Overview of the Recognizing Inference in Text (RITE-2) at NTCIR-10

Yotaro Watanabe
Tohoku University, Japan
yotaro-w@ecei.tohoku.ac.jp

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

Junta Mizuno
Tohoku University, Japan
junta-m@ecei.tohoku.ac.jp

Tomohide Shibata
Kyoto University, Japan
shibata@i.kyoto-u.ac.jp

Hiroshi Kanayama
IBM Research – Tokyo, Japan
HKANA@jp.ibm.com

Cheng-Wei Lee
Academia Sinica, Taiwan
aska@iis.sinica.edu.tw

Chuan-Jie Lin
National Taiwan Ocean University, Taiwan
cjlin@mail.ntou.edu.tw

Shuming Shi
Microsoft Research Asia, P.R.China
shumings@microsoft.com

Teruko Mitamura
Carnegie Mellon University, U.S.A
teruko+@cs.cmu.edu

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

Hideki Shima
Carnegie Mellon University, U.S.A
hideki@cs.cmu.edu

Kohichi Takeda
IBM Research – Tokyo, Japan
TAKEDASU@jp.ibm.com

ABSTRACT

This paper describes an overview of RITE-2 (Recognizing Inference in TExt) task in NTCIR-10. We evaluated systems that automatically recognize semantic relations between sentences such as paraphrase, entailment, contradiction in Japanese, Simplified Chinese and Traditional Chinese. The tasks in RITE-2 are Binary Classification of entailment (BC Subtask), Multi-Class Classification including paraphrase and contradiction (MC Subtask), Entrance Exam Subtasks (Exam BC and Exam Search), Unit Test, and RITE4QA Subtask. We had 28 active participants, and received 215 formal runs (110 Japanese runs, 53 Traditional Chinese runs, 52 Simplified Chinese runs). This paper also describes how the datasets for RITE-2 had been developed, how the systems were evaluated, and reports RITE-2 formal run results.

Keywords

entailment, contradiction, entrance exam, test collections, evaluation

1. INTRODUCTION

Understanding meaning of texts by computers is crucially important to establish advanced information access technologies. Since the number of documents in the Web is rapidly increasing, efficiently finding necessary information from the Web has become quite difficult. However, if deep understanding of texts by computers establishes, it makes it possible to automatically collect only necessary information or organizing the vast amount of information in the Web. One of the promising technologies toward understanding meaning of texts by computers is textual entailment recognition which has attracted the attention of many researchers in recent decades. The task of recognizing textual entailment is, given a pair of texts t_1 and t_2, to recognize whether t_1 entails t_2, in other words, a human reading t_1 would infer that t_2 is most likely true [3]. This technology can be applied for various information access technologies such as Question Answering, Document Summarization, Information Retrieval, etc. In question answering, answers of questions can be detected based on semantic relatedness while absorbing surface difference of texts (e.g. [6]). In document summarization, we can remain necessary information (e.g. [31]) by filtering redundant texts using RTE technologies.

Textual entailment recognition task has attracted the attention of many researchers in recent decades, especially the community of Recognizing Textual Entailment (RTE) [3]. From the third PASCAL RTE challenge (RTE-3), the task has included recognizing not only entailment relations, but also contradiction relations [5]. In the RTE6, the task setting was changed to a more realistic scenario: given a corpus and a set of *candidate* sentences retrieved by a search engine from that corpus, systems are required to identify all the sentences from among the candidate sentences that entail a given hypothesis. The setting of cross/multi-lingual entailment relation recognition has also been explored in [14, 13, 18]. In the SemEval-2012 Cross-lingual Textual Entailment (CLTE) [17], the dataset which consists of text pairs in Spanish-English, Italian-English, French-English, German-English were used in evaluation.

RITE (Recognizing Inference in Text), the first task of evaluating systems which recognize semantic relations between sentences for Japanese and Chinese, was organized in NTCIR-9. The RITE task consists of the four subtasks: Binary-Class (BC) subtask, Multi-Class (MC) subtask, Entrance Exam subtask and RITE4QA subtask. In the BC subtask, given a text pair t_1 and t_2, a system identifies whether t_1 can be inferred from t_2 (i.e. t_1 entails t_2) or not. In the MC subtask, a system is required to recognize not only entailment but also paraphrase and contradiction. The Entrance Exam subtask is similar to the BC subtask, however the dataset for the task was developed from past

Japanese National Center Test for University Admissions. In the RITE4QA subtask, the dataset was developed from Factoid Question Answering datasets. The NTCIR-9 RITE task achieved a great success with the highest number of participants (24 participants) among the NTCIR-9 tasks, however, there are still room to explore. (1) The results submitted by the RITE participants were not enough. Actually, the highest accuracy of the BC subtask for Japanese was only 58%. (2) Also, there are not enough studies on the effects of various linguistic phenomena which affect semantic relations between sentences. In order to tackle such issues, it is necessary to continue to explore the task of entailment relation recognition.

In the NTCIR-10 RITE-2 task, in addition to the four subtasks in NTCIR-9 RITE (BC, MC, ExamBC and RITE4QA), the two new subtasks were added: Exam Search subtask and UnitTest subtask. In the Exam Search subtask, instead of a text t_1, a set of documents are given to systems. Systems are required to search a set of texts in the documents which entails or contradicts t_2. In the UnitTest subtasks, the set of examples were developed by providing a breakdown of linguistic phenomena that are necessary for recognizing relations between t_1 and t_2 in the dataset for the BC subtask. Also, the setting of the MC subtask was slightly changed. We removed backward-entailment relation included in NTCIR-9 RITE, resulting in the MC subtask as a four-way classification problem. Since back-ward-entailment can be recognized by flipping t_1 and t_2 and checking whether the pair has a forward-entailment relation or not. In order to make it easier for people to participate in the subtasks, we provided (1) a baseline tool that can be modified easily, (2) linguistic analysis results for all of the subtasks and the search results for the Entrance Exam Search subtask.

The important dates for NTCIR-10 RITE-2 were as follows:

2012/07/05	Development data released
2012/07/26	Evaluation tool released
2012/08/31	Registration due
2012/09/06	Baseline tool released
2012/10/03	Textbook and Wikipedia search results for Entrance Exam Search subtask released
2013/01/09	Formal run data released
2013/01/16	Formal run results submission due
2013/01/22	Results of the formal run released
2013/01/31	Formal run unofficial results submission due
2013/02/04	Formal run data with the correct labels released

This paper is organized as follows. At first we describe the RITE-2 subtasks (Section 2), and our organization effort (Section 3). Then we report RITE-2 active participants (Section 4), and the formal run results (Section 5). Finally we conclude in Section 6. [34]

2. TASK OVERVIEW

This section describes the subtasks in RITE-2 and how the dataset for each subtask was developed.

2.1 BC Subtask

The BC subtask is, given a pair of sentences t_1 and t_2, to recognize whether t_2 can be inferred from t_1 (i.e. t_1 entails t_2). The BC subtask dataset for Japanese was developed by the 16 college students (from 1st year under-graduate students to 1st year Ph.D students). In order to make text pairs, candidate sentences are extracted from Wikipedia using an off-the-shelf search engine. Topics of sentences are selected so as not to be impartial. We allowed the annotators to edit extracted sentence pairs if some modification is necessary to create a pair of sentences which has an intended semantic relation. However, the number of modification was minimized as much as possible. After the sentence modification, we performed data filtering. A pair of sentences was removed if similar examples are already developed, it requires a special knowledge of particular fields, it includes interrogative sentence, it includes subjective information, etc. After that, five annotators labeled semantic relations for each pair, and the pair was accepted if four or more annotators labeled the same semantic relation.

In terms of Simplified and Traditional Chinese BC dataset, there are one type of development datasets, *-DevBC, and two types of the formal run datasets, i.e. *-BC and *-ArtificialBC. They are converted from their corresponding MC dataset by replacing forward-entailment (F) and bi-directional entailment (B) labels with Y labels, and contradiction (C) and independence (I) labels with N labels. We describe how the dataset for the MC subtask was developed in Section 2.2.

The *-ArtificialBC and *-DevBC datasets are combined as single files named as *-ExtraBC to be distributed to the participants. The pair IDs in 1 ˜ 1321 are assigned to DevBC pairs, while 1322 ˜ 1894 are assigned to ArtificialBC pairs.

BC	Y	N	Total
JA (dev)	240	371	611
JA (test)	256	354	610
CT (dev)	716	605	1321
CT (test)	479	402	881
CT (arti)	341	232	573
CS (dev)	528	286	814
CS (test)	422	359	781
CS (arti)	341	232	573

Table 1: Statistics of the BC dataset.

The statistics of the BC dataset are shown in Table 1.

In the BC subtask, systems were evaluated by macro-F1 score which is defined by

$$macroF1 = \frac{1}{|\mathcal{C}|}\sum_{c \in \mathcal{C}} F1_c = \frac{1}{|\mathcal{C}|}\sum_{c} \frac{2 \times Prec._c \times Rec._c}{Prec._c + Rec._c} \quad (1)$$

where $\mathcal{C}$ is the set of classes and $Prec._c$ and $Rec._c$ is a precision value and a recall value for the class c. Precision and recall are defined as follows.

$$Precision = \frac{N_{correct}}{N_{predicted}} \quad (2)$$

$$Recall = \frac{N_{correct}}{N_{target}} \quad (3)$$

2.2 MC Subtask

MC subtask is a 4-way labeling subtask to detect (forward / bi-directional) entailment or no entailment (contradiction

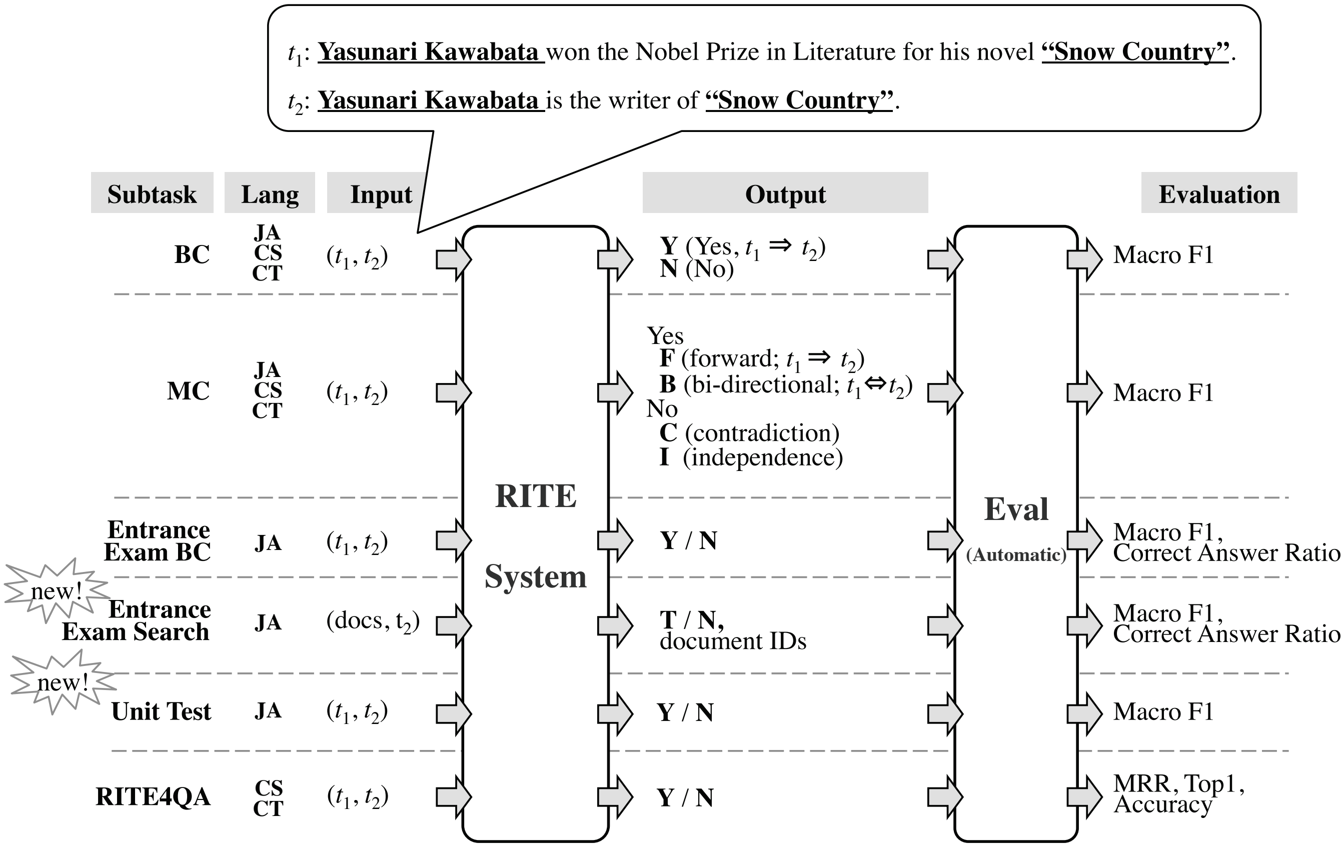

Figure 1: An overview of the RITE-2 subtasks.

/ independence) in a text pair. The four types of relations are defined as follows.

Bi-directional Entailment (Paraphrase) (B): t_1 entails t_2 AND t_2 entails t_1.

Forward Entailment (F): t_1 entails t_2 AND t_2 does not entail t_1.

Contradiction (C): t_1 and t_2 contradicts, or cannot be true at the same time.

Independence (I): otherwise.

In NTCIR-9 RITE, backward-entailment was included in the set of semantic relations. However, backward-entailment can be detected by checking whether the flipped pair holds forward-entailment (i.e. t_1 can be inferred from t_2) or not. So, we decided to exclude backward-entailment relation from the set of semantic relation used in the MC subtask.

The dataset of MC subtask for Japanese were developed as the same as the BC subtask. The sentence pairs were created from Wikipedia, and the pairs were accepted if four or more annotators labeled the same semantic relation. In terms of contradiction relation, we created a set of examples so as to include diverse linguistic phenomena such as negation, factualities, antonyms, numerical mismatch, existence/non-existence, condition mismatch, relation mismatch, definition mismatch, etc. The examples are randomly divided into the development data and the test data.

In terms of Chinese datasets, two development datasets, *-DevMC, were created from merging their corresponding RITE-1 development datasets and formal run datasets.

The formal run dataset of MC subtask for Chinese was developed from two sources and two ways. In terms of sources, pairs are created based on Wikipedia sentences and information retrieval results. We randomly choose Wikipedia pages for annotators to create pairs. A Wikipedia page is first analyzed by an annotator to pick up sentences with important information to the subject of the page. More sentences are collected for pair creation by searching the web with keywords of the Wikipedia sentence.

In the formal run datasets, some special pairs are created and collected for the investigation of two research issues: (a) How to improve the discriminability of textual entailment datasets? (b) Are there any notable performance differences between the development and formal run datasets? For issue (a), instead of only use sentences from the source data, we proposed an idea to create more variation pairs from a source pair in order to lower the possibility of guessing a label right. In other words, a system which only favors certain type of pairs won't get all the variations correct easily. We call these variations "artificial pairs" and put pairs agreed by four annotators into the CT-MC dataset. Unlike traditional approaches, we also collected pairs with low agreement to form an extra dataset called CT-ArtificialMC, hope this dataset can still be useful to discriminate RITE systems. As for the second research issue, the development dataset CT-DevMC is also used as formal run data for further analysis.

The CT-ArtificialMC and CT-DevMC datasets are combined as a single file named as CT-ExtraMC to be distributed to the participants. The pair IDs in 1 ˜ 1321 are assigned to DevMC pairs, while 1322 ˜ 1894 are assigned to ArtificialMC pairs.

After CT datasets are created, they are manually converted to CS datasets, except that CS-ArtificialMC is created by Google translating the CT-ArtificialMC dataset. In the conversion, some improper pairs are removed, which results in less pairs in CS datasets.

Unfortunately, 5 ArtificialMC pairs have “R” labels which should not be included this year. Their pair IDs are 1769, 1792, 1793, 1794, and 1796. They are removed before evaluation.

MC	B	F	C	I	Total
JA (dev)	83	207	65	193	548
JA (test)	70	205	61	212	548
CT (dev)	262	544	254	261	1321
CT (test)	151	328	114	288	881
CT (arti)	186	155	115	112	568
CS (dev)	159	369	146	140	814
CS (test)	145	277	106	253	781
CS (arti)	186	155	115	112	568

Table 2: Statistics of the MC dataset.

