

Overview of the NTCIR-11 Recognizing Inference in Text and Validation (RITE-VAL) Task

Suguru Matsuyoshi
University of Yamanashi,
Japan
sugurum@yamanashi.ac.jp

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

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

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

Cheng-Wei Shih
Academia Sinica, Taiwan
dapi@iis.sinica.edu.tw

Yotaro Watanabe
NEC Corporation, Japan
yotaro.w@gmail.com

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

ABSTRACT

This paper describes an overview of Recognizing Inference in Text and Validation (RITE-VAL) task in NTCIR-11. We evaluated systems that automatically recognize semantic relations between sentences such as entailment, contradiction and independence in Japanese (JA), English (EN), Simplified Chinese (CS) and Traditional Chinese (CT). RITE-VAL task has the following two subtasks: Fact Validation subtask (FV) and System Validation subtask (SV). SV consists of binary classification subtask (SVBC) and multi-classification subtask (SVMC). We had 23 active participating teams, and received 170 formal runs (59 Japanese runs, 9 English runs, 53 Simplified Chinese runs and 49 Traditional Chinese runs). This paper also describes how the datasets for RITE-VAL had been developed, how the systems were evaluated, and reports RITE-VAL formal run results.

Keywords

entailment, contradiction, fact validation, entrance exam, linguistic phenomenon

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 [1]. 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. [3]). In document summarization, we can remain necessary information (e.g. [9]) 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) [1]. From the third PASCAL RTE challenge (RTE-3), the task has included recognizing not only entailment relations, but also contradiction relations [2]. 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 [5, 4, 7]. In the SemEval-2012 Cross-lingual Textual Entailment (CLTE) [6], 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,

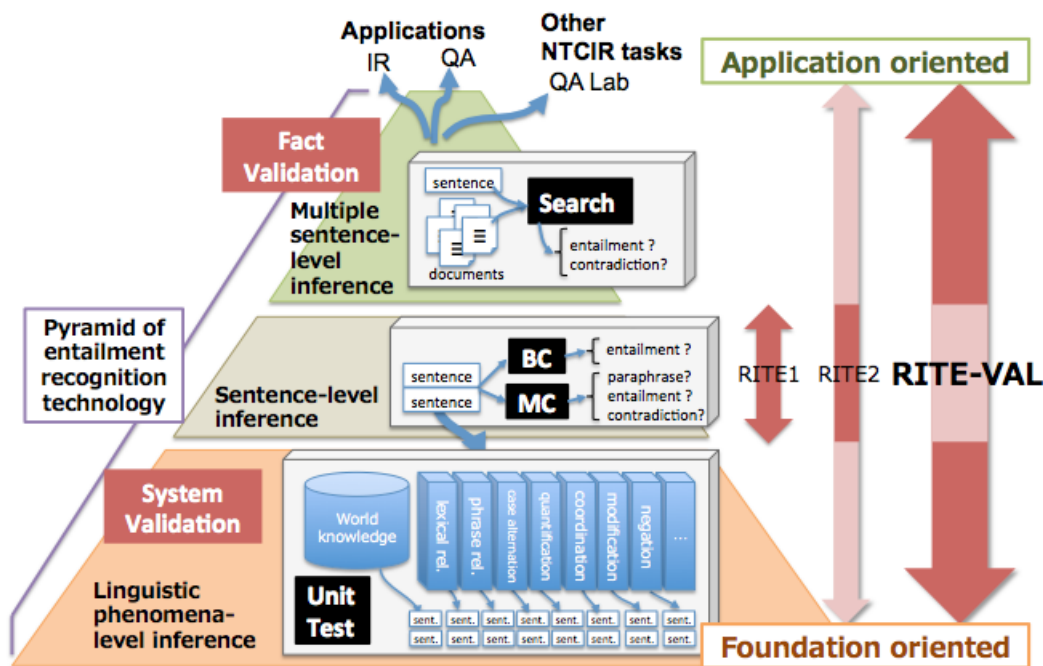


Figure 1: The Two Subtasks of RITE-VAL in a Pyramid of Entailment Recognition Technology.

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 [10], 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.

In the third round of RITE, RITE-VAL, we focus on the two subtask settings as shown in Figure 1: multiple sentence-level inference (Search subtask) and linguistic phenomena-level inference (UnitTest subtask). Exploration of Search

subtask setting is necessary to encourage applying textual inference to several information access applications because they require handling multiple sentences or documents in fact. Also, conducting UnitTest with large-scale dataset provides more plausible evaluation of textual inference systems, and promotes progress of textual inference technologies.

We took priority for sentence-level inference at NTCIR-9 and tackled to foundation-oriented problems and application-oriented problems at NTCIR-10. The former made the limitation of the current NLP technology clear and the latter made the performance of the current RTE technology clear. From this background, we will improve these subtasks at NTCIR-11. More concretely, we will enrich the linguistic variety for UnitTest and consider not only “Y” (entailment) and “N” (non-entailment) but also “C” (contradiction) for Search. We are considering the following subtasks:

Fact Validation Given a text t_2 , a system identifies whether t_2 is entailed from the sentences relevant to t_2 , which are retrieved from Wikipedia or textbook.

System Validation Although the setting is the same as the MC subtask, the text pair includes a breakdown of linguistic phenomena, such as synonymy of word/phrase, modifier, inference, etc.

The improvement of the technology of recognizing inference can contribute to several NLP applications, such as Information Retrieval, Question Answering, Machine Translation, and other tasks in NTCIR-11. Since the data in the Search subtask is constructed from the National Center Test for University admission, the improvement in the Search subtask naturally contributes to Todai Robot Project by NII.

We are going to use thousands of text pairs built in NTCIR-9 and NTCIR-10 RITE as training data. In addition, we are

planning to create training/test data especially for Fact Validation and System Validation subtasks.

We try to address the following important aspects when designing and organizing the task.

