Textual Entailment Recognition using Word Overlap, Mutual Information and Subpath Set

Yuki Muramatsu
Nagaoka University of Technology
muramatsu@jnlp.org

Kunihiro Udaka
Nagaoka University of Technology
udaka@jnlp.org

Kazuhide Yamamoto
Nagaoka University of Technology
yamamoto@jnlp.org

Abstract

When two texts have an inclusion relation, the relationship between them is called entailment. The task of mechanically distinguishing such a relation is called recognising textual entailment (RTE), which is basically a kind of semantic analysis. A variety of methods have been proposed for RTE. However, when the previous methods were combined, the performances were not clear. So, we utilized each method as a feature of machine learning, in order to combine methods. We have dealt with the binary classification problem of two texts exhibiting inclusion, and proposed a method that uses machine learning to judge whether the two texts present the same content. We have built a program capable to perform entailment judgment on the basis of word overlap, i.e. the matching rate of the words in the two texts, mutual information, and similarity of the respective syntax trees (Subpath Set). *Word overlap* was calclated by utilizing BiLingual Evaluation Understudy (BLEU). *Mutual information* is based on co-occurrence frequency, and the *Subpath Set* was determined by using the Japanise WordNet. A Confidence-Weighted Score of 68.6% was obtained in the mutual information experiment on RTE. Mutual information and the use of three methods of SVM were shown to be effective.

1 Introduction

This paper can help solve textual entailment problems. Researchers of natural language processing have recently become interested in the automatic recognition of textual entailment (RTE), which is the task of mechanically distinguishing an inclusion relation. Text implication recognition is the task of taking a text (T) and a hypothesis (H), and judging whether one (the text) can be inferred from the other (hypothesis). Here below is an example task. In case of entailment, we call the relation to be 'true').

Example 1: Textual entailment recognition.
T: Google files for its long-awaited IPO.
H: Google goes public.
Entailment Judgment: True.

For such a task, large applications such as question answering, information extraction, summarization and machine translation are involved. A large-scale evaluation workshop has been conducted to stimulate research on recognition of entailment (Dagan et al., 2005). These authors divided the RTE methods into six methods. We focused on 3 methods of them.

Pérez and Alfonseca's method (Pérez and Alfonseca, 2005) used *Word Overlap*. This method is assumed to have taken place when words or sentences of the text and the hypothesis are similar, hence the relation should be true. Pérez and Alfonseca used the BLEU algorithm to calculate the entailment relationship. Glickman et al's. method was considered as using statistical lexical relations. These authors assumed that the possibility of entailment were high when the co-occurrence frequency of the word in the source and the target were high.

Proceedings of the 2nd Workshop on Cognitive Aspects of the Lexicon (CogALex 2010), pages 18–27,
Beijing, August 2010

While this may be correct, we believe nevertheless that it problematic not to consider the co-occurrence of the hypothesis words. This being so, we proposed to use *mutual information*. Finally, Herrera et al's. method is based on Syntactic matching. They calculated the degree of similarity of the syntax tree.
We combined these three methods using machine learning techniques.

2 Related Works

Dagan et al. (Dagan et al, 2005) conducted research in 2005 on how to evaluate data of RTE; the authors insisted on the need of semantic analysis. As a first step, they considered the problem of textual entailment, proposing how to build evaluation data. Theu also organised and workshop on this topic. Their evaluation data are problems of binary classification of the texts to be compared. They used a sentence extracted from a newspaper corpus, and built a hypothesis from this text using one of seven methods: question answering, sentence comprehension, information extraction, machine translation, paraphrasing, information retrieval and comparable documents. They proposed a method of evaluation using RTE, and they introduced several RTE methods.