The statistics of the MC dataset are shown in Table 2.

In the MC subtask, we evaluate systems with macroF1 so as to treat all of the semantic relations equally. Contradiction recognition is an important problem in RTE, but the number of contradiction instances is less than the other relations. By replacing accuracy to macro-F1, minor relations are given the same weight with the other relations.

2.3 Entrance Exam Subtasks

This subtask aims to answer multiple-choice questions of real university entrance exams, by referring to textual knowledge i.e. Wikipedia and textbooks. This is an attempt to emulate human's process to answer entrance exam questions as the RITE task. This is an interesting challenge that use real entrance exams for the evaluation of intelligent systems.

This subtask provides two types of data.

BC style (ExamBC) The data is provided in the same form as the BC subtask. Systems are asked to recognize inference relations between t_1 and t_2. In this data, t_1 is extracted from Wikipedia, while t_2 is taken from actual questions of the Center Test for University Admissions in Japan.

Search style (ExamSearch) In this data, t_1 is not given. Systems are asked to retrieve texts that can be used as t_1 from Wikipedia or textbooks, and answer whether t_2 is entailed (inferred) from retrieved texts.

The ExamSearch subtask is closer to the actual situation of answering entrance exams, while it is much harder than the ExamBC subtask.

ExamBC and ExamSearch data files include exactly same t_2 sentences, and their IDs are common. Therefore, you can use the ExamBC dataset for the development of retrieval systems for ExamSearch. We also provide document IDs of texts that are retrieved by human annotators as candidates for t_1 (in the formal run this data will not be provided). In the Entrance Exam subtasks, systems are required to provide a confidence score for each decision.

ExamBC	Y	N	Total
dev	210	300	510
test	173	275	448
ExamSearch	Y	N	Total
dev	210	300	510
test	173	275	448

Table 3: Statistics of the Entrance Exam BC and Exam Search dataset.

The statistics of the Entrance Exam BC subtask and Exam Search subtask are shown in Table 3.

In the Entrance Exam subtasks, we evaluate systems with macro-F1 scores in the same way as the BC/MC subtasks, as well as correct answer ratios for multiple-choice questions. In the latter evaluation, Y/N labels are mapped into selection of answers for the original questions according to the confidence scores outputted by systems, and the correct answer ratio is measured. In the ExamSearch subtask, we also evaluate precision/recall of t_1 search results that are retrieved from Wikipedia or textbooks by the systems.

2.4 Unit Test

For recognizing inference in text, various kinds of semantic/contextual processing are necessary. While the RITE task aims at such integrated semantic/context processing systems, it also has a problem that research focused on specific linguistic phenomena is not easy to pursue.

The unit test subtask provides a data set that includes a breakdown of linguistic phenomena that are necessary for recognizing relations between t_1 and t_2. Sentence pairs are sampled from the BC subtask data, and several sentence pairs are created for each sample so that only one linguistic phenomenon appears in each pair.

The unit test data corresponds to a subset of the BC task data. The data is small, but you might want to use it for various research including the analysis of linguistic issues that appear in the RITE data, the evaluation of recognition accuracy for each phenomenon, and the development/training of a recognizer for each linguistic phenomenon.

UnitTest	Y	N	Total
dev	239	33	272
test	212	29	241

Table 4: Statistics of the UnitTest dataset.

The statistics of the UnitTest dataset are shown in Table 4 and the categories of linguistic phenomena included in the Unit Test data are shown in Table 5.

2.5 RITE4QA Subtask

RITE-2 RITE4QA data creation and evaluation are different from NTCIR-9 RITE. In NTCIR-9 RITE, RITE4QA pairs were created by selecting answer passages similar to the question sentence. Such an approach may not be able to show the real performance of a RITE system in a factoid QA setting because answer-bearing sentences are not always

		dev	test
lexical	synonymy	10	10
	hypernymy	6	3
	meronymy	1	1
	entailment	1	0
phrase	synonymy	45	35
	hypernymy	3	0
	entailment	28	45
case alternation		9	7
modifier		30	42
nominalization		2	1
coreference		12	4
clause		29	14
relative clause		10	8
transparent head		2	1
list		11	3
quantity		1	0
scrambling		16	15
inference		4	2
implicit relation		10	18
apposition		3	1
temporal		2	1
spatial		4	1
disagree	lexical	5	2
	phrase	25	25
	modality	2	1
	spatial	1	1
	temporal	0	1
Total		272	241

Table 5: The categories of linguistic phenomena in the UnitTest data.

similar to the question sentence. In order to simulate a more realistic QA situation, we decided to include all the possible answer-bearing sentences to create RITE4QA pairs. We call the resulting dataset CT-CompleteRITE4QA.

Since NTCIR-7 and NTCIR-8 CLQA questions were used in NTCIR-9 RITE, we choose a new factoid QA dataset composed of 150 factoid questions for RITE-2. The final CT-CompleteRITE4QA consists of more than 7,000 pairs. Considering some systems may not be able to handle such a large amount of pairs in the formal run period, CT-CompleteRITE4QA is further divided into CT-RITE4QA and CT-OptionalRITE4QA. Pairs are ranked based on the original answer scores from the QA system. Only pairs of top 20 answers of each question are included in CT-RITE4QA. Others are put in CT-OptionalRITE4QA.

Since RITE4QA is regarded as an extrinsic evaluation, it would be better to use factoid QA evaluation metrics instead of textual entailment metrics. In order to do that, each RITE-2 run is converted to a factoid QA run, which is a ranking of answers with answer-bearing sentences, based on an original QA ranking (SrcRank) from a QA system. The new ranking, RiteRank, is primarily based on the confidence scores of Y labels. If there are Y pairs with the same confidence score, it falls back to the rank in SrcRank. In other words, a "Naive Run", which is a RITE4QA run with all the pairs labeled as "Y" in the same confidence, will result in a RiteRank that is identical to the SrcRank. Two SrcRanks are created for RITE4QA evaluation:

- BetterRanking: This ranking is produced from a good QA system.
- WorseRanking: The reverse ranking of BetterRanking. It is a simulated worse QA result.

Factoid QA evaluation metrics are reported against WorseRanking and BetterRanking respectively for each RITE4QA run. In the evaluation reports, BetterRanking scores show how good a RITE system is in terms of the improvement on the answer ranking of a good-performing factoid QA system, while WorseRanking scores show how good a RITE system is when it is applied to the answer ranking of a bad-performing factoid QA system.

Three factoid QA evaluation metrics are shown in a RITE-4QA report:

- Top1 accuracy: is the rate of questions which top 1 answers are correct.
- MRR: is the average reciprocal rank (1/n) of the highest rank n of a correct answer for each question.
- Top5 accuracy: shows the rate at which at least one correct answer is included in the top 5 answers.

3. TASK ORGANIZATION EFFORTS

3.1 Baseline System

In NTCIR-9 RITE, Shima et al. provided a textual entailment recognition system[1] to make participation for the RITE tasks easy. The system judges textual entailment based on a character overlap ratio for the binary-class subtasks. For the MC subtask, the system determines entailment directions (B, F and R) based on a character overlap. If the pair does not hold entailment relation, then the system randomly assigns either C or I.

In NTCIR-10 RITE-2, we provided a machine learning-based baseline system written in Python which is one of the popular programming languages in Natural Language Processing area. In the baseline system, averaged perceptron can be used to train a model. Also, the ML algorithms implemented in Classias [21] can be used.

The supported algorithms are as follows.

- Averaged Perceptron
- L1-regularized logistic regression
- L2-regularized logistic regression
- L1-regularized L1-loss SVM
- L2-regularized L1-loss SVM

The baseline system supports both the binary-class subtasks ("JA-BC", "JA-ExamBC" and "JA-UnitTest") and the multi-class subtask ("JA-MC"). For the binary-class subtasks, the system outputs "Y" or "N" by performing a binary classification. For the MC subtask, the system at first selects a class from the three classes "F", "I" and "C". If the system classified the pair as "F", then the system checks whether the flipped pair of sentences also holds forward-entailment ("F"). If the system outputs "F", then the pair is classified as "B" and "F" otherwise.

Initial features implemented in the system are listed as follows. The baseline system aligns two words or two chunks (a unit that consists of one or more content words and zero

[1] `http://code.google.com/p/rite-sdk/`

```xml
<?xml version='1.0' encoding='UTF-8' standalone='no' ?>
<dataset type='bc'>
  <pair id='1' label='Y'>
    <t1>
      川端康成は「雪国」などの作品でノーベル文学賞を受賞した。
      <Sentence id="sample_t1" role="text" text="川端康成は「雪国」などの作品でノーベル文学賞を受賞した。">
        <Annotation tool="MeCab" ver="0.994"/>
        <Annotation tool="CaboCha" ver="0.64"/>
        <Annotation tool="UniDic" ver="1.3.12"/>
        <Chunk head="c4" id="c0" score="2.473067" type="D">
          <Token orig="川端" pos1="名詞" pos2="固有名詞" pos3="人名" pos4="姓" ... />
          <Token orig="康成" pos1="名詞" pos2="固有名詞" pos3="人名" pos4="名" ... />
          <Token orig="は" pos1="助詞" pos2="係助詞" ... />
        </Chunk>
        <Chunk head="c2" id="c1" score="1.368909" type="D">
          <Token orig="「" pos1="補助記号" pos2="括弧開" ... />
          <Token orig="雪国" pos1="名詞" pos2="普通名詞" ... />
          <Token orig="」" pos1="補助記号" pos2="括弧閉" ... />
          <Token orig="など" pos1="助詞" pos2="副助詞" ... />
          <Token orig="の" pos1="助詞" pos2="格助詞" ... />
        </Chunk>
        ...
      </Sentence>
    </t1>
    <t2>
      川端康成「雪国」の著者である。
      <Sentence id="sample_t2" role="hypothesis" text="川端康成「雪国」の著者である。">
        <Annotation tool="MeCab" ver="0.994"/>
        <Annotation tool="CaboCha" ver="0.64"/>
        <Annotation tool="UniDic" ver="1.3.12"/>
        <Chunk head="c1" id="c0" score="1.141424" type="D">
          <Token orig="川端" pos1="名詞" pos2="固有名詞" pos3="人名" pos4="姓" ... />
        ...
        </Chunk>
      </Sentence>
    </t2>
  </pair>
</dataset>
```


Figure 2: An example of XML file containing a linguistic analysis result.


```xml
<?xml version="1.0" encoding="UTF-8"?>
  <pair id="1" label="Y">
    <t2>最初にペリーが来航してから翌年再来航するまでに，老中阿部正弘が諸大名に対し，アメリカ大統領国書への対応についての意見を求めた。</t2>
  <ResultSet firstResultPosition="1" totalResultsReturned="5">
  <Result Rank="1" Id="0000105917" OrigId="310335">
    <Title>海岸防禦御用掛</Title>
    <RawStrings>
      <RawString Sid="8" Score="113.756">阿部は将軍を中心とした譜代大名・旗本らによる独裁体制の慣例を破り、水戸藩主徳川斉昭を海防参与に推戴
し、...</RawString>
      <RawString Sid="7" Score="75.168">ペリー来航当時、時の将軍徳川家慶は死の床にあり、国家の一大事に際して執政をとるなど適わない状態であった。
</RawString>
      <RawString Sid="20" Score="74.593">このような諸大名・諸藩の藩士をもおおいに幕政に参画させた政治手法は、結果として諸大名や朝廷が中央政治に進
出する足がかりをつくることとなったといわれ、...</RawString>
      <RawString Sid="19" Score="73.279">徳川斉昭以下、海防掛は海防のあり方について積極的に献策を行ったが、翌年、阿部に代わり老中首座となった堀田
正睦が中心となって...</RawString>
      <RawString Sid="6" Score="67.513">これに伴い、老中首座の阿部正弘らが中心となって幕府として海防のあり方を検討するために設けられた。</RawString>
    </RawStrings>
  </Result>
  <Result Rank="2" Id="0000736812" OrigId="1409177">
    <Title>安政の改革</Title>
...
```


Figure 3: An example of XML file containing a search result.

or more functional words) if two words or two chunks are exactly the same.

The set of features are used to provide the baseline results in Section 5. The users can easily modify the set of features used in the system.

1. bag of content words
2. ratio of aligned content words
3. bag of aligned chunks
4. ratio of aligned chunks
5. bag of aligned head words for each chunks

3.2 Providing Preprocessed Data

The recognition of paraphrase, entailment and contradiction relations requires a wide variety of linguistic analysis and knowledge. The list of available linguistic analysis and knowledge has been provided to the participants on the Web page in the same manner as RITE-1. In addition, we provide the participants with the linguistic analysis results and the search results for Entrance Exam Search subtask.

Linguistic analysis results.

For those who are unfamiliar with Natural Language Processing, such as graduate students and researchers who specialize in another area, it may be difficult to install a linguistic analysis tool, such as morphological analyzer and parser, and obtain some information from its analysis result. Thus, we provide the participants with the linguistic analysis result of development/test data, which avoids the necessity of installing linguistic analysis tools on their machine. The following two sets of Japanese linguistic analysis results are provided:

- morphological analyzer MeCab[2] and parser CaboCha[3]
- morphological analyzer JUMAN[4] and parser KNP[5]

The output format usually varies according to analysis tools, which requires some preprocessings for each analysis tool, and thus, we design a uniform XML format representing the linguistic analysis result. Figure 2 shows an example. The linguistic results of the above two sets of linguistic tools are represented by the same tag set.

Search results for Entrance Exam Search subtask.

In the Entrance Exam Search Subtask, to determine the correctness of statement t2, some relevant sentences with t2 need to be retrieved from a textbase. As a text collection, we provide textbooks and Wikipedia for the participants.

It is not so easy to set up a search engine, such as Apache Solr, to retrieve documents from the textbooks and Wikipedia. Therefore, we provide search results for the participants. We adopt TSUBAKI [29] as a search engine, and in the search results, for each t2, at most five search results from the textbook and Wikipedia are provided in the XML format. An example of XML file is shown in Figure 3.

4. PARTICIPANTS

	ID	Organization	Country /Region	JA	CS	CT
1	OKA1 [16]	Okayama University	Japan	√		
2	JUNLP [23]	Jadavpur University	India	√	√	√
3	JAIST [24]	Japan Advanced Institute of Science and Technology	Japan	√		
4	KitAi [28]	Kyushu Institute of Technology	Japan	√		
5	IASL [27]	Academia Sinica	Taiwan		√	√
6	MIG [8]	National Chengchi University	Taiwan		√	√
7	SKL [7]	Nagoya University	Japan	√		
8	CYUT [36]	Chaoyang University of Technology	Taiwan		√	√
9	WSD [38]	Waseda University	Japan	√		
10	Yuntech [9]	National Yunlin University of Science and Technology	Taiwan		√	√
11	IMTKU [4]	TamKang University	Taiwan		√	√
12	KC99 [2]	National Kaohsiung University of Applied Sciences	Taiwan			√
13	bcNLP [33]	Shanghai Jiao Tong University	China		√	
14	WUST [11]	Wuhan University of Science and Technology	China		√	
15	NTTD [19]	NTT DATA Corporation	Japan	√		
16	NTOUA [10]	National Taiwan Ocean University	Taiwan			√
17	DCUMT [22]	Dublin City University	Ireland	√		
18	EHIME [30]	Ehime University	Japan	√		
19	KDR [1]	NEC Corporation	Japan	√		
20	FLL [12]	Fujitsu Laboratories Ltd.	Japan	√		
21	BnO [32]	National Institute of Informatics	Japan	√		
22	THK [35]	Tohoku University	Japan	√		
23	KYOTO [26]	Kyoto University	Japan	√		
24	IBM [20]	IBM Japan, Ltd.	Japan	√		
25	TKDDI [15]	Tohoku University and KDDI R&D Laboratories	Japan	√		
26	ut12	The University of Tokyo	Japan	√		
27	WHUTE [25]	Wuhan University	China		√	√
28	MCU [37]	Ming-Chuan University	Taiwan			√

Table 6: The participants in RITE-2.

Subtask	JA	CT	CS	Total
BC	41	20	21	82
MC	20	21	21	62
ExamBC	31	–	–	31
ExamSearch	4	–	–	4
UnitTest	14	–	–	14
RITE4QA	–	12	10	22
Total	110	53	52	215

Table 7: Number of Submissions (w/o unofficial results).