- Using crowd sourcing to create dataset. In order to create UnitTest dataset, it is necessary to consider various linguistic phenomena deeply. For this reason, the data size of UnitTest is limited at NTCIR-10. At NTCIR-11, in order to increase the data and to cover more various linguistic phenomena, we are planning to use crowd sourcing service such as Yahoo! Crowd Sourcing to create the dataset.
- Accountability. We will again encourage participants to do an ablation study which is done by removing one resource, tool, or algorithm at a time, and see its impact to the overall system (lower performance indicates higher importance). Remember that participants can take advantage of automatic evaluation in the RITE task, and quickly try out multiple different experiments. In that way, participants can avoid a system from being a complicated black box, but can instead see it as a collection of building blocks.

The important dates for NTCIR-11 RITE-VAL were as follows:

Apr 30, 2014	Development data released
Jul 25, 2014	Registration due
Jul 25, 2014	Formal run data (for CS and CT) released
Aug 4, 2014	Formal run data (for JA and EN) released
Aug 7, 2014	Formal run (for CS, CT and JA) results submission due
Aug 9, 2014	Results of the formal run (for CS, CT and JA) released
Aug 10, 2014	Formal run (for EN) results submission due
Aug 11, 2014	Results of the formal run (for EN) released
Aug 21, 2014	Formal run data (for JA and EN) with the correct labels released
Aug 29, 2014	Formal run data (for CS and CT) with the correct labels released

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

2. TASK OVERVIEW

This section describes the subtasks in RITE-VAL and how the dataset for each subtask was developed. We do not have System Validation subtask for English.

2.1 Fact Validation subtasks for Chinese

Fact Validation Subtask (FV) succeeds the ExamSearch Subtask in NTCIR-10 RITE-2, and it is the first attempt of a Chinese-version FV subtask. This subtask aims to answer single-choice or multiple-choice questions of real university entrance exams, by referring to textual knowledge i.e. Wikipedia.

Unlike BC or MC RITE datasets where both t_1 and t_2 would be given, FV datasets only contain t_2 . FV systems are asked to retrieve texts that can be used as t_1 from Wikipedia, and answer whether t_2 is entailed (inferred) from retrieved

		E	C	U	Total
CS	Training	239	82	155	476
	Test	222	201	190	613
	Total	461	283	345	1,089
CT	Training	243	84	162	489
	Test	222	201	190	613
	Total	465	285	352	1,102

Table 1: Statistics of the Chinese Fact Validation datasets.

texts. Therefore, the FV subtask is closer to the actual situation of answering entrance exams.

Chinese FV subtask is a multi-classification task where labels are defined as follows.

- E (entailment): information provided in Chinese Wikipedia supports the validity of a sentence
- C (contradiction): information provided in Chinese Wikipedia contradicts a sentence
- U (unknown): no known information provided in Chinese Wikipedia supports or contradicts a sentence

The statistics of Chinese Fact Validation datasets are shown in Table 1.

2.1.1 Textual Knowledge Base: Chinese Wikipedia

The textual knowledge base of Chinese FV subtask is Chinese Wikipedia. There are more than 100 thousands of Wikipedia articles in Chinese, not including dialects (such as Cantonese or Hakka) or Classical Chinese. Articles in Chinese Wikipedia can be written in Traditional Chinese or Simplified Chinese. They are translated into appropriate Chinese character sets and terms according to the readers' locations before showing in the browser.

For the sake of building a static FV benchmark, the textual knowledge base is restricted to be the archive of Chinese Wikipedia dumped on Feb. 10, 2014. All main articles in Chinese Wikipedia were translated into one kind of Chinese. We adopted the general translation dictionary for Chinese from Wikipedia to do the translation. Vocabulary was set to be in Taiwan for Traditional Chinese FV subtask and Mainland China for Simplified Chinese FV subtask.

2.1.2 Training and Test sets

Test sets

The formal test set of Traditional Chinese FV Subtask was created from the questions in the University Entrance Examinations¹, General Scholastic Ability Test (GSAT), held in Taiwan in 1994 to 1997. Only tests of Science (Physics, Chemistry, Biology, and Geology) and Social Studies (History, Geography, and Civics) were adopted, for their question were written in a more straight-forward way and it would be more possible to find answers in Wikipedia.

All questions were combined with their options and rewritten into declarative sentences. Experts in science and social studies were hired to seek supporting or contradicting evidences in Chinese Wikipedia. Under the limitation of time, the experts would assign "unknown" labels on questions they could not find evidence in a short time.

¹<http://www.ceec.edu.tw/AbilityExam/AbilityExamPaper.htm>

	Y	N	Total
Training	383	575	958
Test	206	308	514
Total	589	883	1,472

Table 2: Statistics of the Japanese Fact Validation datasets.

When the Traditional Chinese FV test set was ready, it was translated into Simplified Chinese in the same way to translate Chinese Wikipedia (cf. Section 2.1.1).

Training sets

Due to lack of materials to prepare training data, we selected a subset of RITE-2 Chinese test sets and extracted t_2 's as Chinese FV training sets. The creation of these selected RITE2 pairs included manually searching in Wikipedia, therefore the sentences and labels were usable in some extent.

The t_2 's in CT-FV training set come from pairs in RITE2-CT-MC-testset whose IDs are in 389 881, and t_2 's in CT-FV training set come from pairs in RITE2-CS-MC-testset whose IDs are in 302 781.

2.2 Fact Validation subtask for Japanese

The fact validation task (FV) in Japanese aims to answer questions validating a statement (given as a sentence) using documents describing facts, such as Wikipedia and textbooks. This task involves two challenges. One is to retrieve evidence texts from entire documents to prove whether a given statement is true or false. This task looks similar to typical information retrieval, but, in fact, retrieving evidences for false statements is not trivial, because false statements may include words/phrases that do not appear in evidence texts. The other challenge is to judge entailment relations between statements and evidential texts. In general evidential texts are an excerpt from huge documents, and necessary information does not always appear completely in a sentence. Therefore, document-level NLP technologies like coreference resolution are crucial.