Odani et al. (Odani et al, 2005) did research on the construction of evaluation data in Japan, mentionning that there was a problem in the evaluation data of Dagan et al. For example, they stated that 'The evaluation data that he constructed are acting some factors. So it is difficult to discuss the problem'. Next, they did an RTE evaluation data using Japanese. The inference factors for judging entailment judgment were divided into five categories: inclusion, lexicon (words that can't be declined), lexicon (declinable words), syntax and inference. The subclassification was set for each classification, and Japanese RTE evaluation data was constructed. In addition, a dictionary and Web text were used for the entailment judgment. The authors were able to solve entailment judgment with words or phrases containing synonyms and/or a super-sub type relation. However, this classification lacks precision.

For example, they defined the term 'lexicon (words that cannot be declined)' as 'The meaning and the character of the noun that exists in text are data from which information on the truth of hypothesis is given'. Given this lack of clarity, we considered this method to be difficult to reproduce.

However, the evaluation data they built is general and available for public use. Regarding the research using the evaluation data of such RTE, there have been many reports in the workshop.

For example, Pérez and Alfonseca (Pérez and Alfonseca, 2005) assumed that the possibility of entailment was high when the text matched the hypothesis. The concordance rate of the text and the hypothesis was then calculated for judging the text and the hypothesis of the inclusion relation. In their research, they used BiLingual Evaluation Understudy (BLEU) to evaluate machine translation. An entailment judgment of 'true' was given when the BLEU score was above than a given threshold decided in advance. The evaluation data of Dagan et al. was used in the experiment, and its accuracy was about 50%. The evaluation data of comparable document types were the results with the highest accuracy. Hence the authors concluded that this method can be considered as a baseline of RTE. We dealt with it as *word overlap*.

Glickman et al. (Glickman et al, 2005) conducted research using co-occurring words. They assumed that the entailment judgment was 'true' when the probability of co-occurrence between the text and the hypothesis was high. In addition, the content word of the text with the highest co-occurrence probability was calculated from the content word of all of the hypotheses, and it was proposed as a method for entailment judgment. A Web search engine was used to calculate in the co-occurrence probability. This experiment yielded an accuracy of approximately 58%, while the evaluation data of comparable document types was about 83%. This being so, the authors concluded that they have been able to improve the results with the help of other deep analytical tools. We improved this method, and used it as mutual information.

Herrera et al. (Herrera et al., 2005) focused on syntactic similarity. They assumed that the entailment judgment was 'true' when the syntactic similarity of the text and the hypothesis was high. In addition, they used WordNet for considering identifiable expressions. The results

of the experiment yielded an accuracy of approximately 57%. We improved this method, and used it then as subpath set.

Prodromos Malakasiotis and Ion Androutsopoulos (Prodromos Malakasiotis and Ion Androutsopoulos, 2007) used Support Vector Machines. They assumed that the entailment judgment was 'true' when the similarity of words, POS tags and chunk tags were high. The results of the experiment yielded an accuracy of approximately 62%. However, they forgot to combine past RTE methods as feature of SVM.

The authors of this paper present a new RTE method. We propose to combine word overlap, mutual information and subpath sets. We dealt with SVM by using 3 methods equally as features, and we estimated higher precision than when using individual; independent methods.

3 Textual Entailment Evaluation Data

We used the textual entailment evaluation data of Odani et al. for the problem of RTE. This evaluation data is generally available to the public at the Kyoto University[1].

The evaluation data comprises the inference factor, subclassification, entailment judgment, text and hypothesis. Table 1 gives an example. The inference factor is divided into five categories according to the definition provided by Odani et al.: inclusion, lexicon (indeclinable word), lexicon (declinable word), syntax and inference. They define the classification viewpoint of each inference factor as follows:

Example 2: Classification criteria of inference factors

· Inclusion: The text almost includes the hypothesis.

· Lexicon (Indeclinable Word): Information of the hypothesis is given by the meaning or the behaviour of the noun in the text.

· Lexicon (Declinable Word): Information of the hypothesis is given by the meaning or the behaviour of the declinable word in the text.

· Syntax: The text and the hypothesis have a relation of syntactic change.