Table 6 lists the RITE-2 participants. The participants were from Japan (15), Taiwan (8), China (3), India (1) and

[2] `http://mecab.googlecode.com/svn/trunk/mecab/doc/`

Team	OKA1	JUNLP	JAIST	KitAi	IASL	MIG	SKL	CYUT	WSD
Approach	hybrid	statistical (SVM)	statistical	statistical (SVM)	rule	rule	rule	statistical	hybrid
Overlap	*	*	*	*	*	*	*	*	*
Alignment		*	*	*	*			*	
Transformation					*				
Char/Word Overlap	*	*	*	*	*	*	*	*	
Syntactic		*	*	*	*	*		*	
Predicate-Argument					*				*
Named Entity		*	*		*	*	*		*
Entity/Event Rel.		*			*			*	
Temporal/Numeric					*		*	*	*
Entailment Rel.		*	*						*
Modality									*
Polarity			*						
Synonym/Antonym		*	*	*	*	*		*	*
Hypernym/Hyponym		*	*	*	*				
Meronym/Holonym			*		*				
Entity/Event rels					*				
Entailment rules		*	*						*
Resources	Japanese WordNet		Japanese WordNet, WordNet, ALAGIN Ent. DB, Weighted Pol. word list	Japanese WordNet, Nihongo Goi-Taikei	E-Hownet, Chinese WordNet			Wikipedia	Japanese WordNet
Tools	ChaSen	Google Translator	Google Translator, Stanford CoreNLP, CaboCha				Juman		
Provided info. used				knp-morph, syn					

TEAM	Yuntech	IMTKU	KC99	bcNLP	WUST	NTTD	NTOUA	DCUMT	EHIME
Approach	statistical	statistical	statistical	statistical	statistical	rule	statistical (SVM)	hybrid	hybrid (MLN)
Overlap		*	*	*	*		*		
Alignment			*	*			*	*	*
Transformation						*		*	
Char/Word Overlap		*	*	*	*	*	*	*	
Syntactic	*	*	*	*	*			*	
Predicate-Argument	*					*		*	
Named Entity	*			*			*	*	
Entity/Event Rel.	*	*						*	
Temporal/Numeric	*	*	*	*			*	*	
Entailment Rel.					*			*	
Modality									
Polarity									
Synonym/Antonym		*	*	*	*	*	*	*	
Hypernym/Hyponym				*	*		*	*	*
Meronym/Holonym									
Entity/Event Rel								*	
Entailment rules					*			*	
Resources		Chinese WordNet, TongYiCi CiLin	Wikipedia	HowNet, TongYiCi CiLin, negative words, antonyms	HIT synonym forest, Hownet, Negative words, Antonyms	ALAGIN synonym verbs, NICT synonym words	WordNet, Chinese WordNet, Wikipedia	Japanese WordNet, Wikipedia, Google	Wikipedia
Tools	Stanford Parser, CKIP Word Segmentation	Stanford Parser	Google Translator	BaseSeg, Stanford factored parser, BaseNER					Juman, KNP
Provided info. used								knp-morph, syn, pas	

TEAM	KDR	FLL	BnO	THK	KYOTO	IBM	TKDDI	WHUTE	MCU
Approach	statistical	statistical	statistical	hybrid	hybrid	hybrid	rule	rule	statistical
Overlap	*	*	*			*	*	*	*
Alignment	*	*	*	*	*	*	*		
Transformation			*	*				*	
Char/Word Overlap	*	*	*	*		*	*	*	*
Syntactic	*	*	*	*			*	*	
Predicate-Argument	*	*	*		*	*	*	*	
Named Entity	*	*	*			*		*	
Entity/Event Rel.					*		*		
Temporal/Numeric	*	*	*		*	*		*	*
Entailment Rel.				*	*		*		
Modality									
Polarity								*	
Synonym/Antonym	*	*	*	*	*	*	*		*
Hypernym/Hyponym	*	*	*	*	*	*	*		*
Meronym/Holonym	*								
Entity/Event Rels									
Entailment Rules		*	*	*	*		*		
Resources	Japanese WordNet, Wikipedia, Bunruigoihyo, Kougaku Thesaurus	English-Japanese dictionary, Japanese WordNet, ALAGIN Ent.DB	BunruiGoiHyo, Wikipedia, Nihongo Goi-Taikei, Antonym (Kojien), ALAGIN Ent.DB, Tsutsuji Func. Words	Japanese WordNet, Hype.-Hypo. Rels (Wikipedia), ALAGIN Ent.DB, IwanamiDic.	Wikipedia, Japanese dic., Web Corpus	Temporal expression knowledge, ontology (Wikipedia)	Japanese WordNet, Wikipedia, ALAGIN Ent.DB, IwanamiDic, ALAGIN allographic db	CIBA HANYU, TongYiCi CiLin	
Tools	CaboCha, SynCha, normalizeNumExp	MeCab, CaboCha, SynCha, Japanese NER, LDA, normalizeNumexp	MeCab, Cabocha, SynCha, normalizeNumexp	MeCab, CaboCha		In-house syntactic parser		Stanford Word Segmenter, Stanford POS Tagger, ICTCLAS, numeral normalizer	
Provided info. used					knp morph, syn, pas				

Table 8: A summary of the participant's systems.

Ireland (1), and 28 groups in Total. This was 4 more groups than that of NTCIR-9 RITE (24 groups).

Table 7 shows the number of the submitted runs in the RITE-2 formal run. "Unofficial results" denote runs submitted after the submission deadline of the formal run. Compared to NTCIR-9 RITE, the number of submissions for Chinese decreased, however, many participants were attended the subtasks for Japanese. So, the total number of runs were almost the same as NTCIR-9 RITE (212). In the subtasks for Japanese, the most common subtasks was BC subtask (41 runs). In the subtasks for Traditional and Simplified Chinese, the number of submitted runs for BC and MC was almost the same.

Table 8 shows the details of participant's systems including fundamental approaches, used information, used resources, and tools.

The half of the systems can be categorized into statistical approaches (50%), and the rest of the approaches are hybrid (27%) and rule-based (23%). The fundamental approaches used in the participant's systems were overlap-based (77%), alignment-based (63%) and transformation-based (23%). Since some systems include multiple strategies, the total of these ratio exceeds 100%.

The types of information used in the participant's systems are character/word overlap (85%), syntactic information (67%), temporal and numeric information (63%), named entity information (56%), predicate-argument structure (44%), entailment relations (30%), polarity information (7%), and modality information (4%). The types of resources used in the participant's systems are synonym-antonym (81%) hypernym-hyponym (63%), entailment rules (37%), meronym-holonym (11%), entity-event relations (7%). Diverse types of resources and tools were used in the participant's systems. The resources that more than one teams used are Wikipedia (10), Japanese WordNet (9), ALAGIN Entailment DB (5), Chinese WordNet (3), and TongYiCi CiLin (3), Nihongo Goi-Taikei (2), Bunruigoihyo (2), ALAGIN Entailment DB (2), Iwanami Dictionary (2), HowNet (2), WordNet (2).

Also, the tools that more than one teams used are Japanese predicate-argument structure analyzer SynCha (3), Japanese numeric expression normalizer normalizeNumExp (3), Google Translator (3), and Stanford NLP tools (CoreNLP, Segmenter, Parser, POS Tagger) (4).

5. FORMAL RUN RESULTS

In the formal run, participants could submit up to three runs for each subtask. The submission names in the tables follow the naming rule: (TeamID)-(Language)-(Subtask Name)-(Run Number). The run names marked by asterisks in the tables are unofficial results.

5.1 Results on BC subtask

Table 9, Table 10 and Table 11 show the results of BC subtask for Japanese, Simplified Chinese and Traditional Chinese respectively. In terms of the results on the Japanese data, the highest performance was achieved by DCUMT (80.49 in MacroF1 and 81.64 in accuracy). Compared to the previous RITE (NTCIR-9), the difference is noticeable. It is due to the data filtering described in Section 2.1. By conducting the strict filtering, almost of the noisy examples could be removed from the final version of the data. Regarding the Chinese subtasks, the top performances were achieved by bcNLP for the CS dataset, and by IASL for the CT dataset. These accuracies were almost the same as the previous RITE.

Table 12 and 13 show the results of ArtificialBC subtask for Simplified Chinese and Traditional Chinese respectively. Table 14 and 15 show the results of DevBC subtask for Simplified Chinese and Traditional Chinese respectively. This year we mainly focus on BC subtask results. The evaluations of ArtificialBC and DevBC runs are provided for additional observation.

5.2 Results on MC subtask

Table 16, Table 17 and Table 18 show the results of MC subtask for Japanese, Simplified Chinese and Traditional Chinese respectively. In terms of the results on the JA data, the top system achieved over 70% in accuracy which is higher than that of NTCIR-9 RITE (51%). We believe that the difference is due to the data filtering described in Section 2.2. Looking at the performances by the semantic relation types, SKL, WSD and FLL achieved approx. 70 in F1-value for bi-directional entailment recognition, SKL achieved over 75 in F1-value for forward-entailment, and approx. 30 in F1-value for contradiction recognition was achieved by THK.

Regarding the results on CS data, bcNLP achieved the top performances for all of the relations (56.82 of MacroF1), bi-directional (66.67 in F1), forward-entailment (67.86 in F1), and contradiction (38.41 in F1). When we include the unofficial results, the top performance for contradiction was achieved by IASL (38.42 in F1).

In terms of the results on the CT data, IASL achieved the top for all of the relations (46.32 of MacroF1) and contradiction (29.90 in F1). For bi-directional entailment, NTOUA achieved the best performance (62.07 in F1), and for forward-entailment, MCUIM achieved the best performance (70.07 in F1).

Table 19 and 20 show the results of ArtificialMC subtask for Simplified Chinese and Traditional Chinese respectively. Table 21 and 22 show the results of DevMC subtask for Simplified Chinese and Traditional Chinese respectively. This year we mainly focus on MC subtask results. The evaluations of ArtificialMC and DevMC runs are provided for additional observation.

5.3 Results on Exam BC subtask

Table 23 shows the results on Exam BC subtask. Since the accuracy of the top system was about 70% which is lower than that of the BC subtask, the difficulty of this dataset is higher than that of the BC subtask. Regarding correct answer ratio, the highest score was achieved by the BnO team and the score was 57.41.

5.4 Results on Exam Search subtask

Table 24 shows the results on Exam Search subtask. The number of teams participated in this subtask was only four probably due to the difficulty of handling the subtask. The accuracy of the top system 64.5% was lower than that of the other BC subtasks because systems are required to decide the truth of the propositions from search results which consist of multiple sentences in general. Although the t_2s are

index.html
[3]http://code.google.com/p/cabocha/
[4]http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN
[5]http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?KNP

Team	MacroF1	Acc.	Y-F1	Y-Prec.	Y-Rec.	N-F1	N-Prec.	N-Rec.
DCUMT-JA-BC-01	80.49	81.64	75.76	84.95	68.36	85.22	79.95	91.24
WSD-JA-BC-03	80.08	80.66	76.68	77.60	75.78	83.47	82.78	84.18
SKL-JA-BC-02	79.46	79.84	76.66	74.54	78.91	82.25	84.07	80.51
BnO-JA-BC-03	78.93	79.34	75.95	74.25	77.73	81.90	83.33	80.51
WSD-JA-BC-02	78.77	79.67	74.38	78.95	70.31	83.15	80.10	86.44
WSD-JA-BC-01	78.61	79.51	74.23	78.60	70.31	82.99	80.05	86.16
SKL-JA-BC-01	78.61	78.85	76.33	71.97	81.25	80.89	85.05	77.12
BnO-JA-BC-02	78.31	79.02	74.40	76.23	72.66	82.22	80.87	83.62
BnO-JA-BC-01	77.61	78.36	73.49	75.62	71.48	81.72	80.16	83.33
KitAi-JA-BC-01	77.11	77.70	73.44	73.44	73.44	80.79	80.79	80.79
OKA1-JA-BC-02	76.71	77.05	73.88	70.71	77.34	79.53	82.42	76.84
JAIST-JA-BC-02	76.47	76.89	73.35	71.06	75.78	79.59	81.60	77.68
SKL-JA-BC-03	76.40	77.21	72.03	74.27	69.92	80.77	79.13	82.49
KitAi-JA-BC-03	76.16	76.72	72.48	71.92	73.05	79.83	80.29	79.38
JAIST-JA-BC-01	75.56	76.23	71.51	71.94	71.09	79.61	79.27	79.94
OKA1-JA-BC-01	74.59	74.59	74.30	64.55	87.50	74.88	87.83	65.25
KYOTO-JA-BC-02	74.50	75.57	69.28	73.36	65.63	79.73	76.90	82.77
IBM-JA-BC-01	74.49	74.92	71.19	68.73	73.83	77.79	80.00	75.71
IBM-JA-BC-02	73.40	73.77	70.26	67.02	73.83	76.54	79.57	73.73
JAIST-JA-BC-03	73.08	73.77	68.75	68.75	68.75	77.40	77.40	77.40
IBM-JA-BC-03	72.90	73.44	69.08	67.54	70.70	76.72	78.07	75.42
KitAi-JA-BC-02	72.35	72.46	70.63	63.92	78.91	74.07	81.63	67.80
U-TOKYO-JA-BC-01	72.23	73.93	65.36	73.89	58.59	79.11	73.96	85.03
OKA1-JA-BC-03	71.71	72.13	68.28	65.36	71.48	75.15	77.88	72.60
FLL-JA-BC-03	67.99	70.00	59.96	68.16	53.52	76.02	70.90	81.92
*TKDDI-JA-BC-03	63.83	69.02	50.13	77.24	37.11	77.53	66.94	92.09
TKDDI-JA-BC-02	63.55	68.69	49.87	76.00	37.11	77.23	66.80	91.53
*TKDDI-JA-BC-02	63.55	68.69	49.87	76.00	37.11	77.23	66.80	91.53
*TKDDI-JA-BC-01	63.45	68.69	49.60	76.42	36.72	77.29	66.74	91.81
TKDDI-JA-BC-01	63.45	68.69	49.60	76.42	36.72	77.29	66.74	91.81
FLL-JA-BC-01	63.06	68.36	49.08	75.61	36.33	77.05	66.53	91.53
NTTD-JA-BC-03	61.90	62.30	58.03	54.45	62.11	65.77	69.50	62.43
*FLL-JA-BC-05	61.05	63.28	51.72	57.69	46.88	70.37	66.17	75.14
FLL-JA-BC-02	59.73	64.10	46.45	62.09	37.11	73.00	64.77	83.62
U-TOKYO-JA-BC-03	59.01	65.41	42.82	69.91	30.86	75.21	64.39	90.40
U-TOKYO-JA-BC-02	57.84	64.75	40.77	69.16	28.91	74.91	63.82	90.68
NTTD-JA-BC-01	57.59	64.10	40.97	66.09	29.69	74.20	63.64	88.98
*FLL-JA-BC-06	55.69	57.70	46.25	49.55	43.36	65.14	62.44	68.08
EHIME-JA-BC-01	54.34	59.34	39.22	52.63	31.25	69.46	61.57	79.66
*FLL-JA-BC-04	52.58	55.08	41.70	45.79	38.28	63.47	60.10	67.23
THK-JA-BC-01	52.40	53.28	45.92	44.65	47.27	58.87	60.18	57.63
NTTD-JA-BC-02	50.38	62.46	25.89	75.47	15.63	74.86	61.22	96.33
EHIME-JA-BC-02	50.14	51.48	41.96	42.13	41.80	58.31	58.15	58.47
JUNLP-JA-BC-01	48.83	49.02	45.72	41.32	51.17	51.93	57.34	47.46
EHIME-JA-BC-03	48.05	48.36	44.05	40.39	48.44	52.05	56.44	48.31
KYOTO-JA-BC-03	46.42	60.98	18.49	75.00	10.55	74.35	60.10	97.46
KYOTO-JA-BC-01	41.97	60.00	9.63	92.86	5.08	74.32	59.23	99.72
Baseline-JA-BC-01	62.53	63.93	55.28	57.63	53.13	69.78	67.91	71.75

Table 9: Results on BC subtask (JA).