The data for the JA FV task is created from the Center Test, which is a standardized achievement test for university admission in Japan. The test consists of various subjects, including Physics and Mathematics, while the data set for the RITE-VAL fact validation task is taken from social studies; namely, World History, Japanese History, Modern Society, and Politics & Economics. World History and Japanese History have two sub-categories for each (A and B), and, in total, tests from six subjects are used. These subjects mainly ask for knowledge of historical facts in multiple-choice questions with 4 options, which can be regarded as a limited form of fact validation.

The task data is a set of sentences (t_2), which are taken directly from choices of multiple-choice questions in the above-mentioned subjects. In some cases, key phrases are mentioned in a question sentence (e.g. "Choose a correct statement that describes a historical event in 7th century"), and we manually added the key phrase ("In 7th century") to each of the choices.

The statistics of Japanese Fact Validation datasets are shown in Table 2.

We provided three document texts; a snapshot of Japanese Wikipedia, and two textbooks of World/Japanese History. We applied only format conversion and no manual process-

	Y	N	Total
Training	141	238	379
Test	74	114	188
Total	215	352	567

Table 3: Statistics of the English Fact Validation datasets.

	BC		MC				Total
	Y	N	B	F	C	I	
Training	370	211	222	148	152	59	581
Test	600	600	300	300	300	300	1,200
Total	970	811	522	448	452	359	1,781

Table 4: Statistics of the Chinese System Validation datasets.

ing.

2.3 Fact Validation subtask for English

The data set for the English fact validation is created from English translation of the Center Test data. Expert native speakers translated texts of World History B and Politics & Economics of the Center Test. In a way similar to the Japanese FV data, we extracted choices of the questions from the translated texts to obtain the data set for English. By following the Japanese data set, we manually added key phrases in question sentences if necessary.

The statistics of English Fact Validation datasets are shown in Table 3.

2.4 System Validation subtasks for Chinese

System Validation Subtask (SV) succeeds the UnitTest Subtask in NTCIR-10 RITE-2, and it is the first attempt of a Chinese-version SV subtask. This subtask aims to examine the ability of RITE systems to handle different entailment-related linguistic phenomena. Annotating relevant linguistic phenomena on RITE sentence pairs can not only estimate the proposed effects to a specific sub-problem in textual entailment recognition, but also provide participants a more precise diagnostic tool to their system.

In Chinese System Validation Subtask, 28 linguistic phenomena are defined, where 19 of them are related to entailment and 9 of them related to contradiction.

Chinese SV training sets and test sets were created in the same way. We randomly selected over 100 pages from Chinese Wikipedia for extracting source sentences. Based on these wiki contents, our annotators create pairs by following the rule that only one linguistic phenomenon appears in a single pair.

There are two Chinese SV subtasks, binary classification (SVBC) and multi-classification (SVMC). Each created pair was assigned a BC label and a MC label by annotators. Numbers of different classes were balanced in SV test sets.

Similar to FV dataset preparation, after the Traditional Chinese SV training set and test set were ready, they were translated into Simplified Chinese in the same way to translate Chinese Wikipedia (cf. Section 2.1.1). The statistics of Chinese SV dataset are shown in Table 4 and the categories of linguistic phenomena appeared in the subtask are listed in Table 5.

Category	Training	Test
Linguistic phenomena related to entailment		
abbreviation	6	25
apposition	7	25
case_alternation	21	27
clause	25	59
coreference	11	24
hypernymy	30	27
inference	75	184
lexical_entailment	12	29
list	20	37
meronymy	4	23
modifier	37	131
paraphrase	47	49
quantity	11	29
relative_clause	6	36
scrambling	27	35
spatial	18	42
synonymy:lex	48	51
temporal	11	40
transparent_head	13	26
Linguistic phenomena related to contradiction		
antonym	20	35
exclusion:common_sense	8	34
exclusion:modality	12	38
exclusion:modifier	14	33
exclusion:predicate_argument	51	38
exclusion:quantity	6	29
exclusion:spatial	14	32
exclusion:temporal	7	34
negation	20	28
Total	581	1200

Table 5: Linguistic phenomena contained in the Chinese system validation.

	Y	N	Total
Training	1,330	1,362	2,692
Test	339	1,040	1,379
Total	1,669	2,402	4,071

Table 6: Statistics of the Japanese System Validation datasets.

2.5 System Validation subtask for Japanese

System Validation Subtask aims to make the participants validate their own system used in Fact Validation subtask.

In NTCIR-9 RITE-1 and NTCIR-10 RITE-2, the datasets were constructed by the college students, and it takes much cost and time. To alleviate this problem, in the Japanese System Validation Subtask, we attempt to use crowdsourcing, “Lancers”², to construct the dataset.

We first construct a set of “t1” and “t2” by using Japanese fact validation task. For each “t2”, the similar sentences to “t2” are retrieved from Wikipedia and textbooks, and are regarded as “t1”.

Then, we give 2,000 sets of “t1” and “t2” to workers, and they assign the label “Y” or “N” to each set. Each set is labeled by 5 workers. We can construct the dataset for about one day. Since there are some unreliable workers, the sets where more than 3 workers assign the same label are adopted for the dataset, and 1,379 sets were obtained. The statistic of Japanese System Validation dataset is shown in Table 6.

2.6 Data Format

Figure 2 shows XML data format for RITE-VAL. This file format is the same for NTCIR-10 RITE-2. The top ele-

²<http://www.lancers.jp/>

```
<?xml version='1.0' encoding='UTF-8' ?>
<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。
    </t1>
    <t2>
      川端康成「雪国」の著者である。
    </t2>
  </pair>
  <pair id='2' label='N'>
    <t1>
      :
    </t1>
  </pair>
</dataset>
```

Figure 2: XML data format for RITE-VAL.

ment of the XML data is a <dataset> element. It contains <pair> elements each of which represents a text pair. A <pair> element has <t1> and <t2> elements which contain texts t_1 and t_2 respectively, and @id and @label attributes. The @label attribute for a text pair takes a label “Y” or “N” for SVBC, JA-FV and EN-FV; “B,” “F,” “C,” or “I” for SVMC; “E,” “C,” or “U” for CS-FV and CT-FV.