· Inference: Logical form.

They divided the data into 166 subclasses, according to each inference factor. The entailment judgment is a reliable answer in the text and the hypothesis. It is a difficult problem to entailment judgment for the criteria answer. Therefore, when they reported on the RTE workshop, they assumed the following classification criteria:

Example 3: Classification criteria of entailment determination.

· ◎(T_{alw}): When the text is true, the hypothesis is always true.

· ○(T_{alm}): When the text is true, the hypothesis is almost true.

· △(F_{may}): When the text is true, the hypothesis may be true.

· ×(F_{alw}): When the text is true, the hypothesis is false.

In terms of the text and the hypothesis, when we observed the evaluation data, the evaluation data accounted for almost every sentence in both the texts and the hypotheses, and also the hypotheses were shorter than the texts.

There is a bias in the number of problems evaluated by the inference factor and by the subclassification. The number of evaluation data open to the public now stands at 2471.

Inference Factor	Sub-Classification	Entailment Judgment	Text	Hypothesis
Lexicon (Indeclinable Word)	Behavior	◎	Toyota opened a luxury car shop.	Lexus is a luxury car.

Table 1: RTE Evaluation data of Odani et al.

[1] http://www.nlp.kuee.kyoto-u.ac.jp/nl-resource

4 Proposal Method

Up now, a number of methods have been proposed for RTE. However, when the previous methods were combined, the performances were hard to judge. Hence, we used each method as a feature of machine learning, and combined them then.

The input text and the hypothesis were considered as a problem of binary classification ('true' or 'false'). Therefore, we employed support vector machines (Vapnik, 1998), which are often used to address binary classification problems (in fact, we implemented our system with Tiny SVM). With this method we achieved higher precision than with individual independent methods.

Figure 1 shows our proposed method.

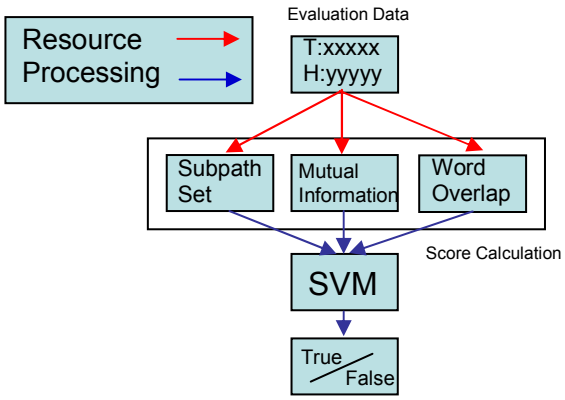

Figure 1: Our Proposed Method

In the following sections, we will describe the three features used in machine learning.

4.1 Word Overlap

It is assumed that when words or sentences of the text and the hypothesis are similar, the relation should be true. Pérez and Alfonseca used a BLEU algorithm to calculate the entailment between the text and the hypothesis. BLEU is often used to evaluate the quality of machine translation. Panieni et al. provided the following definition of BLEU. In particular, the BLEU score between length r of the sentence B and length c of the sentence A is given by the formulas (1) and (2):

$$Bleu(A,B) = BP \exp(\sum_{i=1}^{n} \log(p_i)/n) \quad (1)$$

$$BP = exp\left(1 - max\{1, r/c\}\right) \quad (2)$$

where p_i represents the matching rate of n-gram. The n-gram of this method was calculated as word n-gram. We assumed n = 1 and used the public domain program NTCIR7[2]. Here is an example of the calculation.

Example 4: Calculation by BLEU.
T:月は地球の衛星である。(The moon is Earth's satellite.)
H:月は地球の周りにある。(The moon is around the Earth.)
BLEU:0.75

We estimated n = 1 for the following reasons:

1. The reliability of word overlap is not high when n is large.
2. The calculated result of BLEU often becomes 0 when n is large.

