A Machine Learning based Textual Entailment Recognition System of JAIST Team for NTCIR9 RITE

Quang Nhat Minh Pham
Japan Advanced Institute of Science And Technology
1-1 Asahidai, Nomi, Ishikawa, 923-1292, JAPAN
minhpqn@jaist.ac.jp

Le Minh Nguyen
Japan Advanced Institute of Science And Technology
1-1 Asahidai, Nomi, Ishikawa, 923-1292, JAPAN
nguyenml@jaist.ac.jp

Akira Shimazu
Japan Advanced Institute of Science And Technology
1-1 Asahidai, Nomi, Ishikawa, 923-1292, JAPAN
shimazu@jaist.ac.jp

ABSTRACT

NTCIR9-RITE is the first shared-task of recognizing textual inference in text written in Japanese, Simplified Chinese, or Traditional Chinese. JAIST team participates in three subtasks for Japanese: Binary-class, Entrance exam and RITE4QA. We adopt a machine learning approach for these subtasks, combining various kinds of entailment features by using machine learning techniques. In our system, we use a Machine Translation engine to automatically produce English translation of the Japanese data, and both original Japanese data and its translation are used to train an entailment classifier. Experimental results show the effectiveness of our method. Although our system is lightweight and does not require deep semantic analysis or extensive linguistic engineering, it obtained the first rank (accuracy of 58%) among participant groups on the Binary-class subtask for Japanese.

Categories and Subject Descriptors

I.2.7 [**Natural Language Processing**]: Text Analysis, Language parsing and understanding

General Terms

Theory, Languages

Keywords

Textual Entailment, Machine Learning, Machine Translation

1. INTRODUCTION

Recognizing Textual Entailment (RTE) is a fundamental task in Natural Language Understanding. It has been proposed as an applied semantic framework to model language variability [4]. Given two text portions T (text) and H (hypothesis), the task is to determine whether the meaning of H can be inferred from the meaning of T.

RTE can potentially be applied in many NLP tasks, such as Question Answering or Text Summarization. Applications of RTE have been reported in several tasks: Question Answering [8], Information Extraction [17]. In these studies, RTE has been integrated as an important component. For instance, in Question Answering [8], a RTE component was used to determine if a candidate answer is the right answer for a question or not.

RTE task has been received much attention in NLP research community, recently. There have been several RTE shared tasks hold by TAC conference [1], and many dedicated RTE workshops.

This year, NTCIR9 Workshop holds the RITE (Recognizing Textual Inference in TExt) shared-task which is the first attempt of constructing a common benchmark for evaluating systems which automatically detect entailment, paraphrase, and contradiction of texts written in Japanese, Simplified Chinese, or Traditional Chinese [18]. There are four subtasks offered by the shared-task organizers: Binary-class (BC), Multi-class (MC), Entrance Exam, and RITE4QA subtask.

JAIST team participates in three subtasks for Japanese: BC subtask, Entrance Exam, and RITE4QA subtask. This paper describes our RTE system used in the shared-task.

Our RTE system is based on machine learning. The RTE task is formulated as a binary classification problem and machine learning methods are applied to combine entailment features extracted from each pair of text T and hypothesis H. The advantage of machine learning-based approaches to RTE is that multiple entailment features can be easily combined to learn an entailment classifier. Entailment features in our system are mainly based on distance and similarity measures applied on two text portions.

In our RTE system, for each Japanese pair T/H, we use a Machine Translation (MT) engine to produce its English translation, and both the original pair and its translation are used to determine whether the entailment relationship exists in the pair. Our method is based on a reasonable assumption that if T entails H then the translation T' should entail the translation H'. We expect that this bilingual constraint can be used to improve the performance of the RTE system.