For Fact Validation Subtask, a <pair> element includes no <t1> element. A <pair> element in task data for system training purposes has a @label attribute, while one in task data for formal run does not.

In the Fact Validation Subtask, to determine the correctness of statement t_2 , some relevant sentences with t_2 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 [8] as a search engine, and in the search results, for each t_2 , 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.

2.7 Evaluation Method

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.

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

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

In the Fact Validation Subtask, 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 Fact Validation Subtask, we

```
<?xml version="1.0" encoding="UTF-8"?>
<dataset type='bc'>
  <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>
        :
        :
      </pair>
    </dataset>
```

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

Subtask	JA	EN	CS	CT	Total
FV	30	9	12	15	66
FVsearch	3	0	-	-	3
SVBC	26	-	23	17	66
SVMC	-	-	18	17	35
Total	59	9	53	49	170

Table 8: Number of Submissions.

also evaluate precision/recall of t_1 search results that are retrieved from Wikipedia or textbooks by the systems.

3. PARTICIPANTS

Table 7 lists the RITE-VAL participants. The participants were from Japan (11), Taiwan (7), China (4), Norway (1) and Vietnam (1), and 23 teams in Total³. NTCIR-11 RITE-VAL has five less teams than NTCIR-10 RITE-2 had.

Table 8 shows the number of the submitted runs in the RITE-VAL formal run. Compared to NTCIR-10 RITE-2, the number of submissions decreased, however, many participants attended the subtasks for Chinese. The total number of the formal runs was 170, which is 45 less than that in NTCIR-10 RITE-2. In the subtasks for Japanese and Traditional Chinese, the numbers of the submitted runs for FV and SVBC (or SVMC) were almost the same. In the subtasks for Simplified Chinese, the number of the runs for SVBC was almost twice as much as that for FV.

4. FORMAL RUN RESULTS

In the formal run, participants could submit up to five runs for each subtask. The submission names in the tables follow the naming rule: (TEAMID)-(LANGUAGE)-(SUBTASK NAME)-(RUN NUMBER).

4.1 Results on FV Subtasks

Tables 9 and 10 show the results of FV subtasks for Simplified Chinese and Traditional Chinese respectively. E/C/U

³One team consists of people from Japan and Vietnam.

prec./rec. are for precision/recall of each label (entailment, contradiction, unknown). Table 11 shows the results of FV subtask for Japanese, and Table 12 shows the results of FV subtask for English. Since these data sets are created from university entrance examinations, we additionally show correct answer ratios of exams for Japanese and English subtasks. The results are sorted by Macro F1 scores. JA, JB, MS, PE, WA, WB stand for Japanese History A/B, Modern Society, Politics & Economics, Worls History A/B, respectively.

4.2 Results on CS/CT SVBC, SVMC subtasks

Tables 13 and 14 show the results of SVBC subtasks for Simplified Chinese and Traditional Chinese respectively. Tables 15 and 16 show the results of SVMC subtasks for Simplified Chinese and Traditional Chinese respectively. The definition of BC/MC follows the task definition of RITE2. The tables additionally show precision/recall for each label.

4.3 Results on JA SV subtask

Tables 17 shows the results of SV subtask for Japanese. The evaluation measures are Macro F1, accuracy, and precision/recall/F1 for Y/N labels, and the results are sorted by Macro F1 scores.

5. CONCLUSION

This paper described an overview of NTCIR-11 RITE-VAL task. We have provided datasets for fact validation and system validation in Japanese, English, Simplified Chinese, and Traditional Chinese. We had 23 participating teams and received 170 formal run results.

6. REFERENCES

- [1] 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.
- [2] D. Giampiccolo, B. Magnini, I. Dagan, and B. Dolan. The third pascal recognizing textual entailment

ID	Organization	Country /Region	JA	EN	CS	CT
1	ASNLP	Academia Sinica	Taiwan		✓	✓
2	BnO	National Institute of Informatics	Japan			
3	BUPTTeam	Beijing University of Posts and Telecommunications	China	✓	✓	
4	CL	Nara Institute of Science and Technology	Japan	✓		
5	III&CYUT	Institute for Information Industry and Chaoyang University of Technology	Taiwan		✓	✓
6	IMTKU	Tamkang University	Taiwan		✓	✓
7	JAVN	VNU University of Engineering and Technology	Japan, Vietnam	✓	✓	✓
8	KitAi	Kyushu Institute of Technology	Japan	✓		
9	KJP	Shizuoka University	Japan	✓		
10	KSU	Kyoto Sangyo University	Japan	✓		
11	KTU	Kyoto University	Japan	✓		
12	KUAS	National Kaohsiung University of Applied Sciences	Taiwan			✓
13	MCU	Ming-Chuan University	Taiwan			✓
14	MIG	National Chengchi University	Taiwan		✓	✓
15	NAK	Keio University	Japan	✓		
16	NTOUA	National Taiwan Ocean University	Taiwan			✓
17	NUL	Nihon Unisys, Ltd.	Japan	✓		
18	NWNU	Northwest Normal University	China		✓	
19	SITLP	Shibaura Institute of Technology	Japan	✓		
20	SKL	Nagoya University	Japan	✓		
21	WHUTE	Wuhan University	China		✓	✓
22	WUST	Wuhan University of Science and Technology	China		✓	
23	Yamraj	Norwegian University of Science and Technology	Norway	✓		

Table 7: The participants in RITE-VAL.