Team	MacroF1	Acc.	Y-F1	Y-Prec.	Y-Rec.	N-F1	N-Prec.	N-Rec.
bcNLP-CS-BC-03	73.84	74.65	78.43	72.58	85.31	69.25	78.25	62.12
MIG-CS-BC-02	68.09	68.50	71.72	69.64	73.93	64.45	66.97	62.12
CYUT-CS-BC-03	67.86	68.12	70.74	70.16	71.33	64.98	65.63	64.35
bcNLP-CS-BC-01	67.04	69.65	76.32	65.98	90.52	57.75	80.20	45.13
bcNLP-CS-BC-02	66.89	69.91	76.89	65.71	92.65	56.88	83.33	43.18
MIG-CS-BC-01	65.71	65.81	67.56	69.33	65.88	63.87	62.11	65.74
CYUT-CS-BC-02	63.11	63.12	62.50	69.36	56.87	63.73	58.16	70.47
WHUTE-CS-BC-02	61.65	66.58	75.40	62.60	94.79	47.90	84.51	33.43
CYUT-CS-BC-01	61.17	61.59	57.14	71.94	47.39	65.20	55.86	78.27
*IASL-CS-BC-02	60.45	63.25	70.98	61.90	83.18	49.91	66.82	39.83
WHUTE-CS-BC-01	58.20	64.79	74.79	60.99	96.68	41.61	87.50	27.30
MIG-CS-BC-03	57.19	63.64	73.80	60.42	94.79	40.59	81.51	27.02
IMTKU-CS-BC-03	54.28	62.74	73.95	59.42	97.87	34.61	89.53	21.45
Yuntech-CS-BC-03	53.52	59.54	70.24	58.28	88.39	36.80	65.25	25.63
Yuntech-CS-BC-02	52.10	59.03	70.32	57.77	89.81	33.88	65.60	22.84
Yuntech-CS-BC-01	50.91	58.64	70.39	57.40	91.00	31.42	66.07	20.61
IMTKU-CS-BC-01	50.82	60.31	72.42	57.98	96.45	29.22	81.01	17.83
*IASL-CS-BC-01	50.60	54.03	63.63	55.58	74.41	37.57	50.00	30.08
WUST-CS-BC-02	50.14	58.77	70.89	57.31	92.89	29.39	69.07	18.66
WUST-CS-BC-01	50.14	58.77	70.89	57.31	92.89	29.39	69.07	18.66
*WUST-CS-BC-01	50.14	58.77	70.89	57.31	92.89	29.39	69.07	18.66
WUST-CS-BC-03	50.14	58.77	70.89	57.31	92.89	29.39	69.07	18.66
IMTKU-CS-BC-02	50.12	60.31	72.66	57.87	97.63	27.57	85.51	16.43
JUNLP-CS-BC-01	48.49	48.66	51.39	52.61	50.24	45.59	44.44	46.80

Table 10: Results on BC subtask (CS).

Team	MacroF1	Acc.	Y-F1	Y-Prec.	Y-Rec.	N-F1	N-Prec.	N-Rec.
IASL-CT-BC-02	67.14	67.76	71.66	68.64	74.95	62.63	66.48	59.20
MIG-CT-BC-02	67.07	67.54	70.99	69.03	73.07	63.14	65.51	60.95
MIG-CT-BC-03	66.99	67.54	71.23	68.74	73.90	62.76	65.85	59.95
IMTKU-CT-BC-01	65.99	66.29	69.16	68.80	69.52	62.83	63.22	62.44
WHUTE-CT-BC-01	65.55	66.17	70.20	67.37	73.28	60.89	64.44	57.71
MIG-CT-BC-01	65.42	65.61	67.94	68.88	67.01	62.91	61.93	63.93
IMTKU-CT-BC-03	63.82	64.25	67.76	66.47	69.10	59.87	61.36	58.46
Yuntech-CT-BC-03	62.31	62.54	65.26	65.82	64.72	59.36	58.78	59.95
Yuntech-CT-BC-02	62.02	62.54	66.46	64.75	68.27	57.58	59.57	55.72
Yuntech-CT-BC-01	61.64	62.43	67.13	64.02	70.56	56.16	60.06	52.74
KC99-CT-BC-01	57.67	59.48	66.42	60.45	73.70	48.93	57.58	42.54
CYUT-CT-BC-01	55.16	55.16	55.77	60.14	51.98	54.55	50.75	58.96
CYUT-CT-BC-02	52.64	53.35	58.44	56.67	60.33	46.83	48.79	45.02
IASL-CT-BC-01	51.77	55.85	65.79	56.84	78.08	37.76	52.91	29.35
CYUT-CT-BC-03	51.58	54.71	63.89	56.39	73.70	39.27	50.59	32.09
JUNLP-CT-BC-01	48.72	48.81	50.82	53.20	48.64	46.63	44.47	49.00
IMTKU-CT-BC-02	48.61	51.53	36.36	63.54	25.47	60.86	48.19	82.59
NTOUA-CT-BC-01	32.63	33.48	25.06	32.34	20.46	40.20	34.08	49.00
NTOUA-CT-BC-03	31.71	33.94	19.39	28.81	14.61	44.04	35.89	56.97
NTOUA-CT-BC-02	30.70	34.17	15.20	25.37	10.86	46.20	36.83	61.94

Table 11: Results on BC subtask (CT).

Team	MacroF1	Acc.	Y-F1	Y-Prec.	Y-Rec.	N-F1	N-Prec.	N-Rec.
*IASL-CAD-CS-ArtificialBC-02	59.53	63.53	72.24	66.02	79.77	46.82	57.14	39.66
bcNLP-CS-ArtificialBC-03	57.27	57.94	62.64	66.45	59.24	51.90	48.33	56.03
*IASL-CAD-CS-ArtificialBC-01	56.83	59.34	67.23	64.59	70.09	46.44	49.75	43.53
MIG-CS-ArtificialBC-02	53.29	53.93	58.75	62.88	55.13	47.83	44.16	52.16
MIG-CS-ArtificialBC-01	51.53	51.66	54.06	62.21	47.80	48.99	42.77	57.33
bcNLP-CS-ArtificialBC-01	51.08	53.75	62.52	60.38	64.81	39.64	42.03	37.50
WHUTE-CS-ArtificialBC-01	50.51	55.67	66.49	60.43	73.90	34.54	42.95	28.88
bcNLP-CS-ArtificialBC-02	50.31	53.58	63.06	59.89	66.57	37.56	41.24	34.48
Yuntech-CS-ArtificialBC-02	50.17	53.40	62.87	59.79	66.28	37.47	41.03	34.48
Yuntech-CS-ArtificialBC-01	50.03	54.28	64.59	59.90	70.09	35.47	41.38	31.03
IMTKU-CS-ArtificialBC-03	49.44	56.89	68.85	60.40	80.06	30.03	43.80	22.84
CYUT-CS-ArtificialBC-03	48.97	49.21	52.53	59.19	47.21	45.40	40.20	52.16
Yuntech-CS-ArtificialBC-03	48.57	52.53	62.84	58.82	67.45	34.30	39.01	30.60
IMTKU-CS-ArtificialBC-01	48.44	56.89	69.32	60.13	81.82	27.57	43.12	20.26
IMTKU-CS-ArtificialBC-02	47.65	57.07	69.85	60.00	83.58	25.45	42.86	18.10
WHUTE-CS-ArtificialBC-02	47.36	51.83	62.70	58.15	68.04	32.02	37.36	28.02
CYUT-CS-ArtificialBC-02	47.35	47.64	51.30	57.45	46.33	43.40	38.59	49.57
CYUT-CS-ArtificialBC-01	47.29	47.29	47.39	58.37	39.88	47.20	39.71	58.19
MIG-CS-ArtificialBC-03	46.42	52.88	65.03	58.24	73.61	27.81	36.62	22.41

Table 12: Results on ArtificialBC subtask (CS).

Team	MacroF1	Acc.	Y-F1	Y-Prec.	Y-Rec.	N-F1	N-Prec.	N-Rec.
IASL-CAD-CT-ArtificialBC-02	62.27	63.87	70.04	69.14	70.97	54.51	55.61	53.45
IASL-CAD-CT-ArtificialBC-01	55.44	57.77	65.62	63.64	67.74	45.25	47.62	43.10
Yuntech-CT-ArtificialBC-02	54.29	54.62	58.20	64.41	53.08	50.38	45.21	56.90
Yuntech-CT-ArtificialBC-01	53.68	54.28	58.93	63.30	55.13	48.43	44.57	53.02
MIG-CT-ArtificialBC-02	52.66	53.23	57.86	62.37	53.96	47.45	43.53	52.16
NTOUA-CT-ArtificialBC-01	52.57	53.05	57.37	62.41	53.08	47.77	43.46	53.02
Yuntech-CT-ArtificialBC-03	52.33	52.53	55.41	62.83	49.56	49.25	43.42	56.90
IMTKU-CT-ArtificialBC-01	52.21	52.53	56.13	62.37	51.03	48.29	43.20	54.74
MIG-CT-ArtificialBC-01	51.54	51.66	53.91	62.31	47.51	49.17	42.81	57.76
MIG-CT-ArtificialBC-03	51.51	52.71	59.13	60.87	57.48	43.89	42.23	45.69
IMTKU-CT-ArtificialBC-03	51.29	51.31	50.27	64.09	41.35	52.31	43.34	65.95
KC99-CT-ArtificialBC-01	51.12	51.66	56.24	60.96	52.20	46.00	41.99	50.86
NTOUA-CT-ArtificialBC-03	50.48	50.61	53.07	61.07	46.92	47.88	41.80	56.03
WHUTE-CT-ArtificialBC-01	50.19	51.13	57.06	59.81	54.55	43.32	40.84	46.12
CYUT-CT-ArtificialBC-03	49.50	49.74	52.94	59.78	47.51	46.07	40.73	53.02
NTOUA-CT-ArtificialBC-02	48.17	48.17	48.35	59.40	40.76	47.99	40.41	59.05
CYUT-CT-ArtificialBC-02	47.70	47.99	51.62	57.82	46.63	43.77	38.93	50.00
CYUT-CT-ArtificialBC-01	47.64	47.64	47.55	58.87	39.88	47.74	40.06	59.05
IMTKU-CT-ArtificialBC-02	36.50	43.63	15.22	72.50	8.50	57.78	41.46	95.26

Table 13: Results on ArtificialBC subtask (CT).

Team	MacroF1	Acc.	Y-F1	Y-Prec.	Y-Rec.	N-F1	N-Prec.	N-Rec.
bcNLP-CS-DevBC-03	81.26	83.29	87.43	85.38	89.58	75.09	78.85	71.68
WHUTE-CS-DevBC-01	75.83	80.22	86.13	78.99	94.70	65.52	84.53	53.50
IMTKU-CS-DevBC-01	73.93	79.85	86.36	77.00	98.30	61.50	93.57	45.80
Yuntech-CS-DevBC-01	73.51	78.13	84.58	77.96	92.42	62.45	78.72	51.75
Yuntech-CS-DevBC-03	71.91	76.78	83.61	77.12	91.29	60.21	75.66	50.00
IMTKU-CS-DevBC-03	71.67	77.40	84.41	76.38	94.32	58.93	81.48	46.15
IMTKU-CS-DevBC-02	71.36	78.62	85.78	75.43	99.43	56.93	97.46	40.21
bcNLP-CS-DevBC-02	70.00	76.78	84.26	75.19	95.83	55.74	84.40	41.61
WHUTE-CS-DevBC-02	69.79	76.17	83.67	75.30	94.13	55.91	79.87	43.01
MIG-CS-DevBC-02	69.63	72.85	79.52	77.86	81.25	59.74	62.36	57.34
Yuntech-CS-DevBC-02	69.57	74.69	82.06	75.97	89.20	57.08	70.62	47.90
bcNLP-CS-DevBC-01	69.50	76.04	83.63	75.11	94.32	55.38	80.13	42.31
CYUT-CS-DevBC-03	69.01	73.22	80.43	76.45	84.85	57.59	64.91	51.75
*IASL-CAD-CS-DevBC-01	68.85	72.48	79.49	76.95	82.20	58.21	62.40	54.55
MIG-CS-DevBC-01	68.57	71.25	77.76	78.05	77.46	59.38	58.97	59.79
*IASL-CAD-CS-DevBC-02	67.67	72.48	80.14	75.33	85.61	55.20	64.49	48.25
MIG-CS-DevBC-03	66.96	75.68	83.93	73.44	97.92	50.00	90.00	34.62
CYUT-CS-DevBC-01	63.39	63.88	67.62	80.79	58.14	59.17	49.08	74.48
CYUT-CS-DevBC-02	62.67	64.62	71.20	75.42	67.42	54.14	49.71	59.44

Table 14: Results on DevBC subtask (CS).

Team	MacroF1	Acc.	Y-F1	Y-Prec.	Y-Rec.	N-F1	N-Prec.	N-Rec.
Yuntech-CT-DevBC-01	74.46	75.25	78.94	73.24	85.61	69.97	78.72	62.98
MIG-CT-DevBC-01	72.60	72.90	75.48	74.06	76.96	69.71	71.40	68.10
IMTKU-CT-DevBC-01	72.54	72.90	75.68	73.68	77.79	69.40	71.86	67.11
MIG-CT-DevBC-02	72.23	72.75	76.03	72.65	79.75	68.42	72.90	64.46
Yuntech-CT-DevBC-03	71.94	72.29	75.10	73.21	77.09	68.77	71.08	66.61
Yuntech-CT-DevBC-02	70.90	71.61	75.44	71.02	80.45	66.37	72.55	61.16
MIG-CT-DevBC-03	69.73	71.54	77.13	68.32	88.55	62.32	79.13	51.40
WHUTE-CT-DevBC-01	69.15	69.95	74.10	69.52	79.33	64.20	70.63	58.84
IMTKU-CT-DevBC-03	68.56	68.81	71.39	70.99	71.79	65.72	66.16	65.29
IASL-CAD-CT-DevBC-02	68.38	69.04	72.93	69.31	76.96	63.84	68.63	59.67
CYUT-CT-DevBC-03	68.06	68.89	73.19	68.67	78.35	62.94	69.25	57.69
IASL-CAD-CT-DevBC-01	62.84	64.57	70.86	63.93	79.47	54.83	65.89	46.94
KC99-CT-DevBC-01	61.70	62.76	68.05	63.59	73.18	55.35	61.37	50.41
CYUT-CT-DevBC-02	61.12	61.32	63.89	64.66	63.13	58.35	57.56	59.17
CYUT-CT-DevBC-01	57.86	57.99	55.56	65.10	48.46	60.16	53.17	69.26
IMTKU-CT-DevBC-02	55.64	57.76	45.93	75.00	33.10	65.34	52.34	86.94
NTOUA-CT-DevBC-03	35.02	36.11	26.61	35.25	21.37	43.43	36.53	53.55
NTOUA-CT-DevBC-01	34.64	35.12	29.00	35.64	24.44	40.28	34.82	47.77
NTOUA-CT-DevBC-02	32.52	34.97	19.64	29.75	14.66	45.39	36.88	59.01

Table 15: Results on DevBC subtask (CT).