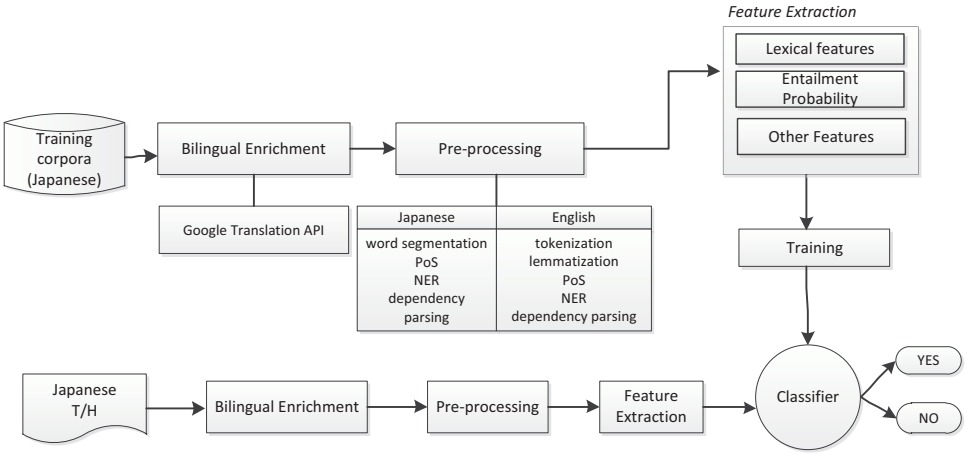

Figure 1: System Architecture of Japanese RTE

The remainder of our paper is organized as follows. Section 2 presents some related work to our research. Section 3 describes our machine-learning-based system. In Section 4, we present experimental results for BC subtask, Entrance Exam subtask and RITEQA subtask. Finally, Section 6 gives conclusions and some remarks.

2. RELATED WORK

Mehadad et al. [14] proposed the cross-lingual textual entailment (CLTE) task in which text T and hypothesis H are written in different languages. A basic solution for CLTE task was proposed, in which a Machine Translation (MT) system is added to the front-end of an existing RTE engine. For instance, for a pair of English text and Spanish hypothesis, the hypothesis will be translated into English and then, the RTE engine will be run on the pair of the text and the translation of the hypothesis.

Mehadad et al. [15] proposed a new approach to CLTE task, which take advantages of bilingual corpora by extracting information from the phrase-table to enrich inference and entailment rules, and using extracted rules for a distance based entailment system. Effects of bilingual corpora to monolingual TE was also analysed. The main idea of that work is to increase the coverage of monolingual paraphrase tables by extracting paraphrases from bilingual parallel corpora and use extracted paraphrases for monolingual RTE. This approach requires monolingual paraphrase tables of the two languages.

Our approach makes use of Machine Translation for monolingual RTE. In our machine-learning-based RTE system, we combine both features extracted from data in original language and from translation data produced by a MT component to learn an entailment classifier.

3. SYSTEM DESCRIPTION

In our paper, we adopt the machine learning based approach to building RTE system. A RTE problem is formulated as a binary classification problem in which each instance consists of a pair of the text T and the hypothesis H.

In this section, we describe our RTE system. Our RTE system is divided into four main modules as shown in Figure 1: bilingual enrichment, preprocessing, feature extraction, and training.

For each Japanese pair T/H, first, it is automatically translated into English using a MT engine. Then in preprocessing, both the Japanese pair and its associated translation pair are analysed. After that, features extracted from the pair and its English translation pair are input to a classifier to determine the label for the pair.

Our system used Support Vector Machines (SVMs) [21, 3], a robust method for classification problems, to train the Entailment classifier which can determine whether the text T entails the hypothesis H for each pair T/H. We tried several machine learning methods, such as Maximum Entropy Model [2], yet SVM obtained the best performance.

3.1 Bilingual Enrichment

In order to make use of English translation data for RTE, original RTE corpus in Japanese is automatically translated into English, using Google Translator Toolkit[1].

3.2 Preprocessing

3.2.1 Japanese Pairs

We used Cabocha tool [20] for data preprocessing. For each pair, preprocessing process consists of tokenizing, chunking, named-entity recognition, and dependency parsing. Parsed content of each sentence is represented in XML format.