Run	MacroF1	Acc.	E-F1	E-Prec.	E-Rec.	C-F1	C-Prec.	C-Rec.	U-F1	U-Prec.	U-Rec.
III&CYUT-CS-FV-05	38.93	44.05	55.71	45.22	72.52	14.56	31.67	9.45	46.51	45.69	47.37
III&CYUT-CS-FV-02	38.78	43.88	56.75	45.73	74.77	15.73	31.82	10.45	43.85	44.57	43.16
WHUTE-CS-FV-02	38.08	41.92	43.41	47.34	40.09	18.73	37.88	12.44	52.09	39.83	75.26
III&CYUT-CS-FV-03	37.00	41.76	54.48	44.67	69.82	14.02	27.14	9.45	42.49	41.84	43.16
III&CYUT-CS-FV-04	36.89	43.39	54.13	44.38	69.37	8.66	33.33	4.98	47.89	43.22	53.68
III&CYUT-CS-FV-01	36.76	41.92	56.61	45.38	75.23	16.00	29.73	10.95	37.67	39.77	35.79
WHUTE-CS-FV-01	35.94	39.97	39.20	44.32	35.14	17.10	33.82	11.44	51.52	39.02	75.79
MIG-CS-FV-02	31.07	35.89	48.78	38.17	67.57	15.27	28.38	10.45	29.17	33.56	25.79
MIG-CS-FV-03	30.88	36.70	49.46	37.77	71.62	26.93	31.76	23.38	16.24	43.18	10.00
MIG-CS-FV-04	28.73	35.89	49.61	37.83	72.07	27.17	29.94	24.88	9.39	43.48	5.26
MIG-CS-FV-01	27.90	35.40	49.28	36.23	77.03	13.90	31.03	8.96	20.51	33.73	14.74
MIG-CS-FV-05	26.02	36.87	52.77	37.72	87.84	13.53	27.69	8.96	11.76	41.94	6.84

Table 9: Results on FV subtask (CS).

- challenge. In *ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, 2007.
- [3] 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 Meeting of the ACL*, 2006.
- [4] 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.
- [5] 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.
- [6] 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.
- [7] 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.
- [8] 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.
- [9] D. Tatar, E. Tamaianu-Morita, A. Mihiş, and D. Lupsa. Summarization by logic segmentation and text entailment. In *Proceedings of CICLING 2008*, 2008.
- [10] 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 NTCIR Conference*, pages 385–404, 2013.

Run	MacroF1	Acc.	E-F1	E-Prec.	E-Rec.	C-F1	C-Prec.	C-Rec.	U-F1	U-Prec.	U-Rec.
III&CYUT-CT-FV-02	39.51	44.70	56.61	45.38	75.23	15.15	31.75	9.95	46.77	47.80	45.79
III&CYUT-CT-FV-05	39.36	44.54	55.61	45.10	72.52	14.79	33.93	9.45	47.69	46.50	48.95
MCU-CT-FV-01	39.27	43.07	37.33	71.79	25.23	30.33	86.05	18.41	50.15	34.76	90.00
MCU-CT-FV-02	39.27	43.07	37.33	71.79	25.23	30.33	86.05	18.41	50.15	34.76	90.00
WHUTE-CT-FV-02	38.08	41.92	43.41	47.34	40.09	18.73	37.88	12.44	52.09	39.83	75.26
III&CYUT-CT-FV-01	38.04	42.90	56.03	44.96	74.32	17.78	34.78	11.94	40.33	41.81	38.95
III&CYUT-CT-FV-03	37.72	42.41	54.58	44.80	69.82	14.87	29.41	9.95	43.70	42.71	44.74
III&CYUT-CT-FV-04	37.69	44.05	53.82	43.79	69.82	9.48	35.48	5.47	49.76	45.61	54.74
WHUTE-CT-FV-01	35.94	39.97	39.20	44.32	35.14	17.10	33.82	11.44	51.52	39.02	75.79
KUAS-CT-FV-01	33.97	36.38	48.99	41.43	59.91	25.59	26.92	24.38	27.33	37.27	21.58
MIG-CT-FV-02	31.07	35.89	48.78	38.17	67.57	15.27	28.38	10.45	29.17	33.56	25.79
MIG-CT-FV-03	30.88	36.70	49.46	37.77	71.62	26.93	31.76	23.38	16.24	43.18	10.00
MIG-CT-FV-01	29.07	34.91	49.53	37.86	71.62	18.75	25.21	14.93	18.94	33.78	13.16
MIG-CT-FV-04	28.73	35.89	49.61	37.83	72.07	27.17	29.94	24.88	9.39	43.48	5.26
MIG-CT-FV-05	26.02	36.87	52.77	37.72	87.84	13.53	27.69	8.96	11.76	41.94	6.84

Table 10: Results on FV subtask (CT).