Team	MacroF1	Acc.	B-F1	B-Prec.	B-Rec.	F-F1	F-Prec.	F-Rec.	C-F1	C-Prec.	C-Rec.	I-F1	I-Prec.	I-Rec.
SKL-JA-MC-01	59.96	69.53	67.18	72.13	62.86	76.47	76.85	76.10	21.15	25.58	18.03	75.06	70.54	80.19
SKL-JA-MC-02	58.25	68.61	69.29	77.19	62.86	74.94	73.36	76.59	13.59	16.67	11.48	75.17	71.49	79.25
SKL-JA-MC-03	55.45	68.07	63.24	65.15	61.43	73.85	69.70	78.54	9.20	15.38	6.56	75.51	73.33	77.83
WSD-JA-MC-03	54.39	69.53	68.29	59.57	80.00	75.29	72.73	78.05	0.00	0.00	0.00	73.99	70.51	77.83
WSD-JA-MC-02	54.18	68.98	68.83	63.10	75.71	73.95	70.67	77.56	0.00	0.00	0.00	73.94	70.04	78.30
WSD-JA-MC-01	53.98	68.80	68.29	59.57	80.00	74.70	72.48	77.07	0.00	0.00	0.00	72.93	69.36	76.89
FLL-JA-MC-01	53.67	64.96	69.01	68.06	70.00	70.35	60.56	83.90	8.82	42.86	4.92	66.50	71.35	62.26
JAIST-JA-MC-01	52.60	66.97	66.67	64.86	68.57	74.55	69.79	80.00	0.00	0.00	0.00	69.20	65.68	73.11
BnO-JA-MC-02	52.44	57.66	58.94	44.53	87.14	62.18	73.03	54.15	20.38	16.67	26.23	68.27	78.53	60.38
JAIST-JA-MC-02	52.27	65.33	63.51	60.26	67.14	71.84	59.68	90.24	5.97	33.33	3.28	67.76	80.52	58.49
BnO-JA-MC-03	52.11	65.69	59.74	54.76	65.71	72.93	67.36	79.51	5.80	25.00	3.28	69.95	69.63	70.28
BnO-JA-MC-01	52.03	66.42	65.82	59.09	74.29	72.23	67.23	78.05	0.00	0.00	0.00	70.05	68.47	71.70
JAIST-JA-MC-03	51.48	65.33	67.92	60.67	77.14	71.13	58.49	90.73	0.00	0.00	0.00	66.86	83.69	55.66
*FLL-JA-MC-04	51.27	64.23	64.94	59.52	71.43	68.84	65.78	72.20	3.17	50.00	1.64	68.15	64.56	72.17
KYOTO-JA-MC-02	50.12	64.78	59.55	49.07	75.71	69.90	69.57	70.24	0.00	0.00	0.00	71.01	67.81	74.53
*FLL-JA-MC-02	35.12	44.71	25.93	36.84	20.00	49.58	43.82	57.07	16.00	42.86	9.84	48.98	47.16	50.94
THK-JA-MC-01	30.98	49.09	21.95	75.00	12.86	60.75	47.77	83.41	28.57	52.17	19.67	43.63	54.61	36.32
EHIME-JA-MC-03	25.89	40.33	4.76	14.29	2.86	49.46	39.26	66.83	8.82	42.86	4.92	40.51	44.38	37.26
EHIME-JA-MC-01	24.47	28.10	14.58	11.48	20.00	37.97	42.01	34.63	12.36	9.40	18.03	32.95	41.43	27.36
*FLL-JA-MC-03	22.47	34.49	11.01	15.38	8.57	40.64	34.59	49.27	0.00	0.00	0.00	38.23	37.79	38.68
EHIME-JA-MC-02	21.99	36.31	2.41	7.69	1.43	43.98	35.78	57.07	3.03	20.00	1.64	38.55	39.41	37.74
JUNLP-JA-MC-01	21.42	22.63	16.83	12.88	24.29	24.42	30.22	20.49	17.17	12.41	27.87	27.27	34.29	22.64
KYOTO-JA-MC-01	17.04	40.33	0.00	0.00	0.00	8.29	75.00	4.39	3.23	100.00	1.64	56.64	39.59	99.53
Baseline-JA-MC-01	26.61	45.44	0.00	0.00	0.00	56.18	43.01	80.98	5.41	15.38	3.28	44.88	54.36	38.21

Table 16: Results on MC subtask (JA).

Team	MacroF1	Acc.	B-F1	B-Prec.	B-Rec.	F-F1	F-Prec.	F-Rec.	C-F1	C-Prec.	C-Rec.	I-F1	I-Prec.	I-Rec.
bcNLP-CS-MC-03	56.82	61.08	66.67	77.27	58.62	67.30	53.91	89.53	38.41	64.44	27.36	54.89	69.28	45.45
*IASL-CS-MC-02	50.94	53.91	55.30	61.34	50.34	64.44	57.51	73.29	38.42	40.21	36.79	45.59	50.00	41.90
WHUTE-CS-MC-01	46.79	54.80	61.54	62.41	60.69	64.36	49.90	90.61	18.71	39.39	12.26	42.58	73.08	30.04
WHUTE-CS-MC-02	46.53	56.59	62.25	59.87	64.83	65.09	51.24	89.17	8.26	33.33	4.72	50.53	75.59	37.94
bcNLP-CS-MC-02	44.88	57.62	59.68	71.84	51.03	67.86	52.58	95.67	0.00	0.00	0.00	51.99	63.79	43.87
MIG-CS-MC-02	44.74	51.60	58.50	49.07	72.41	52.84	57.69	48.74	11.35	22.86	7.55	56.26	52.01	61.26
CYUT-CS-MC-02	42.52	48.78	53.64	51.59	55.86	56.13	48.80	66.06	12.42	18.18	9.43	47.87	55.15	42.29
MIG-CS-MC-01	41.82	49.17	56.44	50.83	63.45	50.70	57.27	45.49	5.48	10.00	3.77	54.64	47.65	64.03
Yuntech-CS-MC-02	40.91	51.22	55.02	51.83	58.62	64.81	49.90	92.42	13.43	32.14	8.49	30.40	65.79	19.76
Yuntech-CS-MC-03	40.89	51.22	53.95	51.57	56.55	65.15	50.20	92.78	13.43	32.14	8.49	31.04	63.41	20.55
WUST-CS-MC-02	40.87	52.37	59.74	56.44	63.45	62.75	47.99	90.61	3.57	33.33	1.89	37.43	71.91	25.30
WUST-CS-MC-03	40.87	52.37	59.74	56.44	63.45	62.75	47.99	90.61	3.57	33.33	1.89	37.43	71.91	25.30
*WUST-CS-MC-01	40.87	52.37	59.74	56.44	63.45	62.75	47.99	90.61	3.57	33.33	1.89	37.43	71.91	25.30
CYUT-CS-MC-01	40.37	47.63	60.34	59.33	61.38	56.72	44.94	76.90	12.31	33.33	7.55	32.12	46.62	24.51
WUST-CS-MC-01	40.33	51.73	59.31	54.65	64.83	62.20	47.86	88.81	3.57	33.33	1.89	36.26	69.66	24.51
Yuntech-CS-MC-01	40.33	50.70	53.42	50.62	56.55	64.56	49.61	92.42	13.64	34.62	8.49	29.70	63.64	19.37
CYUT-CS-MC-03	40.10	51.09	48.73	51.54	46.21	57.07	44.31	80.14	0.00	0.00	0.00	54.59	73.33	43.48
bcNLP-CS-MC-01	39.95	53.91	43.43	81.13	29.66	64.70	49.07	94.95	0.00	0.00	0.00	51.69	59.90	45.45
*IASL-CS-MC-01	34.95	41.74	37.29	48.35	30.34	59.39	47.05	80.51	25.24	26.00	24.53	17.89	28.45	13.04
MIG-CS-MC-03	34.42	43.15	53.90	39.80	83.45	58.58	51.96	67.15	12.94	13.68	12.26	12.27	70.83	6.72
IMTKU-CS-MC-03	27.26	40.20	9.81	10.83	8.97	67.10	52.67	92.42	32.14	25.86	42.45	0.00	0.00	0.00
JUNLP-CS-MC-01	24.38	24.71	22.42	19.59	26.21	27.00	32.49	23.10	22.02	16.29	33.96	26.07	32.54	21.74
IMTKU-CS-MC-01	23.89	37.64	5.85	10.00	4.14	63.20	48.73	89.89	22.82	17.71	32.08	3.69	27.78	1.98
IMTKU-CS-MC-02	19.67	36.11	7.73	12.90	5.52	57.78	41.73	93.86	8.74	10.39	7.55	4.41	31.58	2.37

Table 17: Results on MC subtask (CS).

Team	MacroF1	Acc.	B-F1	B-Prec.	B-Rec.	F-F1	F-Prec.	F-Rec.	C-F1	C-Prec.	C-Rec.	I-F1	I-Prec.	I-Rec.
IASL-CT-MC-02	46.32	51.99	52.35	53.06	51.66	64.63	53.99	80.49	29.90	36.25	25.44	38.41	52.73	30.21
WHUTE-CT-MC-01	45.50	55.16	58.86	56.36	61.59	67.06	54.36	87.50	12.08	25.71	7.89	43.99	63.40	33.68
MIG-CT-MC-02	45.15	51.53	57.68	48.64	70.86	54.49	58.60	50.91	14.19	26.83	9.65	54.25	50.45	58.68
NTOUA-CT-MC-03	44.80	55.73	61.10	50.43	77.48	64.21	55.00	77.13	1.50	5.26	0.88	52.40	70.59	41.67
NTOUA-CT-MC-01	44.63	56.64	62.07	54.82	71.52	65.79	54.01	84.15	0.00	0.00	0.00	50.66	69.28	39.93
MIG-CT-MC-03	44.21	48.92	51.07	39.93	70.86	52.52	61.13	46.04	19.19	22.62	16.67	54.04	54.61	53.47
KC99-CT-MC-01	43.75	50.17	45.48	42.94	48.34	63.61	57.00	71.95	16.67	15.87	17.54	49.24	66.08	39.24
MIG-CT-MC-01	42.16	48.92	53.49	47.67	60.93	51.91	57.14	47.56	10.06	17.78	7.02	53.19	47.30	60.76
Yuntech-CT-MC-03	40.14	51.99	52.70	53.79	51.66	65.27	51.03	90.55	4.38	13.04	2.63	38.19	61.07	27.78
Yuntech-CT-MC-01	39.76	51.76	48.92	53.54	45.03	65.80	51.19	92.07	5.63	14.29	3.51	38.68	60.29	28.47
Yuntech-CT-MC-02	38.77	51.31	46.89	52.46	42.38	65.59	50.75	92.68	4.32	12.00	2.63	38.30	60.00	28.13
IMTKU-CT-MC-01	35.76	50.85	60.56	52.15	72.19	63.47	57.57	70.73	13.41	22.00	9.65	41.38	54.55	33.33
NTOUA-CT-MC-02	33.49	49.94	1.29	25.00	0.66	62.50	50.96	80.79	13.98	18.06	11.40	56.20	56.49	55.90
IASL-CT-MC-01	33.04	42.22	32.20	44.71	25.17	60.71	47.58	83.84	21.72	22.43	21.05	17.54	31.53	12.15
MCUIM-CT-MC-01	32.51	46.42	59.21	58.82	59.60	70.07	61.33	81.71	25.00	20.69	31.58	8.29	20.27	5.21
IMTKU-CT-MC-03	32.36	50.06	52.83	44.55	64.90	65.58	59.02	73.78	1.65	14.29	0.88	41.75	52.36	34.72
CYUT-CT-MC-02	26.26	30.76	21.39	18.97	24.50	43.68	38.43	50.61	17.00	15.79	18.42	22.98	38.84	16.32
CYUT-CT-MC-01	25.60	31.90	22.58	19.00	27.81	45.85	38.62	56.40	12.44	15.19	10.53	21.54	41.18	14.58
JUNLP-CT-MC-01	24.21	25.31	21.22	17.70	26.49	32.28	41.23	26.52	16.72	12.67	24.56	26.61	30.49	23.61
CYUT-CT-MC-03	23.51	32.24	27.22	19.76	43.71	45.62	39.51	53.96	0.00	0.00	0.00	21.19	41.41	14.24
IMTKU-CT-MC-02	19.37	36.55	0.00	0.00	0.00	30.63	64.08	20.12	15.95	26.53	11.40	50.26	35.79	84.38

Table 18: Results on MC subtask (CT).

Team	MacroF1	Acc.	B-F1	B-P.	B-R.	F-F1	F-P.	F-R.	C-F1	C-P.	C-R.	I-F1	I-P.	I-R.
bcNLP-CS-ArtificialMC-03	40.23	42.23	41.69	52.89	34.41	52.17	43.22	65.81	30.97	60.00	20.87	36.11	29.55	46.43
IASL-CAD-CS-ArtificialMC-02	39.23	40.49	43.56	50.71	38.17	47.93	41.83	56.13	32.61	43.48	26.09	32.84	28.21	39.29
WHUTE-CS-ArtificialMC-01	36.89	40.31	41.46	43.27	39.78	53.26	44.74	65.81	22.08	43.59	14.78	30.77	28.15	33.93
IASL-CAD-CS-ArtificialMC-01	33.41	36.30	28.47	42.11	21.51	50.82	39.78	70.32	25.67	33.33	20.87	28.69	26.52	31.25
Yuntech-CS-ArtificialMC-01	33.28	38.74	40.77	41.81	39.78	53.24	42.37	71.61	12.90	25.00	8.70	26.21	28.72	24.11
Yuntech-CS-ArtificialMC-03	31.70	37.17	37.08	38.82	35.48	53.49	42.69	71.61	10.53	21.62	6.96	25.69	26.42	25.00
Yuntech-CS-ArtificialMC-02	31.27	37.17	37.91	38.76	37.10	53.37	42.53	71.61	9.59	22.58	6.09	24.19	25.24	23.21
IMTKU-CS-ArtificialMC-02	30.05	36.65	27.30	37.38	21.51	54.55	43.35	73.55	35.06	27.98	46.96	3.28	20.00	1.79
IMTKU-CS-ArtificialMC-01	29.97	36.65	25.87	37.00	19.89	54.63	43.23	74.19	36.13	28.72	48.70	3.23	16.67	1.79
CYUT-CS-ArtificialMC-02	29.94	31.94	37.65	44.20	32.80	40.00	34.42	47.74	11.21	12.12	10.43	30.90	29.75	32.14
WHUTE-CS-ArtificialMC-02	29.79	35.43	39.78	40.33	39.25	48.26	41.28	58.06	3.08	13.33	1.74	28.04	23.90	33.93
IMTKU-CS-ArtificialMC-03	29.34	35.43	30.77	34.21	27.96	50.38	41.32	64.52	34.48	28.57	43.48	1.72	25.00	0.89
CYUT-CS-ArtificialMC-01	29.15	32.64	31.52	36.11	27.96	42.64	32.33	62.58	20.00	35.56	13.91	22.45	26.19	19.64
bcNLP-CS-ArtificialMC-02	28.95	35.60	27.14	40.43	20.43	50.00	39.41	68.39	1.69	33.33	0.87	36.99	28.50	52.68
CYUT-CS-ArtificialMC-03	28.17	33.86	33.52	34.88	32.26	41.49	32.48	57.42	0.00	0.00	0.00	37.66	35.43	40.18
MIG-CS-ArtificialMC-02	27.99	30.72	35.08	34.18	36.02	33.33	41.75	27.74	8.54	14.29	6.09	35.01	26.22	52.68
MIG-CS-ArtificialMC-01	27.43	30.37	33.89	35.06	32.80	30.45	42.05	23.87	8.59	14.58	6.09	36.80	26.24	61.61
bcNLP-CS-ArtificialMC-01	25.97	33.68	18.93	40.35	12.37	48.14	36.42	70.97	0.00	0.00	0.00	36.81	28.04	53.57
MIG-CS-ArtificialMC-03	25.87	32.11	41.93	34.36	53.76	46.10	48.57	43.87	10.40	9.63	11.30	5.04	42.86	2.68

Table 19: Results on ArtificialMC subtask (CS).

Team	MacroF1	Acc.	B-F1	B-P.	B-R.	F-F1	F-P.	F-R.	C-F1	C-P.	C-R.	I-F1	I-P.	I-R.
IASL-CAD-CT-ArtificialMC-02	40.37	41.36	43.90	50.70	38.71	49.04	42.79	57.42	36.96	49.28	29.57	31.58	27.27	37.50
WHUTE-CT-ArtificialMC-01	37.81	39.79	42.31	43.26	41.40	50.43	45.79	56.13	27.85	51.16	19.13	30.66	25.93	37.50
CYUT-CT-ArtificialMC-02	36.89	37.00	37.04	43.48	32.26	39.80	32.14	52.26	40.68	58.06	31.30	30.04	28.93	31.25
KC99-CT-ArtificialMC-01	35.65	36.13	39.78	40.33	39.25	43.61	52.25	37.42	25.75	25.42	26.09	33.45	28.22	41.07
IMTKU-CT-ArtificialMC-03	34.48	35.60	41.23	42.77	39.78	47.18	51.94	43.23	13.53	50.00	7.83	35.99	29.38	46.43
Yuntech-CT-ArtificialMC-03	34.20	39.44	40.44	41.11	39.78	55.50	46.70	68.39	9.93	26.92	6.09	30.95	27.86	34.82
IMTKU-CT-ArtificialMC-01	33.67	34.73	40.11	40.44	39.78	45.90	46.67	45.16	16.22	36.36	10.43	32.45	28.10	38.39
IASL-CAD-CT-ArtificialMC-01	33.55	36.30	28.78	43.48	21.51	50.94	40.15	69.68	26.60	34.25	21.74	27.89	25.18	31.25
Yuntech-CT-ArtificialMC-01	32.25	36.82	34.96	37.42	32.80	54.08	44.73	68.39	12.08	26.47	7.83	27.89	25.18	31.25
Yuntech-CT-ArtificialMC-02	32.13	37.00	34.03	38.26	30.65	53.56	43.25	70.32	11.11	27.59	6.96	29.80	26.57	33.93
NTOUA-CT-ArtificialMC-03	31.75	35.60	41.90	37.61	47.31	46.32	53.85	40.65	6.67	14.29	4.35	32.11	25.67	42.86
NTOUA-CT-ArtificialMC-01	31.00	36.30	39.41	36.36	43.01	47.87	48.67	47.10	1.64	14.29	0.87	35.06	27.55	48.21
CYUT-CT-ArtificialMC-01	28.59	32.11	33.03	37.41	29.57	41.33	31.53	60.00	18.87	34.09	13.04	21.11	24.14	18.75
CYUT-CT-ArtificialMC-03	28.29	34.03	33.98	35.26	32.80	41.69	32.72	57.42	0.00	0.00	0.00	37.50	35.16	40.18
MCUIM-CT-ArtificialMC-01	28.16	30.72	42.20	45.63	39.25	51.55	49.70	53.55	11.34	10.61	12.17	7.55	12.77	5.36
MIG-CT-ArtificialMC-01	27.72	30.37	33.52	34.88	32.26	29.63	40.91	23.23	10.91	18.00	7.83	36.80	26.24	61.61
MIG-CT-ArtificialMC-02	27.09	30.02	34.39	33.85	34.95	33.33	41.75	27.74	6.13	10.42	4.35	34.50	25.65	52.68
MIG-CT-ArtificialMC-03	25.75	29.67	38.27	33.20	45.16	23.08	45.28	15.48	4.37	5.88	3.48	37.30	29.15	51.79
NTOUA-CT-ArtificialMC-02	22.96	27.23	1.06	50.00	0.54	38.12	32.02	47.10	19.25	25.00	15.65	33.42	23.62	57.14
IMTKU-CT-ArtificialMC-02	20.75	23.73	7.77	40.00	4.30	20.65	65.52	12.26	17.14	48.00	10.43	37.45	23.89	86.61

Table 20: Results on ArtificialMC subtask (CT).