3.2.2 English Pairs

Each Japanese T/H pair in our corpus is associated with its English translation. We use Stanford-CoreNLP tool to perform preprocessing for English pairs[2]. Stanford-CoreNLP provides a set of fundamental natural language processing

[1]Google Translator Toolkit: `http://translate.google.com/toolkit`

[2]Stanford CoreNLP is available on: `http://nlp.stanford.edu/software/corenlp.shtml`

tools which can take raw English text input. At lexical level, we use the tool for tokenization, lemmatization, part-of-speech tagging, named-entity recognition. At syntactic level, dependency parsing is performed.

3.3 Entailment Classifier

Our system trains an entailment classifier which can decide whether the meaning of a hypothesis H can be inferred from a text T. Each pair T/H is represented by a feature vector $\langle f_1, ..., f_m \rangle$ which contains multiple similarity measures of the pair and some other features. For each training instance consisting of a pair T/H, features are extracted from both the original pair in Japanese and its associated English translation pair. In the rest of this section, we describe features used in the entailment classifier.

3.3.1 Similarity Features

A large part of lexical features used in the entailment classifier are similar to features used in [13]. We used different kinds of text similarity/distance measures applied on the pair and its English translation. These measures capture how H is covered by T.

For each pair T/H (Japanese pair or English translation pair), text similarity/distance measures are applied on two pairs of strings:

- **Pair 1**: Two strings which consist of words of T and H in surface forms. Punctuations and special characters are removed. Stop words are removed for English pairs.
- **Pair 2**: Two strings which consist of base forms of words in T and H, respectively. Punctuations and special character are removed. Stop words are removed for English pairs.

We give a brief description of lexical features used in Entailment classifier as follows.

a) Word overlap

Word-overlap feature captures lexical-based semantic overlap between T and H, which is a score based on matching each word in H with some words in T [5]. When computing lexical matching, Japanese WordNet and English WordNet [6] are used. Matching criterion for two English words are the same as in [5]. For Japanese, a word h_w in H is considered as a match with a word t_w in T if they have the same surface or base form, or h_w is hypernym, meronym, or entailed word or of t_w.

b) Levenshtein distance

Levenshtein distance [11] of two strings is the minimum number of edit operations needed to transform a string to the other. Allowable edit operations are deletion, insertion, or substitution of single character. In our system, edit distances from T to H are computed.

c) BLEU measures

BLEU score is a popular evaluation metric used in automatic machine translation [16]. It measures how a translation generated by a MT system is close to reference translations. The main idea is to compute n-gram matching between automatically generated translations and references translations. In RTE problem, we used BLEU precision of H and T (T is cast as a reference translation) based on uni-gram, 2-gram, and 3-gram. Both baseline BLEU precision and modified n-gram precision are used.

d) Longest Common Subsequence String (LCS)

LCS feature computes the length of the longest common subsequence string between T and H [9]. The LCS feature is normalized by dividing by the length of H.

e) Other similarity/distance measures

We compute various similarity/distance measures which have been used for RTE: Jaccard coefficient, Mahatan distance, Euclidean distance, Jaro-Winkler distance, Cosine Similarity, and Dice Coefficient. For details of these measures, see [13].

3.3.2 Entailment Probability

The entailment probability that T entails H is computed based on the probabilistic entailment model in [7]. The main idea is that the probability that the entailment relationship exists in the pair, $P(H|T)$ is computed via the probability that each individual word in H is entailed by T. The probability $P(H|T)$ is computed by the following equation:

$$P(H|T) = \prod_j P(h_j|T) \quad (1)$$

where the probability $P(h_j|T)$ is defined as the probability that the word h_j in H is entailed by T.

$$P(h_j|T) = \max_i P(h_j|t_i) \quad (2)$$

In Equation 2, $P(h_j|t_i)$ can be interpreted as the lexical entailment score between words t_i and h_j. By this decomposition, the overall probability $P(H|T)$ is computed by the following equation.

$$P(H|T) = \prod_j \max_i P(h_j|t_i) \quad (3)$$

The lexical entailment score of two words w_1 and w_2 is computed by using the word similarity score between them. For English, lexical entailment scores are computed based on Levenshtein distance as in [12]

$$sim(w_1, w_2) = 1 - \frac{dist(w_1, w_2)}{max\,(length(w_1), length(w_2))} \quad (4)$$

For Japanese pairs, we use the Japanese thesaurus, Nihongo goitaikei [10] to compute the similarity of two words.