Run ID	MacroF1	Accuracy	JA	JB	MS	PE	WA	WB
RITEVAL-NUL-JA-FV-03	61.93	63.23	0.474	0.375	0.250	0.400	0.357	0.435
RITEVAL-NUL-JA-FV-01	60.41	62.06	0.526	0.438	0.300	0.440	0.429	0.435
RITEVAL-NUL-JA-FV-05	60.16	62.26	0.579	0.438	0.300	0.440	0.393	0.435
RITEVAL-NUL-JA-FV-02	59.84	62.06	0.526	0.438	0.300	0.440	0.464	0.435
RITEVAL-NUL-JA-FV-04	59.46	61.28	0.474	0.438	0.350	0.480	0.357	0.348
RITEVAL-SKL-JA-FV-01	57.28	59.53	0.474	0.375	0.300	0.200	0.321	0.348
RITEVAL-KJP-JA-FV-05	57.00	57.59	0.579	0.375	0.250	0.280	0.286	0.174
RITEVAL-KitAi-JA-FV-02	56.37	57.59	0.211	0.188	0.200	0.240	0.179	0.174
RITEVAL-KJP-JA-FV-01	56.04	56.42	0.421	0.313	0.200	0.160	0.250	0.043
RITEVAL-CL-JA-FV-01	55.61	59.34	0.263	0.188	0.250	0.160	0.250	0.304
RITEVAL-SKL-JA-FV-04	55.50	56.42	0.368	0.188	0.250	0.360	0.286	0.304
RITEVAL-SKL-JA-FV-05	55.33	60.51	0.263	0.250	0.450	0.160	0.321	0.435
RITEVAL-CL-JA-FV-02	55.00	59.92	0.211	0.125	0.150	0.160	0.321	0.348
RITEVAL-SKL-JA-FV-02	54.87	59.92	0.316	0.313	0.300	0.360	0.286	0.522
RITEVAL-KJP-JA-FV-03	54.77	55.45	0.421	0.250	0.100	0.160	0.357	0.348
RITEVAL-KitAi-JA-FV-03	54.65	57.00	0.211	0.313	0.350	0.240	0.321	0.261
RITEVAL-KTU-JA-FV-02	54.31	59.73	0.158	0.250	0.200	0.160	0.286	0.435
RITEVAL-KTU-JA-FV-01	53.88	59.92	0.158	0.250	0.150	0.240	0.286	0.217
RITEVAL-SKL-JA-FV-03	53.84	56.03	0.316	0.250	0.250	0.440	0.286	0.261
RITEVAL-CL-JA-FV-03	53.58	59.73	0.211	0.063	0.150	0.160	0.357	0.348
RITEVAL-KSU-JA-FV-02	53.51	63.81	0.316	0.250	0.250	0.400	0.393	0.435
RITEVAL-NAK-JA-FV-01	53.16	55.36	0.263	0.125	0.250	0.280	0.250	0.217
RITEVAL-KSU-JA-FV-03	53.15	63.62	0.316	0.250	0.250	0.400	0.357	0.522
RITEVAL-KJP-JA-FV-04	52.92	52.92	0.263	0.188	0.100	0.280	0.286	0.261
RITEVAL-KJP-JA-FV-02	52.75	53.11	0.316	0.250	0.100	0.160	0.214	0.348
RITEVAL-NAK-JA-FV-02	51.81	61.21	0.368	0.250	0.200	0.200	0.214	0.348
RITEVAL-KSU-JA-FV-01	50.97	51.36	0.368	0.500	0.300	0.280	0.393	0.391
RITEVAL-KitAi-JA-FV-01	50.94	58.37	0.263	0.250	0.350	0.200	0.429	0.435
RITEVAL-KTU-JA-FV-03	49.57	59.14	0.211	0.188	0.150	0.160	0.321	0.261
RITEVAL-SITLP-JA-FV-01	44.89	60.89	0.316	0.375	0.250	0.120	0.143	0.391

Table 11: Results on FV subtask (JA).

Run ID	MacroF1	Accuracy	PE	WB
RITEVAL-BnO-EN-FV-01	53.17	55.85	0.160	0.304
RITEVAL-MIG-EN-FV-05	50.86	51.60	0.320	0.304
RITEVAL-MIG-EN-FV-04	49.60	51.06	0.320	0.261
RITEVAL-MIG-EN-FV-01	48.87	50.00	0.160	0.304
RITEVAL-MIG-EN-FV-02	48.87	50.00	0.160	0.304
RITEVAL-MIG-EN-FV-03	47.27	47.34	0.160	0.261
RITEVAL-WHUTE-EN-FV-02	46.03	53.72	0.200	0.261
RITEVAL-WHUTE-EN-FV-01	45.40	52.13	0.160	0.217
RITEVAL-BnO-EN-FV-02	45.29	52.66	0.160	0.174

Table 12: Results on FV subtask (EN).

Run	MacroF1	Acc.	Y-F1	Y-Prec.	Y-Rec.	N-F1	N-Prec.	N-Rec.
BUPTTeam-CS-SVBC-05	61.51	62.33	67.15	59.54	77.00	55.86	67.45	47.67
BUPTTeam-CS-SVBC-04	61.42	62.83	68.81	59.28	82.00	54.02	70.81	43.67
BUPTTeam-CS-SVBC-02	60.82	62.33	68.52	58.85	82.00	53.11	70.33	42.67
BUPTTeam-CS-SVBC-01	60.54	62.08	68.34	58.66	81.83	52.75	69.97	42.33
BUPTTeam-CS-SVBC-03	60.54	62.08	68.34	58.66	81.83	52.75	69.97	42.33
NWNU-CS-SVBC-05	59.71	59.75	60.95	59.18	62.83	58.47	60.39	56.67
NWNU-CS-SVBC-04	58.83	58.83	58.90	58.80	59.00	58.76	58.86	58.67
NWNU-CS-SVBC-03	58.03	59.00	64.40	56.91	74.17	51.67	62.92	43.83
III&CYUT-CS-SVBC-04	56.75	56.75	57.07	56.65	57.50	56.42	56.85	56.00
III&CYUT-CS-SVBC-03	56.03	56.50	60.57	55.39	66.83	51.49	58.19	46.17
WHUTE-CS-SVBC-01	53.48	54.58	60.65	53.50	70.00	46.31	56.63	39.17
III&CYUT-CS-SVBC-02	52.60	53.67	59.71	52.82	68.67	45.49	55.24	38.67
WHUTE-CS-SVBC-02	51.96	52.83	58.44	52.23	66.33	45.47	53.88	39.33
NWNU-CS-SVBC-02	51.83	55.00	64.19	53.30	80.67	39.46	60.27	29.33
Yamraj-CS-SVBC-01	49.24	49.25	48.69	49.23	48.17	49.79	49.27	50.33
NWNU-CS-SVBC-01	45.82	51.75	63.74	51.05	84.83	27.90	55.17	18.67
ASNLP-CS-SVBC-01	44.95	51.50	63.94	50.89	86.00	25.95	54.84	17.00
IMTKU-CS-SVBC-03	42.80	53.25	67.25	51.75	96.00	18.34	72.41	10.50
IMTKU-CS-SVBC-02	42.54	53.17	67.25	51.70	96.17	17.84	72.62	10.17
JAVN-CS-SVBC-01	42.32	51.17	64.91	50.65	90.33	19.73	55.38	12.00
IMTKU-CS-SVBC-01	41.77	52.75	67.05	51.47	96.17	16.49	70.89	9.33
WUST-CS-SVBC-01	39.14	52.25	67.39	51.17	98.67	10.89	81.40	5.83
III&CYUT-CS-SVBC-01	34.32	50.25	66.67	50.13	99.50	1.97	66.67	1.00

Table 13: Results on SVBC subtask (CS).