Team	MacroF1	Acc.	B-F1	B-P.	B-R.	F-F1	F-P.	F-R.	C-F1	C-P.	C-R.	I-F1	I-P.	I-R.
bcNLP-CS-DevMC-03	77.60	81.08	82.10	80.61	83.65	87.44	81.50	94.31	69.75	90.22	56.85	71.11	73.85	68.57
WHUTE-CS-DevMC-01	66.61	73.46	72.05	66.49	78.62	85.11	79.07	92.14	44.74	62.20	34.93	64.57	71.93	58.57
Yuntech-CS-DevMC-01	61.28	68.80	63.03	60.82	65.41	81.15	72.49	92.14	46.55	62.79	36.99	54.39	70.45	44.29
Yuntech-CS-DevMC-03	59.05	66.83	59.08	57.83	60.38	79.72	70.89	91.06	45.22	61.90	35.62	52.17	66.67	42.86
Yuntech-CS-DevMC-02	58.65	66.83	60.53	57.30	64.15	79.81	71.04	91.06	42.73	63.51	32.19	51.53	66.29	42.14
WHUTE-CS-DevMC-02	57.59	69.78	68.47	56.28	87.42	84.65	80.15	89.70	13.02	47.83	7.53	64.21	66.41	62.14
IASL-CAD-CS-DevMC-02	55.84	61.67	63.23	64.90	61.64	78.02	79.11	76.96	43.92	51.38	38.36	38.21	32.82	45.71
CYUT-CS-DevMC-01	54.06	61.92	62.60	55.94	71.07	73.68	67.19	81.57	39.44	62.69	28.77	40.51	49.48	34.29
bcNLP-CS-DevMC-02	53.18	67.44	60.52	55.85	66.04	81.58	71.26	95.39	3.97	60.00	2.05	66.67	70.08	63.57
bcNLP-CS-DevMC-01	49.16	64.50	53.29	55.86	50.94	78.79	66.98	95.66	0.00	0.00	0.00	64.54	64.08	65.00
CYUT-CS-DevMC-02	47.13	53.32	61.95	58.33	66.04	64.94	62.13	68.02	17.10	18.70	15.75	44.53	51.40	39.29
IASL-CAD-CS-DevMC-01	46.69	56.63	44.19	54.63	37.11	76.19	69.33	84.55	40.44	43.65	37.67	25.93	26.92	25.00
MIG-CS-DevMC-01	46.61	53.93	53.21	42.75	70.44	68.78	82.82	58.81	13.13	25.00	8.90	51.32	40.76	69.29
MIG-CS-DevMC-02	46.28	53.81	52.49	40.99	72.96	69.07	82.09	59.62	12.00	22.22	8.22	51.58	43.06	64.29
CYUT-CS-DevMC-03	45.95	58.11	56.10	42.53	82.39	70.32	66.35	74.80	0.00	0.00	0.00	57.39	73.33	47.14
MIG-CS-DevMC-03	39.49	51.60	52.76	36.82	93.08	70.94	78.81	64.50	17.28	21.65	14.38	16.99	100.00	9.29
IMTKU-CS-DevMC-01	30.86	50.74	1.65	2.41	1.26	85.68	78.68	94.04	32.17	26.43	41.10	3.94	6.35	2.86
IMTKU-CS-DevMC-02	30.57	50.61	1.65	2.41	1.26	86.03	79.09	94.31	31.64	25.99	40.41	2.94	4.69	2.14
IMTKU-CS-DevMC-03	30.09	47.91	4.91	5.56	4.40	81.36	76.00	87.53	34.09	29.13	41.10	0.00	0.00	0.00

Table 21: Results on DevMC subtask (CS).

Team	MacroF1	Acc.	B-F1	B-P.	B-R.	F-F1	F-P.	F-R.	C-F1	C-P.	C-R.	I-F1	I-P.	I-R.
WHUTE-CT-DevMC-01	64.76	70.02	66.07	61.79	70.99	82.96	77.30	89.52	48.16	64.05	38.58	61.85	64.98	59.00
Yuntech-CT-DevMC-01	61.30	66.99	63.10	61.07	65.27	78.87	70.26	89.89	44.33	64.18	33.86	58.90	65.88	53.26
Yuntech-CT-DevMC-03	58.45	64.88	61.59	57.43	66.41	77.94	70.15	87.68	38.59	62.28	27.95	55.67	60.27	51.72
Yuntech-CT-DevMC-02	57.46	63.82	58.47	57.09	59.92	76.74	67.89	88.24	39.05	59.20	29.13	55.58	61.68	50.57
CYUT-CT-DevMC-02	55.23	58.82	55.60	54.38	56.87	68.08	63.25	73.71	44.86	55.17	37.80	52.40	54.81	50.19
IASL-CAD-CT-DevMC-02	53.10	57.68	55.25	56.35	54.20	74.52	74.05	75.00	43.08	50.80	37.40	39.53	35.35	44.83
CYUT-CT-DevMC-01	53.00	58.59	56.23	52.67	60.31	69.09	60.50	80.51	38.36	63.06	27.56	48.32	58.06	41.38
KC99-CT-DevMC-01	52.18	56.09	52.10	45.22	61.45	72.92	78.85	67.83	33.79	33.33	34.25	49.90	52.54	47.51
NTOUA-CT-DevMC-01	49.60	59.20	55.89	44.47	75.19	76.35	75.00	77.76	11.33	36.96	6.69	54.82	54.10	55.56
NTOUA-CT-DevMC-03	48.93	57.23	55.54	43.28	77.48	74.03	78.28	70.22	9.70	21.05	6.30	56.47	53.82	59.39
MIG-CT-DevMC-03	48.59	52.31	53.22	38.41	86.64	63.69	84.50	51.10	24.06	33.10	18.90	53.38	53.91	52.87
MIG-CT-DevMC-02	47.49	52.76	53.93	44.07	69.47	64.88	77.69	55.70	19.13	36.26	12.99	52.03	41.92	68.58
MIG-CT-DevMC-01	46.51	52.23	52.98	44.95	64.50	64.31	77.12	55.15	16.31	35.06	10.63	52.43	40.50	74.33
IMTKU-CT-DevMC-01	46.45	58.67	64.86	53.47	82.44	67.97	72.75	63.79	44.67	62.86	34.65	54.75	64.58	47.51
IASL-CAD-CT-DevMC-01	45.81	54.13	38.80	49.12	32.06	74.74	66.96	84.56	40.96	41.80	40.16	28.75	31.51	26.44
CYUT-CT-DevMC-03	40.63	51.25	51.49	41.04	69.08	62.41	55.15	71.88	0.00	0.00	0.00	48.61	61.40	40.23
NTOUA-CT-DevMC-02	38.31	50.57	5.05	46.67	2.67	65.92	58.46	75.55	26.60	33.53	22.05	55.67	44.50	74.33
IMTKU-CT-DevMC-03	37.30	49.36	51.04	43.84	61.07	65.51	69.04	62.32	27.43	50.00	18.90	42.51	45.06	40.23
MCUIM-CT-DevMC-01	32.46	51.85	67.13	61.39	74.05	84.89	83.60	86.21	5.13	6.29	4.33	5.16	6.67	4.21
IMTKU-CT-DevMC-02	24.53	32.40	3.93	13.95	2.29	50.12	76.32	37.32	23.41	33.09	18.11	45.17	34.26	66.28

Table 22: Results on DevMC subtask (CT).

Team	MacroF1	Acc.	Correct Answer Ratio	Y-F1	Y-Prec.	Y-Rec.	N-F1	N-Prec.	N-Rec.
BnO-JA-ExamBC-02	67.15	70.31	55.56	56.96	64.71	50.87	77.34	72.76	82.55
BnO-JA-ExamBC-03	66.97	68.75	57.41	59.30	59.65	58.96	74.64	74.37	74.91
KDR-JA-ExamBC-02	66.90	68.75	51.85	59.06	59.76	58.38	74.73	74.19	75.27
BnO-JA-ExamBC-01	66.86	69.87	57.41	56.87	63.57	51.45	76.84	72.73	81.45
KDR-JA-ExamBC-03	66.64	68.30	47.22	59.20	58.86	59.54	74.09	74.36	73.82
WSD-JA-ExamBC-01	64.90	67.86	52.78	54.72	60.00	50.29	75.09	71.62	78.91
IBM-JA-ExamBC-03	64.18	64.51	45.37	60.74	53.02	71.10	67.62	76.85	60.36
WSD-JA-ExamBC-03	64.71	67.63	52.78	54.55	59.59	50.29	74.87	71.52	78.55
SKL-JA-ExamBC-02	64.04	65.63	49.07	56.50	55.25	57.80	71.59	72.66	70.55
WSD-JA-ExamBC-02	63.96	67.63	51.85	52.46	60.61	46.24	75.47	70.57	81.09
KDR-JA-ExamBC-01	63.31	64.51	49.07	56.68	53.61	60.12	69.94	72.83	67.27
SKL-JA-ExamBC-01	61.65	67.63	29.63	46.49	64.29	36.42	76.80	68.57	87.27
SKL-JA-ExamBC-03	60.47	63.17	42.59	50.15	52.53	47.98	70.80	68.97	72.73
KitAi-JA-ExamBC-01	59.84	63.17	36.11	48.28	52.74	44.51	71.40	68.21	74.91
IBM-JA-ExamBC-02	59.33	61.83	46.29	49.26	50.61	47.98	69.41	68.31	70.55
KitAi-JA-ExamBC-03	59.05	61.38	45.37	49.27	50.00	48.55	68.83	68.21	69.45
JAIST-JA-ExamBC-02	59.04	63.39	41.67	45.70	53.49	39.88	72.39	67.40	78.18
JAIST-JA-ExamBC-03	58.65	64.96	42.59	42.49	58.00	33.53	74.80	66.95	84.73
IBM-JA-ExamBC-01	58.53	61.61	44.44	47.24	50.33	44.51	69.82	67.46	72.36
JAIST-JA-ExamBC-01	57.55	63.17	40.74	42.11	53.57	34.68	73.00	66.37	81.09
KitAi-JA-ExamBC-02	57.16	58.71	39.81	49.04	46.84	51.45	65.29	67.44	63.27
KYOTO-JA-ExamBC-02	56.82	62.05	43.52	41.78	51.26	35.26	71.85	65.96	78.91
NTTD-JA-ExamBC-02	55.57	55.58	34.26	54.88	45.15	69.94	56.26	71.11	46.55
NTTD-JA-ExamBC-03	53.12	54.02	34.26	46.63	42.25	52.02	59.61	64.68	55.27
NTTD-JA-ExamBC-01	52.02	58.93	31.48	33.81	44.76	27.17	70.23	63.27	78.91
JUNLP-JA-ExamBC-01	50.46	50.89	30.56	45.81	39.91	53.76	55.10	62.79	49.09
*TKDDI-JA-ExamBC-03	49.08	62.50	28.70	22.94	55.56	14.45	75.22	63.28	92.73
TKDDI-JA-ExamBC-01	48.62	62.28	26.85	22.12	54.55	13.87	75.11	63.12	92.73
TKDDI-JA-ExamBC-02	48.62	62.28	26.85	22.12	54.55	13.87	75.11	63.12	92.73
THK-JA-ExamBC-01	43.77	62.28	26.85	11.52	61.11	6.36	76.03	62.33	97.45
KYOTO-JA-ExamBC-03	38.57	61.38	21.30	1.14	50.00	0.58	76.01	61.43	99.64
KYOTO-JA-ExamBC-01	37.86	60.94	20.37	0.00	0.00	0.00	75.73	61.21	99.27
Baseline-JA-ExamBC-01	54.77	56.47	32.41	45.98	44.15	47.98	63.55	65.38	61.82

Table 23: Results on Exam BC subtask (JA).

Team	MacroF1	Acc.	Correct Answer Ratio	Y-F1	Y-Prec.	Y-Rec.	N-F1	N-Prec.	N-Rec.
*KDR-JA-ExamSearch-02	58.12	64.51	32.41	41.76	57.00	32.95	74.48	66.67	84.36
*KDR-JA-ExamSearch-01	57.59	63.84	33.33	41.30	55.34	32.95	73.87	66.38	83.27
*KDR-JA-ExamSearch-03	57.39	63.17	34.26	41.70	53.64	34.10	73.08	66.27	81.45
NTTD-JA-ExamSearch-01	55.02	58.04	25.93	43.37	45.28	41.62	66.67	65.05	68.36
*BnO-JA-ExamSearch-02	54.77	56.47	31.48	45.98	44.15	47.98	63.55	65.38	61.82
*BnO-JA-ExamSearch-01	52.45	54.91	26.85	41.62	41.62	41.62	63.27	63.27	63.27
*BnO-JA-ExamSearch-03	51.78	51.79	31.48	51.57	42.12	66.47	52.00	66.86	42.55
NTTD-JA-ExamSearch-02	49.15	49.33	25.93	52.21	41.06	71.68	46.08	66.44	35.27
KYOTO-JA-ExamSearch-01	46.57	62.95	28.70	17.00	62.96	9.83	76.15	62.95	96.36
KYOTO-JA-ExamSearch-02	45.41	62.28	26.85	15.08	57.69	8.67	75.75	62.56	96.00

Table 24: Results on Exam Search subtask (JA).

the same as the ExamBC subtask, the highest value of the correct answer ratio was lower than 40% where as the highest value of the correct answer ratio on ExamBC subtask was approx. 57. This difference suggests that automatically retrieving t_1 from a set of documents is essentially difficult task.

5.5 Results on UnitTest

Table 25 shows the results on UnitTest. Since each sentence pair in the UnitTest dataset includes only one linguistic phenomena, recognizing their semantic relations are easier than the other dataset. Actually, the top system achieved approximately 90% in accuracy. In addition, most of the examples are "Y", it is important to detect "N" in the dataset. The top performance of detecting "N" is achieved by the FLL team (60.71 in F1).

In order to explore the difficulties of handling the linguistic phenomena listed in Table 5, we show the performances of participant's systems for each category of linguistic phenomena in Table 26. Due to the space constraint, we show the performances regarding "N" label. The category into which most of the "N" examples categorized is "disagree:phrase", and many systems could correctly recognize these examples as "N". On the other hand, the examples categorized into "synonymy:phrase" and "entailment:phrase" are misclassified by many systems. This result implies the essential difficulty of handling phrase-level semantic relations.

5.6 Results on RITE4QA subtask

Table 27 and Table 28 show the results on RITE4QA for Simplified Chinese and Traditional Chiense respectively. Regarding the WorseRanking, IMTKU improved the baseline QA system the most in CS (28% in Top1 accuracy) and WHUTE in CT (27.33% in Top1 accuracy). When considering the BetterRanking, unfortunately no RITE systems can improve the baseline QA system.