Table 1: Data statistics

Dataset	Y	N	Total
BC Subtask - Dev set	250	250	500
BC Subtask - Test set	250	250	500
Exam Subtask - Dev set	204	295	499
Exam Subtask - Test set	181	261	442
RITE4QA Test set	106	858	964

3.3.3 Dependency-parse-based Features

Dependency relation overlap has been used in paraphrase identification [22]. For RTE task, we use dependency relation precision of H and T which is computed using the following equation:

$$precision_d = \frac{|relations(H) \cap relations(T)|}{|relations(H)|} \quad (5)$$

where $relations(s)$ denotes the set of head-modifier relations for the sentence s.

3.3.4 Named-Entity mismatch

In a pair T/H, if the hypothesis contains a named-entity which does not occur in the text, the text may not entail the hypothesis. We use an indicator function π to compute the named-entity mismatch feature of T and H: $\pi(T, H) = 1$ if H contains a named-entity that does not occur in T and $\pi(T, H) = 0$, otherwise. We compute named-entity mismatch for both Japanese pairs and their associated English translation pairs.

3.3.5 Polarity Mismatch

The polarity mismatch in a pair T/H may indicate that T does not entail H. We compute polarity mismatch in a pair T/H using the Polarity Weighted Word List [19]. In that list, each Japanese word is associated with a weight that indicates whether the word has positive meaning or negative meaning. We use an indicator function to capture if words in the root nodes of dependency parses of T and H have opposite polarity.

4. EXPERIMENTS AND RESULTS

4.1 Data set

Thanks to NTCIR workshop organizers for providing benchmark data to evaluate RTE systems.

For BC subtask and Exam subtask, we trained entailment classification models on the development portion and evaluated on the test portion provided for each subtask. Development set was not provided for RITE4QA subtask, so we used the development set of the Exam subtask to train entailment classifiers. Table 1 provides some statistical information of the data sets. While label distribution of BC subtask's data sets is balanced, in Exam subtask and RITE4QA subtask, the number of "N" pairs is much greater than the number of "Y" pairs.

4.2 Submitted runs

JAIST team submitted three runs for BC subtask (Japanese) as follows:

Table 2: BC Subtask Results

Methods	Accuracy
SVM_bi	0.580 (290/500)
SVM_mono	0.566 (283/500)
MEM_mono	0.552 (276/500)

Table 3: Exam Subtask Results

Methods	Accuracy
LLM	0.622 (275/442)
SVM_bi	0.652 (288/442)
SVM_mono	0.652 (288/442)

Table 4: RITE4QA Subtask Results

Methods	Acc	Top1	MMR5
LLM	0.560	0.180	0.276
SVM_bi	0.676	0.151	0.260
SVM_mono	0.694	0.166	0.273

- **Run 1** (SVM_bi) used libSVM [3] as the machine-learning tool and all features extracted from original Japanese pairs and their associated English translation pairs. We tuned parameters for learning on the development set by using parameter selection tool in the libSVM package.
- **Run 2** (SVM_mono) used libSVM as the machine-learning tool and monolingual features extracted from original Japanese pairs. We compare the result obtained in Run 2 with the result of Run 1 to see whether bilingual constraints can improve performance of the system.
- **Run 3** (MEM_mono) used Maximum Entropy Model as the machine-learning tool and monolingual features extracted from original Japanese pairs.

For Exam subtask and RITE4QA subtask, we submitted the results obtained by SVM_mono, SVM_bi, and a baseline method based on lexical matching (LLM method).

4.3 Experimental Results

Official results achieved on test sets of BC subtask, Exam subtask, and RITE4QA subtask are shown on Table 2, Table 3 and Table 4, respectively. Classification accuracy was used as the evaluation measure for all three subtasks. For the RITE4QA subtask, to evaluate the impact of the RTE engine on the QA system, Top1 and MMR5 were used.