Run	MacroF1	Acc.	Y-F1	Y-Prec.	Y-Rec.	N-F1	N-Prec.	N-Rec.
III&CYUT-CT-SVBC-04	56.24	56.25	56.72	56.12	57.33	55.77	56.39	55.17
III&CYUT-CT-SVBC-03	56.00	56.50	60.69	55.36	67.17	51.31	58.26	45.83
KUAS-CT-SVBC-01	54.10	55.67	62.59	54.14	74.17	45.60	58.99	37.17
WHUTE-CT-SVBC-01	53.48	54.58	60.65	53.50	70.00	46.31	56.63	39.17
III&CYUT-CT-SVBC-02	51.99	53.17	59.51	52.41	68.83	44.47	54.61	37.50
WHUTE-CT-SVBC-02	51.96	52.83	58.44	52.23	66.33	45.47	53.88	39.33
ASNLP-CT-SVBC-01	44.74	51.42	63.95	50.84	86.17	25.54	54.64	16.67
IMTKU-CT-SVBC-01	44.03	50.63	63.19	50.35	84.83	24.87	52.13	16.33
NTOUA-CT-SVBC-03	42.89	52.33	66.11	51.29	93.00	19.66	62.50	11.67
IMTKU-CT-SVBC-03	42.71	44.25	52.11	45.67	60.67	33.30	41.44	27.83
JAVN-CT-SVBC-01	42.21	51.00	64.75	50.56	90.00	19.67	54.55	12.00
IMTKU-CT-SVBC-02	42.18	52.75	66.90	51.48	95.50	17.47	68.97	10.00
NTOUA-CT-SVBC-04	41.01	49.83	63.82	49.91	88.50	18.21	49.26	11.17
NTOUA-CT-SVBC-05	39.73	51.67	66.55	50.88	96.17	12.91	65.15	7.17
NTOUA-CT-SVBC-01	39.26	50.67	65.58	50.36	94.00	12.94	55.00	7.33
NTOUA-CT-SVBC-02	39.26	50.67	65.58	50.36	94.00	12.94	55.00	7.33
III&CYUT-CT-SVBC-01	34.46	50.25	66.63	50.13	99.33	2.29	63.64	1.17

Table 14: Results on SVBC subtask (CT).

Run	MF1	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
WUST-CS-SVMC-01	44.39	51.83	59.91	45.77	86.67	58.44	50.86	68.67	59.20	68.72	52.00	0.00	0.00	0.00
WUST-CS-SVMC-02	44.22	51.67	59.91	45.77	86.67	58.18	50.12	69.33	58.80	70.05	50.67	0.00	0.00	0.00
WUST-CS-SVMC-03	43.84	51.50	60.85	45.22	93.00	57.54	49.52	68.67	56.96	79.64	44.33	0.00	0.00	0.00
III&CYUT-CS-SVMC-04	40.41	42.33	46.99	40.48	56.00	54.66	52.80	56.67	41.21	39.57	43.00	18.76	29.93	13.67
III&CYUT-CS-SVMC-05	40.32	43.08	46.85	41.78	53.33	52.22	49.70	55.00	47.09	41.75	54.00	15.11	30.93	10.00
III&CYUT-CS-SVMC-01	37.64	41.08	40.41	41.55	39.33	49.82	53.01	47.00	47.79	36.19	70.33	12.53	34.33	7.67
ASNLP-CS-SVMC-01	33.74	40.33	46.92	35.03	71.00	58.12	50.75	68.00	24.84	33.71	19.67	5.08	53.33	2.67
III&CYUT-CS-SVMC-03	32.95	42.17	53.26	37.37	92.67	56.75	52.56	61.67	0.65	14.29	0.33	21.16	43.30	14.00
III&CYUT-CS-SVMC-02	31.06	41.67	52.93	37.43	90.33	57.83	50.12	68.33	0.65	12.50	0.33	12.81	38.98	7.67
WHUTE-CS-SVMC-01	25.74	36.83	46.57	42.16	52.00	50.41	34.72	92.00	5.99	29.41	3.33	0.00	0.00	0.00
WHUTE-CS-SVMC-02	24.30	35.08	41.80	39.02	45.00	49.46	33.82	92.00	5.93	27.03	3.33	0.00	0.00	0.00
Yamraj-CS-SVMC-01	23.71	25.08	25.78	23.20	29.00	32.18	26.21	41.67	23.32	28.64	19.67	13.57	21.13	10.00
NWNU-CS-SVMC-03	23.19	29.69	32.19	40.61	26.67	40.13	26.86	79.33	3.56	16.22	2.00	16.89	40.51	10.67
NWNU-CS-SVMC-01	22.57	29.36	35.33	31.43	40.33	40.12	28.39	68.33	3.56	16.22	2.00	11.27	36.36	6.67
NWNU-CS-SVMC-02	21.83	28.86	33.28	34.66	32.00	39.37	26.69	75.00	3.56	16.22	2.00	11.11	45.24	6.33
IMTKU-CS-SVMC-03	19.54	27.92	53.77	38.98	86.67	7.11	10.67	5.33	17.28	15.40	19.67	0.00	0.00	0.00
IMTKU-CS-SVMC-01	19.02	29.17	53.07	37.02	93.67	5.39	14.08	3.33	17.61	15.95	19.67	0.00	0.00	0.00
IMTKU-CS-SVMC-02	18.67	29.08	52.76	36.67	94.00	4.56	15.69	2.67	17.35	15.53	19.67	0.00	0.00	0.00

Table 15: Results on SVMC subtask (CS).

