-C-G5.txt	[NVA_long_fullset.result]
srcb-C-G1.txt	Seq2seq model was used to generate results
srcb-C-G2.txt	Seq2seq model was used to generate results
srcb-C-G3.txt	Seq2seq model was used to generate results
srcb-C-G4.txt	word-based share embedding
srcb-C-G5.txt	char-based share embedding
TUA1-C-G1.txt	Generation-based Comments by RNN-ranking
TUA1-C-G2.txt	Generation-based Comments with Beam-search by RNN-ranking
TUA1-C-G3.txt	Generation-based Comments with and without Beam-search by RNN-ranking
TUA1-C-G4.txt	Generation-based Comments by RNN+COS-ranking
TUA1-C-G5.txt	Generation-based Comments with Beam-search by RNN+COS-ranking
WIDM-C-G1.txt	[NULL]