As can be seen in the Table 2, the SVM_bi method achieved the best accuracy. The performance of SVM_mono is slightly above MEM_mono. However, the improvement achieved with SVM_bi is not statistically significant (we used McNemar Test with $p < 0.05$).

In Exam subtask, the accuracy of SVM_mono is as good as SVM_bi; and in RITE4QA subtask, SVM_mono achieved better accuracy than SVM_bi. In experiments, initial parameters used in SVM training (the cost C and the gamma

Table 5: Error Statistics

Method (Subtask)	False-positive	False-negative
SVM_bi (BC)	107	103
SVM_mono (BC)	101	116
SVM_bi (Exam)	57	97
SVM_mono (Exam)	57	97
SVM_bi (RITE4QA)	243	69
SVM_mono (RITE4QA)	238	57

in the RBF kernel function) affect the performance of entailment classifiers in testing. When we use default parameters provided by the tool libSVM, SVM_bi achieved better accuracy than SVM_mono on the test set of Exam subtask[3]. Unexpectedly, in Exam subtask, and RITE4QA the experimental results did not show any superior improvement when using Machine Translation. However, we argue that we can improve overall performance of RTE system if the quality of the MT component is improved.

4.4 Result Analysis

Table 5 compares the number of false-positive pairs and false-negative pairs predicted by SVM_mono and SVM_bi on the test set of each subtask. False-positive pairs are pairs which are predicted as "Y" pairs by a system while in gold standard, they are "N" pairs. False-negative pairs are pairs which are predicted as "N" pairs by a system while in gold standard, they are "Y" pairs

Analysing false-positive pairs predicted by SVM_bi, we see that false-positive pairs mainly come from "N" pairs in which H is highly covered by T in terms of lexical. A possible explanation for this might be that features used to train classifiers are mainly based on text similarity/distance measures.

In many "N" pairs, the label may be decided by a "cue-difference" between H and T. For instance, in the pair 9 in Figure 2, the "cue-difference" is in two phrases "Ig Nobel Prize" and "Nobel Prize". We argue that in order to detect these false-entailment pairs, we need to develop an alignment component to align corresponding constituents between T and H and design an algorithm for weighting importance of "differences" in the pair based on the alignment.

Among true-entailment pairs which our systems do not correctly detect, many pairs use complicated entailment and paraphrasing rules, such as pair 1 and pair 148 as shown in Figure 2. For instance, in pair 148, we need a rule "housewives and seeking-job people do not have workplaces". Therefore, a large paraphrase table of paraphrasing phrases and a database of entailment rules are needed for the task.

As can be seen in Table 2, the number of false-positive pairs predicted by SVM_bi is greater than the one predicted by SVM_mono. It may indicate that the MT component used in SVM_bi provides more evidences for detecting entailment relationship in "Y" pairs which have high word overlap. Pair 28 in Figure 2 is an example pair which correctly predicted by SVM_bi but incorrectly predicted by SVM_mono.

[3]SVM_mono and SVM_bi achieved accuracies of 65.6% and 69.4%, respectively

Table 7: Ablation Tests

Ablated Resource	BC	Exam
JWordNet	0%	0%
Goi Taikei	0.2%	0.2%
Polarity Words	-0.2%	0.7%
JWordNet + Goi Taikei	0.2%	0.2%
JWordNet + Polarity Words	-0.2%	0.5%
Goi Taikei + Polarity Words	-0.2%	-0.4%
JWordNet + Goi Taikei + Polarity Words	0%	0%

4.5 Feature Analysis

We conduct feature analyses in order to understand impact of features on the performance of machine-learning-based RTE systems.

We divide features set into three categories as follows.

- **LemmaSim** consists of similarity features computed on base/lemma form of each pair T/H.
- **SurSim** consists of similarity features applied on surface form of each pair T/H.
- **SynSem** consists of other features: entailment probability, dependency-parse based features, named-entity mismatch and polarity features.