Table 29 and Table 30 show the results on CompleteRITE-4QA for Simplified Chinese and Traditional Chiense respectively. This year we mainly focus on RITE4QA subtask results. The evaluations of CompleteRITE4QA runs are provided for additional observation.

6. CONCLUSION

This paper described an overview of NTCIR-10 RITE-2 task. In RITE-2, in addition to the four subtasks in NTCIR-9 RITE (BC, MC, ExamBC and RITE4QA), the Exam Search subtask and the UnitTest subtask were added. We had 28 active participants, and received 215 runs (110 Japanese runs, 53 Traditional Chinese runs, 52 Simplified Chinese runs) for the formal run. The results of the formal have quite different trends especially in Japanese runs compared to NTCIR-9 RITE. In the Japanese BC subtask, the top system achieved over 80% in accuracy (in contrast to 58% in NTCIR-9 RITE). One of the reasons that causes such difference is that we filtered examples with less agreement from the dataset. Introducing the subtask of UnitTest enabled detailed analysis of RTE systems, because all of the examples contain only single linguistic phenomenon, and are annotated one of the defined categories of linguistic phenomena. Due to its difficulty, we could have only four participants for the Entrance Exam Subtask. However, since the task setting is more realistic than the traditional setting like the BC and MC subtasks, we believe that there is still room to explore such task setting.

7. REFERENCES

[1] D. Andrade, M. Tsuchida, T. Onishi, and K. Ishikawa. Detecting Contradiction in Text by Using Lexical Mismatch and Structural Similarity. In *Proceedings of the 10th NTCIR Conference*, 2013.

[2] T.-H. Chang, Y.-C. Hsu, C.-W. Chang, Y.-C. Hsu, and J.-I. Chang. KC99: A Prediction System for Chinese Textual Entailment Relation using Decision Tree. In *Proceedings of the 10th NTCIR Conference*, 2013.

[3] I. Dagan, O. Glickman, and B. Magnini. The Pascal Recognising Textual Entailment Challenge. In *Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment*, 2005.

[4] M.-Y. Day, C. Tu, S.-J. Huang, H.-C. Vong, and S.-W. Wu. IMTKU Textual Entailment System for Recognizing Inference in Text at NTCIR-10 RITE2. In *Proceedings of the 10th NTCIR Conference*, 2013.

[5] D. Giampiccolo, B. Magnini, I. Dagan, and B. Dolan. The third pascal recognizing textual entailment challenge. In *ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, 2007.

[6] S. Harabagiu and A. Hickl. Methods for using textual entailment in open-domain question answering. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meetion of the ACL*, 2006.

[7] S. Hattori and S. Sato. Team SKL's Strategy and Experience in RITE2. In *Proceedings of the 10th NTCIR Conference*, 2013.

[8] W.-J. Huang and C.-L. Liu. NCCU-MIG at NTCIR-10: Using Lexical, Syntactic, and Semantic Features for the RITE Tasks. In *Proceedings of the 10th NTCIR Conference*, 2013.

[9] L. Ku, E. T.-H. Chu, and N. Han. Extracting Features for Machine Learning in NTCIR-10 RITE Task. In *Proceedings of the 10th NTCIR Conference*, 2013.

[10] C.-J. Lin and Y.-C. Tu. The Description of the NTOU RITE System in NTCIR-10 . In *Proceedings of the 10th NTCIR Conference*, 2013.

[11] M. Liu, Y. Wang, Y. Li, and H. Hu. WUST at NTCIR-10 RITE-2 Task：Feature Combination Approach to Textual Entailment. In *Proceedings of the 10th NTCIR Conference*, 2013.

[12] T. Makino, S. Okajima, and T. Iwakura. FLL: Local Alignments based Approach for NTCIR-10. In *Proceedings of the 10th NTCIR Conference*, 2013.

[13] Y. Mehdad, M. Negri, and M. Federico. Towards cross-lingual textual entailment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 321–324, 2010.

[14] Y. Mehdad, M. Negri, and M. Federico. Using bilingual parallel corpora for cross-lingual textual entailment. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, pages 1336–1345, 2011.

[15] J. Mizuno, A. Sumida, and K. Inui. TKDDI group at NTCIR10-RITE2: Recognizing Texual Entailment

Team	MacroF1	Acc.	Y-F1	Y-Prec.	Y-Rec.	N-F1	N-Prec.	N-Rec.
*FLL-JA-UnitTest-01	77.77	90.87	94.84	94.39	95.28	60.71	62.96	58.62
*FLL-JA-UnitTest-03	76.98	91.29	95.13	93.61	96.70	58.82	68.18	51.72
JAIST-JA-UnitTest-02	74.52	89.21	93.87	93.87	93.87	55.17	55.17	55.17
*TKDDI-JA-UnitTest-03	74.00	85.89	91.58	96.35	87.26	56.41	44.90	75.86
*TKDDI-JA-UnitTest-01	73.51	85.48	91.32	96.34	86.79	55.70	44.00	75.86
*TKDDI-JA-UnitTest-02	73.51	85.48	91.32	96.34	86.79	55.70	44.00	75.86
TKDDI-JA-UnitTest-01	73.51	85.48	91.32	96.34	86.79	55.70	44.00	75.86
TKDDI-JA-UnitTest-02	73.51	85.48	91.32	96.34	86.79	55.70	44.00	75.86
BnO-JA-UnitTest-01	73.44	85.89	91.63	95.88	87.74	55.26	44.68	72.41
NTTD-JA-UnitTest-02	68.98	84.23	90.73	93.94	87.74	47.22	39.53	58.62
BnO-JA-UnitTest-03	68.78	80.08	87.56	97.13	79.72	50.00	35.82	82.76
JAIST-JA-UnitTest-01	67.36	79.67	87.40	96.05	80.19	47.31	34.38	75.86
NTTD-JA-UnitTest-01	61.99	74.27	83.60	95.18	74.53	40.38	28.00	72.41
THK-JA-UnitTest-01	53.26	71.37	82.35	89.94	75.94	24.18	17.74	37.93
*FLL-JA-UnitTest-02	51.35	77.59	87.08	88.35	85.85	15.63	14.29	17.24
NTTD-JA-UnitTest-03	47.91	53.94	65.63	95.50	50.00	30.19	18.46	82.76
BnO-JA-UnitTest-02	46.80	87.97	93.60	87.97	100.00	0.00	0.00	0.00
KYOTO-JA-UnitTest-02	45.35	48.96	59.41	98.90	42.45	31.28	18.67	96.55
KYOTO-JA-UnitTest-01	37.27	38.59	46.38	100.00	30.19	28.16	16.38	100.00
JAIST-JA-UnitTest-03	29.46	30.71	38.83	86.89	25.00	20.10	11.67	72.41
Baseline-JA-UnitTest-01	51.70	86.31	92.58	88.41	97.17	10.81	25.00	6.90

Table 25: Results on Unit Test (JA).

Based on Dependency Structure Alignment. In *Proceedings of the 10th NTCIR Conference*, 2013.

[16] H. Morita and K. Takeuchi. Construction of a Simple Inference System of Textural Similarity. In *Proceedings of the 10th NTCIR Conference*, 2013.

[17] M. Negri, A. Marchetti, Y. Mehdad, L. Bentivogli, and D. Giampiccolo. Semeval-2012 task 8: Cross-lingual textual entailment for content synchronization. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 399–407, 2012.

[18] M. Negri and Y. Mehdad. Creating a bi-lingual entailment corpus through translations with mechanical turk: $100 for a 10-day rush. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 212–216, 2010.

[19] M. Ohki, T. Suenaga, D. Satoh, Y. Nomura, and T. Takaki. Expanded Dependency Structure based Textual Entailment Recognition System of NTTDATA for NTCIR10-RITE2. In *Proceedings of the 10th NTCIR Conference*, 2013.

[20] M. Ohno, Y. Tsuboi, H. Kanayama, and K. Yoshikawa. IBM Group at NTCIR-10 RITE2: Textual Entailment using Temporal Dimension Reduction. In *Proceedings of the 10th NTCIR Conference*, 2013.

[21] N. Okazaki. Classias: a collection of machine-learning algorithms for classification, 2009.

[22] T. Okita. Local Graph Matching with Active Learning for Recognizing Inference in Text at NTCIR-10. In *Proceedings of the 10th NTCIR Conference*, 2013.

[23] P. Pakray, S. Bandyopadhyay, and A. Gelbukh. Binary-class and Multi-class based Textual Entailment System. In *Proceedings of the 10th NTCIR Conference*, 2013.

[24] M. Q. N. Pham, M. L. Nguyen, and A. Shimazu. JAIST Participation at NTCIR-10 RITE-2. In *Proceedings of the 10th NTCIR Conference*, 2013.

[25] H. Ren, H. Wu, C. Lv, D. Ji, and J. Wan. The WHUTE System in NTCIR-10 RITE Task. In *Proceedings of the 10th NTCIR Conference*, 2013.

[26] T. Shibata, S. Kurohashi, S. Kohama, and A. Yamamoto. Predicate-argument Structure based Textual Entailment Recognition System of KYOTO Team for NTCIR-10 RITE-2. In *Proceedings of the 10th NTCIR Conference*, 2013.

[27] C.-W. Shih, C. Liu, C.-W. Lee, and W.-L. Hsu. IASL RITE System at NTCIR-10. In *Proceedings of the 10th NTCIR Conference*, 2013.

[28] K. Shimada, Y. Seto, M. Omura, and K. Kurihara. KitAi: Textual Entailment Recognition System for NTCIR-10 RITE2. In *Proceedings of the 10th NTCIR Conference*, 2013.

[29] K. Shinzato, T. Shibata, D. Kawahara, and S. Kurohashi. TSUBAKI: An open search engine infrastructure for developing information access methodology. *Journal of Information Processing*, 52(12):216–227, 2011.12.

[30] Y. Takesue and T. Ninomiya. EHIME Textual Entailment System Using Markov Logic in NTCIR-10 RITE-2. In *Proceedings of the 10th NTCIR Conference*, 2013.

[31] D. Tatar, E. Tamaianu-Morita, A. Mihis, and D. Lupsa. Summarization by logic segmentation and text entailment. In *Proceedings of CICLING 2008*, 2008.

[32] R. Tian, Y. Miyao, T. Matsuzaki, and H. Komatsu. BnO at NTCIR-10 RITE: A Strong Shallow Approach and an Inference-based Textual Entailment Recognition System. In *Proceedings of the 10th NTCIR Conference*, 2013.

[33] X.-L. Wang, H. Zhao, and B.-L. Lu. BCMI-NLP Labeled-Alignment-Based Entailment System for NTCIR-10 RITE-2 Task. In *Proceedings of the 10th NTCIR Conference*, 2013.

[34] Y. Watanabe, Y. Miyao, J. Mizuno, T. Shibata, H. Kanayama, C.-W. Lee, C.-J. Lin, S. Shi, T. Mitamura, N. Kando, H. Shima, and K. Takeda. Overview of the Recognizing Inference in Text (RITE-2) at NTCIR-10. In *Proceedings of the 10th*

	*FLL-JA-UnitTest-01			*FLL-JA-UnitTest-03			JAIST-JA-UnitTest-02			TKDDI-JA-UnitTest-03			TKDDI-JA-UnitTest-01		
Macro F1	77.77			76.98			74.52			74.00			73.51		
	N-Prec	Rec	F1	N-Prec	Rec	F1	N-Prec	Rec	F1	N-Prec	Rec	F1	N-Prec	Rec	F1
case_alternation	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–
inference	–(0/ 0)	–	–	0(0/ 1)	–	–	0(0/ 1)	–	–	0(0/ 2)	–	–	0(0/ 2)	–	–
spatial	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–	0(0/ 1)	–	–	0(0/ 1)	–	–
implicit_relation	0(0/ 1)	–	–	–(0/ 0)	–	–	0(0/ 1)	–	–	0(0/ 2)	–	–	0(0/ 2)	–	–
list	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–
disagree:lex	100(1/ 1)	50	67	–(0/ 0)	0	0	–(0/ 0)	0	0	100(2/ 2)	100	100	100(2/ 2)	100	100
synonymy:phrase	0(0/ 3)	–	–	0(0/ 2)	–	–	0(0/ 4)	–	–	0(0/ 6)	–	–	0(0/ 7)	–	–
meronymy:lex	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–
apposition	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–
modifier	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–
transparent_head	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–
synonymy:lex	0(0/ 1)	–	–	0(0/ 1)	–	–	0(0/ 1)	–	–	0(0/ 1)	–	–	0(0/ 1)	–	–
nominalization	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–
coreference	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–
disagree:phrase	100(16/ 16)	64	78	100(15/ 15)	60	75	100(16/ 16)	64	78	100(20/ 20)	80	89	100(20/ 20)	80	89
temporal	0(0/ 1)	–	–	0(0/ 1)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–
disagree:modality	–(0/ 0)	0	0	–(0/ 0)	0	0	–(0/ 0)	0	0	–(0/ 0)	0	0	–(0/ 0)	0	0
entailment:phrase	0(0/ 3)	–	–	0(0/ 1)	–	–	0(0/ 6)	–	–	0(0/ 13)	–	–	0(0/ 13)	–	–
disagree:temporal	–(0/ 0)	0	0	–(0/ 0)	0	0	–(0/ 0)	0	0	–(0/ 0)	0	0	–(0/ 0)	0	0
hypernymy:lex	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–
scrambling	0(0/ 1)	–	–	0(0/ 1)	–	–	–(0/ 0)	–	–	0(0/ 2)	–	–	0(0/ 2)	–	–
clause	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–
relative_clause	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–

	TKDDI-JA-UnitTest-02			BnO-JA-UnitTest-01			NTTD-JA-UnitTest-02			BnO-JA-UnitTest-03			JAIST-JA-UnitTest-01		
Macro F1	73.51			73.44			68.98			68.78			67.36		
	N-Prec	Rec	F1	N-Prec	Rec	F1	N-Prec	Rec	F1	N-Prec	Rec	F1	N-Prec	Rec	F1
case_alternation	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–	0(0/ 1)	–	–	0(0/ 2)	–	–
inference	0(0/ 2)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–	0(0/ 2)	–	–	0(0/ 2)	–	–
spatial	0(0/ 1)	–	–	0(0/ 1)	–	–	–(0/ 0)	–	–	0(0/ 1)	–	–	0(0/ 1)	–	–
implicit_relation	0(0/ 2)	–	–	–(0/ 0)	–	–	0(0/ 1)	–	–	0(0/ 1)	–	–	0(0/ 5)	–	–
list	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–	0(0/ 1)	–	–	–(0/ 0)	–	–
disagree:lex	100(2/ 2)	100	100	100(1/ 1)	50	67	100(1/ 1)	50	67	100(1/ 1)	50	67	100(1/ 1)	50	67
synonymy:phrase	0(0/ 7)	–	–	0(0/ 8)	–	–	0(0/ 7)	–	–	0(0/ 10)	–	–	0(0/ 11)	–	–
meronymy:lex	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–	0(0/ 1)	–	–
apposition	–(0/ 0)	–	–	0(0/ 1)	–	–	–(0/ 0)	–	–	0(0/ 1)	–	–	0(0/ 1)	–	–
modifier	–(0/ 0)	–	–	0(0/ 1)	–	–	–(0/ 0)	–	–	0(0/ 2)	–	–	0(0/ 2)	–	–
transparent_head	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–
synonymy:lex	0(0/ 1)	–	–	0(0/ 1)	–	–	0(0/ 3)	–	–	–(0/ 0)	–	–	0(0/ 2)	–	–
nominalization	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–
coreference	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–	0(0/ 2)	–	–
disagree:phrase	100(20/ 20)	80	89	100(20/ 20)	80	89	100(16/ 16)	64	78	100(22/ 22)	88	94	100(21/ 21)	84	91
temporal	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–
disagree:modality	–(0/ 0)	0	0	–(0/ 0)	0	0	–(0/ 0)	0	0	100(1/ 1)	100	100	–(0/ 0)	0	0
entailment:phrase	0(0/ 13)	–	–	0(0/ 14)	–	–	0(0/ 11)	–	–	0(0/ 20)	–	–	0(0/ 8)	–	–
disagree:temporal	–(0/ 0)	0	0	–(0/ 0)	0	0	–(0/ 0)	0	0	–(0/ 0)	0	0	–(0/ 0)	0	0
hypernymy:lex	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–	0(0/ 1)	–	–
scrambling	0(0/ 2)	–	–	–(0/ 0)	–	–	0(0/ 2)	–	–	0(0/ 2)	–	–	0(0/ 4)	–	–
clause	–(0/ 0)	–	–	–(0/ 0)	–	–	0(0/ 1)	–	–	0(0/ 1)	–	–	–(0/ 0)	–	–
relative_clause	–(0/ 0)	–	–	–(0/ 0)	–	–	0(0/ 1)	–	–	0(0/ 1)	–	–	–(0/ 0)	–	–