Run	MF1	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.
III&CYUT-CT-SVMC-05	40.54	43.33	47.21	42.15	53.67	52.06	49.70	54.67	47.76	42.20	55.00	15.11	30.93	10.00
III&CYUT-CT-SVMC-04	40.52	42.42	47.50	40.71	57.00	54.28	52.66	56.00	41.16	39.75	42.67	19.13	30.22	14.00
KUAS-CT-SVMC-01	39.08	42.58	50.50	43.95	59.33	49.09	42.21	58.67	43.00	42.04	44.00	13.74	39.06	8.33
III&CYUT-CT-SVMC-01	35.43	38.67	48.34	37.50	68.00	50.00	52.99	47.33	30.42	29.56	31.33	12.97	34.29	8.00
ASNLP-CT-SVMC-01	33.79	40.50	47.42	35.35	72.00	58.04	50.62	68.00	24.63	33.92	19.33	5.08	53.33	2.67
III&CYUT-CT-SVMC-03	32.95	42.17	53.26	37.37	92.67	56.75	52.56	61.67	0.65	14.29	0.33	21.16	43.30	14.00
III&CYUT-CT-SVMC-02	31.27	41.83	52.98	37.48	90.33	58.11	50.37	68.67	0.65	12.50	0.33	13.33	40.00	8.00
NTOUA-CT-SVMC-03	31.03	39.17	49.45	35.22	83.00	52.86	47.24	60.00	16.79	35.48	11.00	5.02	42.11	2.67
NTOUA-CT-SVMC-04	29.31	37.17	49.35	35.18	82.67	50.08	45.96	55.00	10.47	24.39	6.67	7.34	24.07	4.33
NTOUA-CT-SVMC-05	29.03	38.58	48.45	34.57	81.00	53.90	45.71	65.67	9.50	43.24	5.33	4.26	24.14	2.33
NTOUA-CT-SVMC-01	28.89	38.33	48.68	35.59	77.00	53.44	43.74	68.67	10.47	40.91	6.00	2.98	13.89	1.67
NTOUA-CT-SVMC-02	28.83	38.25	48.47	35.50	76.33	53.42	43.58	69.00	10.47	40.91	6.00	2.98	13.89	1.67
WHUTE-CT-SVMC-01	25.74	36.83	46.57	42.16	52.00	50.41	34.72	92.00	5.99	29.41	3.33	0.00	0.00	0.00
WHUTE-CT-SVMC-02	24.30	35.08	41.80	39.02	45.00	49.46	33.82	92.00	5.93	27.03	3.33	0.00	0.00	0.00
IMTKU-CT-SVMC-03	19.63	28.08	53.81	38.96	87.00	7.19	11.03	5.33	17.52	15.58	20.00	0.00	0.00	0.00
IMTKU-CT-SVMC-01	19.01	29.11	53.13	37.14	93.33	5.39	14.08	3.33	17.51	15.78	19.67	0.00	0.00	0.00
IMTKU-CT-SVMC-02	18.48	28.94	52.62	36.59	93.67	4.01	14.29	2.33	17.30	15.45	19.67	0.00	0.00	0.00

Table 16: Results on SVMC subtask (CT).

Run	MacroF1	Acc.	Y-F1	Y-Prec.	Y-Rec.	N-F1	N-Prec.	N-Rec.
NUL-JA-SV-04	69.59	77.81	53.78	55.11	52.51	85.40	84.75	86.06
NUL-JA-SV-05	68.94	77.96	52.20	55.89	48.97	85.67	84.01	87.40
NUL-JA-SV-01	68.73	77.81	51.89	55.56	48.67	85.58	83.92	87.31
NUL-JA-SV-02	68.33	75.34	53.42	49.87	57.52	83.23	85.43	81.15
NUL-JA-SV-03	67.87	76.65	51.06	52.66	49.56	84.67	83.87	85.48
KSU-JA-SV-01	66.96	79.84	46.33	67.04	35.40	87.59	81.75	94.33
KSU-JA-SV-03	65.72	75.78	47.15	50.85	43.95	84.29	82.50	86.15
CL-JA-SV-02	65.27	71.86	50.13	44.42	57.52	80.40	84.68	76.54
CL-JA-SV-03	65.27	71.86	50.13	44.42	57.52	80.40	84.68	76.54
KSU-JA-SV-02	64.87	76.00	45.11	51.52	40.12	84.64	81.79	87.69
SKL-JA-SV-01	64.37	74.33	45.54	47.59	43.66	83.21	82.12	84.33
CL-JA-SV-01	64.17	76.50	43.16	53.25	36.28	85.19	81.18	89.62
KTU-JA-SV-01	63.67	71.65	46.66	43.40	50.44	80.69	82.94	78.56
NAK-JA-SV-02	63.19	74.55	42.74	47.81	38.64	83.64	81.18	86.25
JAVN-JA-SV-03	63.05	74.18	42.77	47.00	39.23	83.33	81.20	85.58
NAK-JA-SV-01	62.02	73.89	40.79	46.10	36.58	83.26	80.63	86.06
KitAi-JA-SV-01	62.02	68.02	46.93	39.63	57.52	77.11	83.77	71.44
KitAi-JA-SV-02	59.93	65.41	45.11	36.98	57.82	74.75	83.16	67.88
JAVN-JA-SV-02	55.88	77.66	24.88	71.83	15.04	86.88	77.98	98.08
NAK-JA-SV-03	54.14	72.23	25.34	37.36	19.17	82.94	77.26	89.52
JAVN-JA-SV-01	52.69	59.32	34.99	28.82	44.54	70.40	78.01	64.13
SITLP-JA-SV-01	51.78	72.23	20.37	34.51	14.45	83.18	76.56	91.06
JAVN-JA-SV-04	49.68	76.94	12.64	92.00	6.78	86.72	76.66	99.81
JAVN-JA-SV-05	47.44	54.97	27.54	22.78	34.81	67.33	74.33	61.54
Yamraj-JA-SVBC-01	44.92	47.43	33.18	24.13	53.10	56.66	74.88	45.58
KitAi-JA-SV-03	32.12	33.14	40.44	25.89	92.33	23.80	84.71	13.85

Table 17: Results on SV subtask (JA).