Entailment classifiers are trained using above features subsets and combination of them on the development sets. In order to avoid affects of selecting initial parameters in SVM training on performance of RTE systems, we used default parameters of libSVM package. Table 6 shows accuracies of various settings on the test sets of BC and Exam subtasks.

Feature analyses indicated that similarity features significantly contribute to the performance of RTE systems. As shown in Table 6, without using similarity features, the accuracies of SVM_bi and SVM_mono decrease much. Similarity features applied on base form representation of each pair T/H and its English translation (in the group LemmaSim) are important in exam subtask while the contribution of features in SynSem group are not so significant in both two subtasks.

4.6 Ablation Tests

In RTE task, it is interesting to know how additional resources or components contribute to the performance of our Japanese RTE system. This section presents ablation tests for two subtasks. We only analyse the effects of RTE resources and components to SVM_mono method to avoid affects of unpredictable errors propagated from the Machine Translation component.

Table 7 provides performance differences between of the SVM_mono using complete additional resources and the system without using some resources. The percentages shown in Table 7 indicate the contribution of resources to the performance of the system. As indicated in the table, the impact of additional resources on the performance of our system is not so significant. A possible explanation for this may

# ID	Text	Hypothesis	Dataset	Label
1	石垣島は、冬でもハイビスカスが咲き乱れる楽園だ。 Ishigaki Island is a paradise of bloomed hibiscus even in winter.	石垣島の冬の気温は高い。 Temperature of winter in Ishigaki Island is high.	BC-test	Y
9	「イグ・ノーベル賞」（愚かなノーベル賞）の化学賞 に、広瀬幸雄氏（６２）が選ばれた。 Chemistry “Ig Nobel” prize was awarded to Yukio Hirose (62 years old)	ノーベル賞に広瀬幸雄氏が選ばれた。 Nobel” prize was awarded to Yukio Hirose.	BC-test	N
148	主婦や求職中の人も２割いる。 20% of people are housewives and people who are seeking jobs.	２割が、「職場を持たない人」だ。 20% of people are people who do not have workplace.	BC-test	
28	宝塚歌劇団はチャリティーコンサートを開催した。 Takarazuka Revue Company held a charity concert.	宝塚歌劇団は慈善活動を行った。 Takarazuka Revue Company conducted charity activities.	BC-test	Y

Figure 2: Some examples in BC-test set

Table 6: Feature Analysis

Setting	BC	Exam
SVM_mono + LemmaSim	56.2% (-0.4)	65.1% (-0.5)
SVM_mono + SurSim	56.6% (+0)	64.5% (-1.1)
SVM_mono + SynSem	53.4% (-3.2)	64.0% (-1.6)
SVM_mono + LemmaSim + SurSim	56.8% (+0.2)	64.5% (-1.1)
SVM_mono + LemmaSim + SynSem	56.2% (-0.4)	65.6% (+0)
SVM_mono + SurSim + SynSem	56.0% (-0.6)	66.1% (+0.5)
SVM_mono + All Features	**56.6%**	**65.6%**
SVM_bi + LemmaSim	57.2% (+0.4)	68.1% (-1.3)
SVM_bi + SurSim	57.0% (+0.2)	65.8% (-3.6)
SVM_bi + SynSem	53.4% (-3.4)	65.6% (-3.8)
SVM_bi + LemmaSim + SurSim	58.2% (+1.4)	68.3% (-1.1)
SVM_bi + LemmaSim + SynSem	55.8% (-1.0)	69.2% (-0.2)
SVM_bi + SurSim + SynSem	56.2% (-0.6)	69.9% (+0.5)
SVM_bi + All Features	**56.8%**	**69.4%**

be that resources were used only in computing a small subset of features in our system. Specifically, Japanese WordNet was used to compute word-overlapping features, Nihongo goi taikei was used to compute entailment probability feature, and Polarity Word List was used to compute polarity mismatch in a pair.

5. DISCUSSION

This section discusses entailment phenomena in the RTE corpus. We have observed the data and tried to classify linguistic phenomena of textual entailment. We distinguish true-entailment pairs and false-entailment pairs. Table 3 shows some example T/H pairs in BC-subtask's development set.