First, we will explain the reason 1 mentioned above. The report of Kasahara et al. (Kasahara et al., 2010) is a reproduction of the one provided by Pérez et al (Pérez et al., 2005). They prepared an original RTE evaluation set of reading comprehension type, and proposed a new RTE system using a BLEU algorithm. When they experimented by increasing the maximum number of elements n of word n-gram from 1 to 4, the optimum maximum number of elements n is 3. They proposed the following analysis: if the hypothesis is shorter than the text, with n = 4, then the frequency is low in word 4-gram. However, the accidental coincidence of the word 4-gram significantly affected BLEU. When n is large, the reliability of the word overlap decreases.

Next, as an explanation of reason 2, when the length of the targeted sentence is short, the numerical result of BLEU sometimes becomes 0. For example, the number of agreements of 4-gram becomes 0 when calculating with n = 4, and the BLEU value sometimes becomes 0.

[2] http://www.nlp.mibel.cs.tsukuba.ac.jp/bleu_kit/

Such calculations accounted for approximately 69% of the Odani et al. evaluation set.

4.2 Mutual Information

Glickman et al. (Glickman et al. 2005) assumed that the possibility of entailment is high when the co-occurrence frequency of the word in the text and the hypothesis is high. Therefore, they proposed a method of total multiplication, by searching for the word with the highest co-occurrence frequency from all the words of the hypothesis, as shown in formulas (3) and (4):

$$P(Trh=1|t)=\prod_{u\in h}\max_{V\in t} lep(u,v) \quad (3)$$

$$lep(u,v)\approx\frac{n_{u,v}}{n_v} \quad (4)$$

$P(Trh=1|t)$ expresses the probability of entailment between the text and the hypothesis. In these formulas, u is the content word of the hypothesis (noun, verb, adjective or unknown word); v the content word of the text; n represents the number of Web search hits; $n_{u,v}$ is the number of hits when the words u and v are searched on the Web. But, when the content word of the text is low frequency, the numerical result of the *lep(u, v)* increases for $P(Trh=1|t)$. We believe that it was a problem not to take into account the co-occurrence of the hypothesis words. In addition, their method to handle long sentences and reaching the conclusion 'false' is problematic. This is why, we considered Rodney et al's. method (Rodney et al. 2006) and proposed the use of mutual information, which is calculates on the basis of the formulas (5) and (6):

$$P(Trh=1|t)=\frac{1}{\mathbf{u}}\prod_{u\in h}\max_{V\in t} lep(u,v) \quad (5)$$

$$lep(u,v)\approx-\log\frac{p(n_{u,v})}{p(n_u)\cdot p(n_v)} \quad (6)$$

u is the number of the content words of the hypothesis. Hence, 1/**u** averages product of max *lep(u,v)*. This being so we considered that this model can do entailment judgments independantly of the length of the hypothesis.

It searches for the word of the text considering that the mutual information reaches the maximum value from each of the hypothesis words. When $P(Trh=1|t)$ is higher than an arbitrary threshold value, it is judged to be 'true', and 'false' in the opposite case. Glickman assumed the co-occurrence frequency to be the number of Web-search hits. However, we estimated that the reliability of the co-occurrence frequency was low, because the co-occurrence of the Web search engine was a wide window. This is why, we used the Japanese Web N-gram[3]. In particular, we used 7-gram data, and calculated the co-occurrence frequency $n_{u,v}$, frequency n_u and n_v of the word. $p(n_i)$ was calculated by (?) the frequency n_i divided the number of all words. Japanese Web N-gram was made from 20,036,793,177 sentences, including 255,198,240,937 words. The unique number of 7-gram is 570,204,252.