	NTTD-JA-UnitTest-01			THK-JA-UnitTest-01			Baseline-JA-UnitTest-01			*FLL-JA-UnitTest-02			NTTD-JA-UnitTest-03		
Macro F1	61.99			53.26			51.7			51.35			47.91		
	N-Prec	Rec	F1	N-Prec	Rec	F1	N-Prec	Rec	F1	N-Prec	Rec	F1	N-Prec	Rec	F1
case_alternation	0(0/ 2)	–	–	0(0/ 1)	–	–	–(0/ 0)	–	–	0(0/ 1)	–	–	0(0/ 3)	–	–
inference	0(0/ 1)	–	–	0(0/ 2)	–	–	0(0/ 1)	–	–	0(0/ 1)	–	–	0(0/ 1)	–	–
spatial	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–	0(0/ 1)	–	–
implicit_relation	0(0/ 2)	–	–	0(0/ 15)	–	–	–(0/ 0)	–	–	0(0/ 3)	–	–	0(0/ 9)	–	–
list	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–	0(0/ 1)	–	–
disagree:lex	100(1/ 1)	50	67	100(1/ 1)	50	67	–(0/ 0)	–	–	100(1/ 1)	50	67	100(1/ 1)	50	67
synonymy:phrase	0(0/ 15)	–	–	0(0/ 10)	–	–	0(0/ 3)	–	–	0(0/ 8)	–	–	0(0/ 24)	–	–
meronymy:lex	0(0/ 1)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–	0(0/ 1)	–	–	0(0/ 1)	–	–
apposition	–(0/ 0)	–	–	0(0/ 1)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–
modifier	0(0/ 2)	–	–	0(0/ 3)	–	–	–(0/ 0)	–	–	0(0/ 3)	–	–	0(0/ 8)	–	–
transparent_head	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–	0(0/ 1)	–	–
synonymy:lex	0(0/ 4)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–	0(0/ 2)	–	–	0(0/ 6)	–	–
nominalization	–(0/ 0)	–	–	0(0/ 1)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–	0(0/ 1)	–	–
coreference	–(0/ 0)	–	–	0(0/ 4)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–	0(0/ 1)	–	–
disagree:phrase	100(20/ 20)	80	89	100(9/ 9)	36	53	100(2/ 2)	8	15	100(4/ 4)	16	28	100(22/ 22)	88	94
temporal	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–
disagree:modality	–(0/ 0)	0	0	–(0/ 0)	0	0	–(0/ 0)	0	0	–(0/ 0)	0	0	100(1/ 1)	100	100
entailment:phrase	0(0/ 21)	–	–	0(0/ 10)	–	–	0(0/ 1)	–	–	0(0/ 7)	–	–	0(0/ 37)	–	–
disagree:temporal	–(0/ 0)	0	0	100(1/ 1)	100	100	–(0/ 0)	0	0	–(0/ 0)	0	0	–(0/ 0)	0	0
hypernymy:lex	0(0/ 1)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–	0(0/ 2)	–	–
scrambling	0(0/ 2)	–	–	0(0/ 1)	–	–	0(0/ 1)	–	–	0(0/ 2)	–	–	0(0/ 6)	–	–
clause	0(0/ 2)	–	–	0(0/ 2)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–	0(0/ 3)	–	–
relative_clause	0(0/ 1)	–	–	–(0/ 0)	–	–	–(0/ 0)	–	–	0(0/ 2)	–	–	0(0/ 1)	–	–

	BnO-JA-UnitTest-02			KYOTO-JA-UnitTest-02			KYOTO-JA-UnitTest-01			JAIST-JA-UnitTest-03					
Macro F1	46.8			45.35			37.27			29.46					
	N-Prec	Rec	F1	N-Prec	Rec	F1	N-Prec	Rec	F1	N-Prec	Rec	F1			
case_alternation	–(0/ 0)	–	–	0(0/ 6)	–	–	0(0/ 5)	–	–	0(0/ 6)	–	–			
inference	–(0/ 0)	–	–	0(0/ 2)	–	–	0(0/ 2)	–	–	0(0/ 1)	–	–			
spatial	–(0/ 0)	–	–	–(0/ 0)	–	–	0(0/ 1)	–	–	–(0/ 0)	–	–			
implicit_relation	–(0/ 0)	–	–	0(0/ 12)	–	–	0(0/ 14)	–	–	0(0/ 16)	–	–			
list	–(0/ 0)	–	–	–(0/ 0)	–	–	0(0/ 1)	–	–	0(0/ 3)	–	–			
disagree:lex	–(0/ 0)	0	0	100(2/ 2)	100	100	100(2/ 2)	100	100	100(2/ 2)	100	100			
synonymy:phrase	–(0/ 0)	–	–	0(0/ 28)	–	–	0(0/ 31)	–	–	0(0/ 20)	–	–			
meronymy:lex	–(0/ 0)	–	–	–(0/ 0)	–	–	0(0/ 1)	–	–	0(0/ 1)	–	–			
apposition	–(0/ 0)	–	–	0(0/ 1)	–	–	0(0/ 1)	–	–	0(0/ 1)	–	–			
modifier	–(0/ 0)	–	–	0(0/ 11)	–	–	0(0/ 13)	–	–	0(0/ 42)	–	–			
transparent_head	–(0/ 0)	–	–	–(0/ 0)	–	–	0(0/ 1)	–	–	0(0/ 1)	–	–			
synonymy:lex	–(0/ 0)	–	–	0(0/ 5)	–	–	0(0/ 8)	–	–	0(0/ 8)	–	–			
nominalization	–(0/ 0)	–	–	0(0/ 1)	–	–	0(0/ 1)	–	–	0(0/ 1)	–	–			
coreference	–(0/ 0)	–	–	0(0/ 3)	–	–	0(0/ 4)	–	–	0(0/ 4)	–	–			
disagree:phrase	–(0/ 0)	0	0	100(24/ 24)	96	98	100(25/ 25)	100	100	100(17/ 17)	68	81			
temporal	–(0/ 0)	–	–	0(0/ 1)	–	–	–(0/ 0)	–	–	0(0/ 1)	–	–			
disagree:modality	–(0/ 0)	0	0	100(1/ 1)	100	100	100(1/ 1)	100	100	100(1/ 1)	100	100			
entailment:phrase	–(0/ 0)	–	–	0(0/ 38)	–	–	0(0/ 43)	–	–	0(0/ 24)	–	–			
disagree:temporal	–(0/ 0)	0	0	100(1/ 1)	100	100	100(1/ 1)	100	100	100(1/ 1)	100	100			
hypernymy:lex	–(0/ 0)	–	–	–(0/ 0)	–	–	0(0/ 1)	–	–	0(0/ 3)	–	–			
scrambling	–(0/ 0)	–	–	0(0/ 4)	–	–	0(0/ 6)	–	–	0(0/ 6)	–	–			
clause	–(0/ 0)	–	–	0(0/ 6)	–	–	0(0/ 9)	–	–	0(0/ 14)	–	–			
relative_clause	–(0/ 0)	–	–	0(0/ 4)	–	–	0(0/ 6)	–	–	0(0/ 7)	–	–			

Table 26: The details of the results on the UnitTest data (performance for "N").

	WorseRanking						BetterRanking					
CS	R			R+U			R			R+U		
Run	Top1	MRR	Top5	Top1	MRR	Top5	Top1	MRR	Top5	Top1	MRR	Top5
IMTKU-CS-RITE4QA-03	28.00	33.77	42.67	33.33	40.87	53.33	28.00	33.77	42.67	33.33	40.87	53.33
WHUTE-CS-RITE4QA-01	18.67	27.59	43.33	22.00	33.67	54.00	18.67	27.64	43.33	22.67	34.06	54.00
IMTKU-CS-RITE4QA-02	14.67	21.44	36.00	20.00	29.08	47.33	14.67	21.44	36.00	20.00	29.08	47.33
*IASL-CS-RITE4QA-02	12.67	18.80	32.00	15.33	23.13	39.33	38.00	42.71	50.00	42.00	47.27	56.00
*IASL-CS-RITE4QA-01	12.00	22.71	42.00	15.33	27.96	50.67	31.33	38.09	48.67	36.00	43.74	56.67
IMTKU-CS-RITE4QA-01	10.67	19.91	38.67	17.33	27.21	47.33	10.67	19.91	38.67	17.33	27.21	47.33
CYUT-CS-RITE4QA-02	7.33	11.82	23.33	10.67	17.32	32.67	39.33	46.13	55.33	44.00	51.52	62.67
CYUT-CS-RITE4QA-01	6.67	11.49	23.33	10.00	16.99	32.67	38.00	45.19	54.67	43.33	51.08	62.00
bcNLP-CS-RITE4QA-03	6.67	8.78	11.33	7.33	9.78	12.67	8.00	9.33	10.67	9.33	10.67	12.00
bcNLP-CS-RITE4QA-01	3.33	3.67	4.00	4.67	6.17	8.00	2.67	3.22	4.00	6.00	6.78	8.00
bcNLP-CS-RITE4QA-02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
orgQAsys-CS-RITE4QA-01	7.33	11.32	22.00	10.67	16.99	31.33	40.67	47.60	57.33	44.67	52.32	64.00

Table 27: Results on RITE4QA subtask (CS).

	WorseRanking						BetterRanking					
CT	R			R+U			R			R+U		
Run	Top1	MRR	Top5	Top1	MRR	Top5	Top1	MRR	Top5	Top1	MRR	Top5
WHUTE-CT-RITE4QA-01	27.33	34.57	46.67	30.67	38.76	52.67	26.67	34.29	46.67	30.00	38.48	52.67
IMTKU-CT-RITE4QA-03	16.67	25.70	40.67	19.33	29.39	46.00	17.33	26.03	40.67	20.00	29.72	46.00
IMTKU-CT-RITE4QA-01	14.67	22.69	37.33	22.00	31.84	49.33	14.67	22.58	37.33	22.00	31.73	49.33
CYUT-CT-RITE4QA-03	12.67	17.69	28.67	18.67	26.60	42.67	12.00	17.50	28.67	18.67	26.99	43.33
IASL-CT-RITE4QA-01	12.00	22.14	39.33	14.67	26.86	48.00	30.00	38.27	50.67	32.67	42.34	57.33
IMTKU-CT-RITE4QA-02	12.00	19.82	32.00	19.33	29.43	44.67	12.00	19.84	32.67	20.00	29.79	45.33
IASL-CT-RITE4QA-02	10.67	16.01	27.33	14.67	22.09	37.33	38.00	42.74	49.33	42.00	47.66	56.67
NTOUA-CT-RITE4QA-03	8.00	9.97	13.33	12.67	17.19	24.00	9.33	10.63	13.33	14.67	18.30	24.00
NTOUA-CT-RITE4QA-01	8.00	9.28	11.33	13.33	17.06	22.67	8.67	9.61	11.33	15.33	18.17	22.67
NTOUA-CT-RITE4QA-02	7.33	8.78	10.67	11.33	14.11	18.00	8.00	9.22	10.67	12.00	14.56	18.00
CYUT-CT-RITE4QA-01	6.67	11.71	24.00	10.00	16.99	32.67	38.00	45.19	54.67	43.33	51.08	62.00
CYUT-CT-RITE4QA-02	6.67	11.71	24.00	10.00	16.99	32.67	38.00	45.19	54.67	43.33	51.08	62.00
orgQAsys-CT-RITE4QA-01	7.33	11.54	22.67	10.67	16.99	31.33	40.67	47.60	57.33	44.67	52.32	64.00

Table 28: Results on RITE4QA subtask (CT).

	WorseRanking						BetterRanking					
CS	R			R+U			R			R+U		
Run	Top1	MRR	Top5	Top1	MRR	Top5	Top1	MRR	Top5	Top1	MRR	Top5
WHUTE-CS-CompleteRITE4QA-01	14.00	20.79	34.00	15.33	23.74	40.67	14.00	20.84	34.00	16.00	24.22	40.67
*IASL-CAD-CS-CompleteRITE4QA-01	6.00	14.59	28.67	6.67	16.26	32.00	28.67	34.37	44.00	33.33	39.59	50.00
*IASL-CAD-CS-CompleteRITE4QA-02	4.00	6.63	14.00	4.00	7.47	16.00	38.00	42.93	50.67	42.00	47.49	56.67
CYUT-CS-CompleteRITE4QA-03	1.33	3.78	10.00	1.33	5.84	16.00	40.67	47.60	57.33	44.67	52.32	64.00
CYUT-CS-CompleteRITE4QA-01	1.33	3.09	8.00	1.33	3.48	9.33	38.00	45.19	54.67	43.33	51.08	62.00
CYUT-CS-CompleteRITE4QA-02	1.33	3.09	8.00	1.33	3.48	9.33	39.33	46.13	55.33	44.00	51.52	62.67
orgQAsys-CS-CompleteRITE4QA-01	1.33	3.16	7.33	1.33	3.32	8.00	40.67	47.60	57.33	44.67	52.32	64.00

Table 29: Results on CompleteRITE4QA subtask (CS).

	WorseRanking						BetterRanking					
CT	R			R+U			R			R+U		
Run	Top1	MRR	Top5	Top1	MRR	Top5	Top1	MRR	Top5	Top1	MRR	Top5
WHUTE-CT-CompleteRITE4QA-01	20.67	29.26	42.67	22.00	31.72	47.33	20.00	28.98	42.67	22.00	31.89	47.33
CYUT-CT-CompleteRITE4QA-03	8.67	13.48	22.00	14.00	21.89	35.33	11.33	14.96	22.00	17.33	23.94	36.00
IASL-CAD-CT-CompleteRITE4QA-01	5.33	13.68	27.33	6.00	15.01	29.33	26.67	34.23	46.67	29.33	37.68	51.33
IASL-CAD-CT-CompleteRITE4QA-02	4.00	7.50	16.00	5.33	10.00	20.67	38.00	42.74	49.33	42.00	47.66	56.67
NTOUA-CT-CompleteRITE4QA-03	2.67	4.97	8.67	5.33	8.99	15.33	4.00	5.67	8.67	6.00	9.52	15.33
NTOUA-CT-CompleteRITE4QA-01	2.67	4.83	8.67	4.67	8.27	15.33	3.33	5.17	8.67	5.33	8.72	15.33
NTOUA-CT-CompleteRITE4QA-02	2.67	4.80	8.00	4.67	8.62	14.67	3.33	5.28	8.00	5.33	9.10	14.67
CYUT-CT-CompleteRITE4QA-02	2.00	3.76	8.67	2.00	4.14	10.00	38.00	45.19	54.67	43.33	51.08	62.00
CYUT-CT-CompleteRITE4QA-01	1.33	3.09	8.00	1.33	3.48	9.33	38.00	45.19	54.67	43.33	51.08	62.00
orgQAsys-CT-CompleteRITE4QA-01	1.33	3.16	7.33	1.33	3.32	8.00	40.67	47.60	57.33	44.67	52.32	64.00

Table 30: Results on CompleteRITE4QA subtask (CT).

NTCIR Conference, 2013.

[35] Y. Watanabe, J. Mizuno, and K. Inui. THK's Natural Logic-based Compositional Textual Entailment Model at NTCIR-10 RITE-2. In *Proceedings of the 10th NTCIR Conference*, 2013.

[36] S.-H. Wu, S.-S. Yang, L.-P. Chen, H.-S. Chiu, and R.-D. Yang. CYUT Chinese Textual Entailment Recognition System for NTCIR-10 RITE-2. In *Proceedings of the 10th NTCIR Conference*, 2013.

[37] Y.-C. Wu, Y.-S. Lee, and J.-C. Yang. Combining Multiple Lexical Resources for Chinese Textual Entailment Recognition. In *Proceedings of the 10th NTCIR Conference*, 2013.

[38] H. Yamana, D. Ito, and M. Tanaka. WSD Team's Approaches for Textual Entailment Recognition at the NTCIR10 (RITE2). In *Proceedings of the 10th NTCIR Conference*, 2013.