5.1 True-Entailment Pairs

5.1.1 Type 1: World Knowledge based Inference

To determine label for a pair in this type, world knowledge is indispensable. In the pair, we cannot make a decision based on only textual evidences conveyed in the text and the hypothesis. For instance, in the pair 26 shown in Figure 3, we cannot determine whether the text entails the hypothesis if we do not know that the person called Oyama Nobuyo is a woman.

5.1.2 Type 2: Inference based on paraphrasing and entailment words/phrases

In pairs of this type, the decision can be made based on paraphrasing phrases or entailment words. For instance, in the pair 25 (Figure 3), it uses paraphrasing phrases pair “captured the heart of the public” and “attracted the public.”

5.1.3 Type 3: Hypotheses are facts extracted from texts

In a pair of this type, information conveyed in the hypothesis is a fact which can be extracted from the text. An example is the pair 496 as shown in Figure 3.

5.2 False-Entailment Pairs

5.2.1 Type 1: Negation structure

In a pair of this type, the hypothesis may use negation structures, and the meaning of the hypothesis contrasts with

# ID	Text	Hypothesis	Label
17	省エネは、生活レベルを落として原始時代のような生活をしなければならないという思い込みがあるが、そうとも限らないことに気づく必要がある。 There is a belief that in order to conserve energy, we must lower the level of life to primitive ages, but we need to aware that it is not necessary.	省エネは、生活レベルを落として原始時代のような生活をしなければならない。 In order to conserve energy, we must lower the level of life to primitive ages.	N
25	歌舞伎は大衆の心をとらえてきた。 Kabuki has captured the heart of the public.	歌舞伎は大衆を魅了してきた。 Kabuki has attracted the public.	Y
26	大山のぶ代は『太陽にほえろ！』の脚本家だった。 Oyama Nobuyo is the writer of “Bark at the Sun.”	『太陽にほえろ！』の脚本家は女性である。 The writer of “Bark at the Sun” is a woman.	Y
188	ベラルーシとポーランドは国境を接し合う隣国同士である。 Belarus and Poland have the same national border.	ポーランドとベラルーシは近隣ではない。 Poland and Belarus are not neighbors.	N
198	８割弱の大学でインターンシップが導入されている。 Internship has been introduced in nearly 80% of universities.	希望するすべての学生はインターンシップを体験できる。 All students who want can experience an internship.	N
206	人間の脳は生まれつき、言葉を理解する機能を備えている。 The human brain naturally has the ability to understand language.	人間は動物の中で唯一言語を獲得した。 Human is the only animal can communication by using language	N
357	日本人の平均所得はイギリスのそれよりかなり上である。 Japanese average income is considerably higher than English people.	日本人はイギリス人より幸福だ。 Japanese people are happier than English people	N
496	６月１０日の「時の記念日」を控え、時計メーカーが、電波を使って正確な時刻に修正する電波時計の新製品を相次いで投入する。 Before the Time Day’s June 10th, the watchmaker introduces a series of new products of radio clocks which fix the time exactly using radio waves.	６月１０日は時の記念日だ。 June 10th is the Time Day.	Y

Figure 3: Example pairs in BC subtask’s development set

meaning of the text. An example is the pair 188 as shown in Figure 3.

5.2.2 *Type 2: Hypothesis discusses an aspect of a topic, which is not mentioned in text*

In the pair 206 (Figure 3), the text said that human being can understand language. However, the hypothesis said that human being is the only animal that can acquire language, which is not mentioned in the text.

5.2.3 *Type 3: Ambiguity*

In the pair 17, the hypothesis is completely covered by the text, but the remaining part of the text inverses the label of the pair.

5.2.4 *Type 4: Wrong inference rules*

In pairs of this type, there are inference rules that are not necessarily true. For instance, in the pair 357 (figure 3), the text said that average income in Japan is higher than in England, but it is not necessarily true that Japanese people are happier than English people.