To perform morphological analysis, we used Mecab[4], for example:

Example 5: Calculation by mutual information.
T:この部屋はクーラーが効いている。(The air conditioner works in this room.)
H:涼しい。(It is cool.)
Mutual Information:10.0

$$P(Trh=1|t)=\frac{1}{1}\prod_{u\in}\max_{V\in t} lep(u,v) \quad (7)$$

$$lep(u,v)=-\log\frac{p\left(n_{\text{涼しい,クーラー(cool, the air conditioner)}}\right)}{p\left(n_{\text{涼しい(cool)}}\right)\cdot p\left(n_{\text{クーラー(the air conditioner)}}\right)}\approx 10.0 \quad (8)$$

This method actually standardises the result by dividing by the maximum value of *lep(u, v)*. As a result, p reaches the value 1 from 0. We used the discounting for n_u, n_v,and $n_{u,v}$, because a zero-frequency problem had occurred when calculating the frequency. There are some methods for discounting. We used the additive method reported by Church and Gale (Church and Gale, 1991). They compared some discounting methods by using the newspaper corpus. The addition method is shown as follows.

$$P(w)=\frac{C(w)+1}{N+V} \quad (9)$$

[3] http://www.gsk.or.jp/catalog/GSK-2007-C/
[4] http://mecab.sourceforge.net/

The additive method assumed N to be the number of all words in a corpus. *C(w)* is the frequency of word *w* in the corpus. V is a constant to adjust the total of the appearance probability to 1. It is equal to the unique number of words w. The additive method is very simple, it adds a constant value to occurrence count *C(w)*. The method of adding 1 to the occurrence count is called Laplace method also.

4.3 Subpath Set

Herrera et al. (Herrera et al., 2005) parsed the hypothesis and the text, and they calculated the degree of similarity of the syntax tree from both. Our method also deals with the degree of similarity of the syntax tree. The tree kernel method of Collins and Duffy （M. Collins and N. Duffy, 2002）shows the degree of similarity of the syntax tree; however, it requires much time to calculate the degree of similarity. Therefore, we employed the subpath set of Ichikawa et al. This latter calculates partial routes from the root to the leaf of the syntax tree. Our method assumes the node to be a content word (noun, verb, adjective or unknown word) in the syntax tree, while the branch is a dependency relation. For parsing we relied on Cabocha[5] .

The frequency vector was assumed to comprise a number of partial routes, similar to the approach of Ichikawa et al. (Ichikawa et al., 2005). The number of partial routes is unique. However, even if the same expression is shown for the word with a different surface form, it is not possible to recognise it as the same node. Therefore, we used the Japanese version of WordNet (Bond et al., 2009), in which a word with a different surface can be treated as the same expression, because Japanese WordNet contains synonyms. The same expressions of our method were hypernym words, hyponym words and synonym words in Japanese Word Net, because RTE sometimes considered the hierarchical dictionary of the hypernym and the hyponym word to be the same expression. However, our hypernym and hyponym words were assumed to be a parent and a child node of the object word, as shown in Figure 3.

[5] http://chasen.org/~taku/software/cabocha/

Example 6: Calculation by subpath set.

T:キャンペーン中なので、ポイントが 2 倍付く。

(T:The point adheres by the twice because it is campaigning.)

H:キャンペーン中なので、ポイントが普段より 2 倍付く。

(H:The point adheres usually by the twice because it is campaigning.)

Subpath:0.86

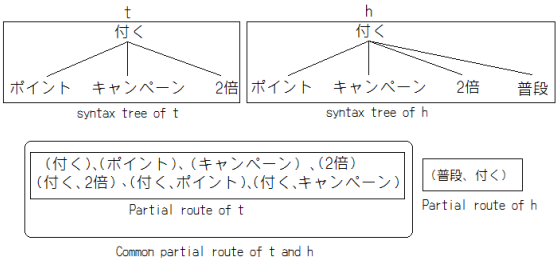

Figure 2: Partial route chart of subpath set.

The number of partial routes is 7, and 6 partial routes overlap in T and H. So, the subpath is 0.86 (6/7).

5 Evaluation

The textual entailment evaluation data of Odani et al., described in Section 3, was used in the experiment. The entailment judgment of four values is manually given to the textual entailment evaluation data. In our experiment we considered 'T_{alw}' and 'T_{alm}' to be 'true' and 'F_{may}' and 'F_{alw}' as 'false'. The evaluation method used was a Confidence-Weight Score (CWS, also known as Average Precision), proposed by Dagan et al.. As for the closed test, the threshold value with the maximum CWS was used.

$$Accuracy = Correct / All \quad (10)$$

$$CWS = \frac{1}{k} \sum_{1 \le i \le All} r_k \cdot precision(k) \quad (11)$$

$$precision(k) = \frac{1}{k} \sum_{1 \le i \le k} r_i \quad (12)$$

All = Number of all evaluation data. Correct = Number of correct answer data. If *k* is a correct answer, *rk* = 1. If *k* is an incorrect answer, *rk* = 0.

When the Entailment judgment annotated in evaluation data matches with the Entailment judgment of our method, the answer is true.

The threshold of the Closed test was set beforehand ($0 \leqq th \leqq 1$). When it was above the threshold, it was judged "true". When it was higher than the threshold, it was judged "false". SVM was used to calculate the value of three methods (word overlap, mutual information and subpath set) as the features for learning data, was experimented.

Open test was experimented 10-fold cross-validations. 9 of the data divided into 10 were utilized as the learning data. Remaining 1 was used as an evaluation data. It looked for the threshold that CWS becomes the maximum from among the learning data. It experimented on the threshold for which it searched by the learning data to the evaluation data. It repeats until all data that divides this becomes an evaluation data, averaged out. (Or we experimented Leave-one-out cross validation.)

Using the SVM, experiments were conducted on the numerical results of Sections 4.1 to 4.3 as the features.

The textual entailment evaluation data numbered 2472: 'T_{alw}': 924, 'T_{alm}': 662, 'F_{may}': 262 and 'F_{alw}': 624, and there were 4356 words. The total number of words was 43421. Tables 2 and 3 show the results of the experiment, which focused respectively on the closed and open tests,. When the 'true' textual entailment evaluation data 'T_{alw}' only and 'T_{alw} and T_{alm}' was used, mutual information achieved the best performance. When the true data 'T_{alm}' only was used, SVM achieved the best performance.

6 Discussion

In this section, we discuss the relation between each 3 method value assumed to be the criterion of judgment and CWS in the closed test. When the 'true' evaluation data was assumed to be 'T_{alm}' only in the open test, the result of SVM exceeded the results of the closed test. We then consider the relation between SVM and CWS.

6.1 Close Test of Ward Overlap

We believe that the results of the experiments of word overlap were more effective than other methods, because they achieved the best performance excluding 'T_{alm}' and 'T_{alw} and T_{alm}' in 3 methods. Figure 3 shows the relation to CWS when BLEU value changes.

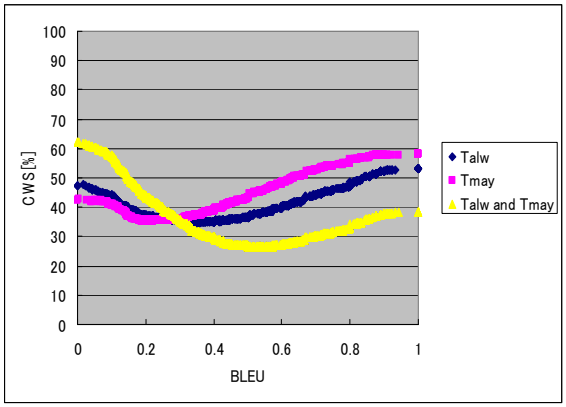

Figure 3: Results of the closed test of the RTE experiments by word overlap.

The tendency shown in Figure 3 did not change much when the relation between the threshold value and CWS was observed, even though the 'true' evaluation data was changed.