Textual entailment phenomenon discussed above indicated that the RTE task has very complicated nature, and extensive encoded world knowledge in the machine-readable form is indispensable for the RTE task.

6. CONCLUSION

We have presented our system which participated in the Binary-class, Entrance exam, and RITE4QA subtasks. Our system is based on machine learning, and multiple entailment features extracted from both original Japanese pairs and their English translation are combined to learn the Entailment classifier. Experimental results has shown some interesting points. First, although our system does not require deep semantic analysis and extensive linguistic engineering, it obtained the best accuracy (58%) in the Binary-class subtask for Japanese on the test set among participant groups. Second, our study has indicated that Machine Translation may be used to improve performance of the RITE system.

Our study still has several major limitations. First, the system is not very precise at detecting *hard* false-entailment pairs in which H is highly covered by T. Second, due to the lack of a entailment rule database and a large paraphrasing tables, our system fail to detect entailment relationship in pairs that use complex inference rules. We plan to address these problems by developing an alignment component and acquiring entailment/paraphrasing rules from large text corpus.

7. REFERENCES

[1] L. Bentivogli, I. Dagan, H. T. Dang, D. Giampiccolo, and B. Magnini. The fifth pascal recognizing textual entailment challenge. In *In Proceedings of TAC Workshop*, 2009.

[2] A. L. Berger, V. J. D. Pietra, and S. A. D. Pietra. A maximum entropy approach to natural language processing. *Comput. Linguist.*, 22:39–71, 1996.

[3] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[4] I. Dagan, O. Glickman, and B. Magnini. The pascal recognising textual entailment challenge. *MLCW, LNAI*, 3944:177–190, 2006.

[5] I. Dagan, D. Roth, and F. Massimo. A tutorial on textual entailment, 2007.

[6] C. Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

[7] O. Glickman, I. Dagan, and M. Koppel. Web based probabilistic textual entailment. In *In Proceedings of the 1st RTE Workshop*, Southampton, UK, 2005.

[8] S. Harabagiu and A. Hickl. Methods for using textual entailment in open-domain question answering. In *In Proceedings of ACL*, pages 905–912, 2006.

[9] D. S. Hirschberg. Algorithms for the longest common subsequence problem. *J. ACM*, 24:664–675, 1977.

[10] S. Ikehara, M. Miyazaki, S. Sirai, A. Yokoo, H. Nakaiwa, K. Ogura, Y. Ooyama, and Y. Hayashi. *Nihon-go goi taikei*. Iwanami, Japan (in Japanese), 1997.

[11] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966.

[12] B. MacCartney. *Natural Language Inference*. PhD thesis, Stanford University, 2009.

[13] P. Malakasiotis and I. Androutsopoulos. Learning textual entailment using svms and string similarity measures. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 42–47, 2007.

[14] 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*, pages 321–324, June 2010.

[15] 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*, pages 1336–1345, June 2011.

[16] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, 2002.

[17] L. Romano, M. Kouylekov, I. Szpektor, I. Dagan, and A. Lavelli. Investigating a generic paraphrase-based approach for relation extraction. In *In Proceedings of EACL*, pages 401–408, 2006.

[18] H. Shima, H. Kanayama, C.-W. Lee, C.-J. Lin, T. Mitamura, Y. Miyao, S. Shi, and K. Takeda. Overview of ntcir-9 rite: Recognizing inference in text. In *In NTCIR-9 Proceedings, to appear*, 2011.

[19] H. Takamura, T. Inui, and M. Okumura. Extracting semantic orientations of words using spin model. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 133–140, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[20] Y. M. Taku Kudo. Japanese dependency analysis using cascaded chunking. In *CoNLL 2002: Proceedings of the 6th Conference on Natural Language Learning 2002 (COLING 2002 Post-Conference Workshops)*, pages 63–69, 2002.

[21] V. N. Vapnik. *Statistical learning theory*. John Wiley, 1998.

[22] S. Wan, M. Dras, R. Dale, and C. Paris. Using dependency-based features to take the "para-farce" out of paraphrase. In *In Proceedings of ALTW*, 2006.