	CWS					
	Closed Test			Open Test		
	T_{alw}	T_{alm}	T_{alw} and T_{alm}	T_{alw}	T_{alm}	T_{alw} and T_{alm}
Word Overlap	53.0%	57.9%	62.1%	39.0%	60.2%	59.3%
Mutual Informaition	**55.9%**	52.9%	**68.6%**	**53.4%**	55.6%	**67.4%**
Subpath Set	54.5%	57.0%	61.8%	45.0%	59.7%	61.1%
SVM	51.4%	**61.2%**	63.5%	49.9%	**61.9%**	64.1%

Table 2: Results of the RTE experiments

However, the entailment judgment of the word overlap method becomes nearly 'false' when the BLEU value is 1 (or 'true' when BLEU score is 0.) Table 3 shows the entailment judgment when the BLEU value is 0 or 1.

We assumed that BLEU value that CWS becomes the maximum depends on the ratio of number of T and F in the evaluation set. However, when true condition is "T_{alw}" only, T is more than F (T:924,F:886). And when true condition is "T_{alm}" only, F is more than T (T:662,F:886). For this reason, The possibility of our assumption is low because both true conditions are BLEU value that CWS becomes the maximum is 1.

6.2 Close Test of Mutual Information

We believe that the results of the experiments of mutual information were more effective than other methods, because they achieved the best performance excluding 'T_{alm}' in 3 methods. Figure 4 shows the relation to CWS when mutual information value changes.

The tendency shown in Figure 4 did not change much when the relation between mutual information value and CWS was observed, even though the 'true' evaluation data was changed. When mutual information values are from 0.2 (or 0.3) to 1, CWS increased. However, the entailment judgment of the mutual information method becomes almost 'true' when mutual information score is near 1 (or 'false' when mutual information score value is near 0.)

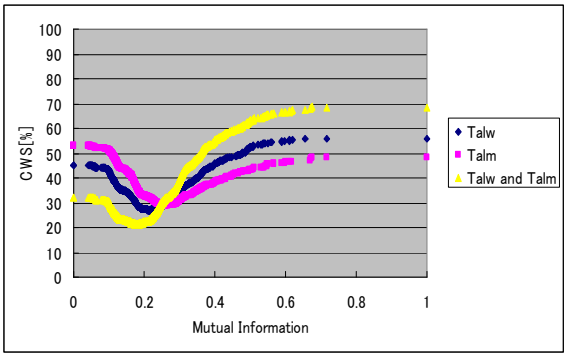

Figure 4: Results of the closed test of the RTE experiments by mutual information.

Table 4 shows the entailment judgment when the mutual information value is near 0 or 1. Our results showed most entailment judgment results to be almost 'true' (or almost 'false') for the optimal threshold value in the evaluation data. Therefore, we considered that the method of RTE using mutual information should be reviewed.

6.3 Close Test of Subpath Set

We believe that the results of the experiments of subpath set were not better than other methods. Figure 5 shows the relation to CWS when subpath set (SS) value changes.

The tendency shown in Figure 5 changed much when the relation between the threshold value and CWS was observed, even though the 'true' evaluation data was changed. When the true conditions are "T_{alw}" and "T_{alm}", the tendencies were very near.

	Answer/System	T/T	T/F	F/T	F/F	CWS
True Condition	T_{alw} (Bleu=1)	5	919	12	874	53.0
	T_{alm} (Bleu=1)	0	662	12	874	57.9
	T_{alw} and T_{alm} (Bleu=0)	1586	0	886	0	62.1

Table 3: Entailment judgment in closed test of word overlap (T=True, F=False).

	Answer/System	T/T	T/F	F/T	F/F	CWS
True Condition	T_{alw} (MI=0.72)	924	0	884	2	55.9
	T_{alm} (MI=0)	0	2	662	884	52.9
	T_{alw} and T_{alm} (MI=0.68)	1586	0	884	2	68.6

Table 4: Entailment judgment in closed test of mutual information (T=True, F=False, MI=mutual information).

However, when the true conditions were "T_{alw}" and "T_{alw} and T_{alm}", the tendencies were different. The tendency of "T_{alw}" was rising. The tendency of "T_{alw} and T_{alm}" was dropping until the subpath set value was 0.2. The entailment judgment of the mutual information method becomes almost 'true' when subpath set value was near 1)

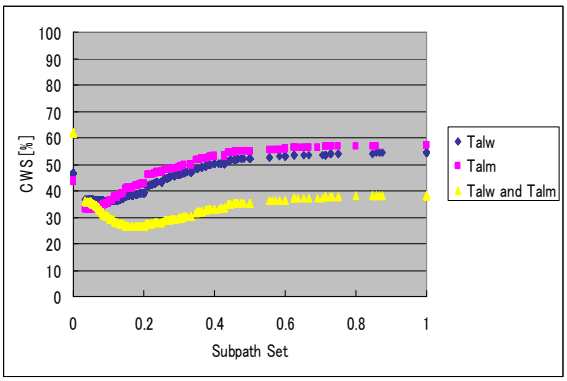

Figure 5: Results of the closed test of the RTE experiments by subpath set.

Table 5 shows the entailment judgment when the threshold value is near 0 or 1. Our results showed most entailment judgment results to be almost 'true' (or almost 'false') for the optimal subpath set value in the evaluation data.

6.4 Open Test of SVM

The open tests were conducted in 10-fold cross-validation , and the experimental result is their average. Figure 6 shows the related chart 10-fold cross-validation.

When the true data were assumed to be 'T_{alm}' only, the maximum value of CWS was 70.3%. As a result, the result of 10–fold cross validation exceeded the closed test.

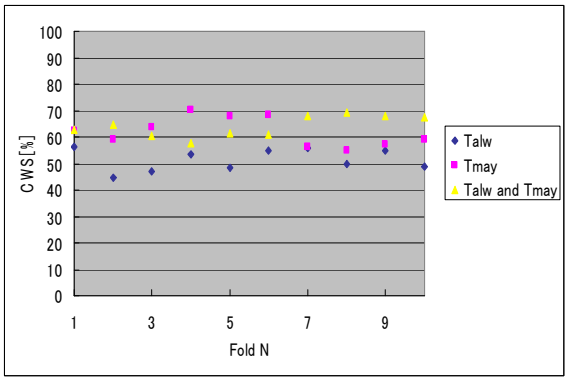

Figure 6: Results of the open test of the RTE experiments by SVM.

When the true data was assumed to be 'T_{alw}' only, the minimum value of CWS was 42.7%. We focused on the difference between the maximum and minimum value in 10-fold cross-validation. When the true answer was assumed to be 'T_{alm}', the difference between the maximum and minimum value is the greatest (15.3 points) in the open tests, and 'T_{alw} and T_{alm}' was the lowest with 11.6 points.

We believe that when the result 'T_{alm}' was 'true', it was consequently more unstable than 'T_{alw} and T_{alm}', because there was a larger amount of evaluation data 'T_{alw} and T_{alm}'.

7 Conclusion

We built a Japanese textual entailment recognition system based on the past methods of RTE. We considered the problem of RTE as a problem of binary classification, and built a new model of RTE for machine learning. We proposed machine learning to consider the matching rate of the words of the text and the hypothesis, using mutual information and similarity of the syntax tree. The method of using mutual information and the use of three methods of SVM tunrned out to be effective.

	Answer/System	T/T	T/F	F/T	F/F	CWS
True Condition	T_{alw} (SS=1)	9	915	14	872	54.5
	T_{alm} (SS=1)	1	661	14	872	57.0
	T_{alw} and T_{alm} (SS=0)	1586	0	886	0	61.8

Table 5: Entailment judgment in closed test of subpath set (T=True, F=False, SS=subpath set).

In the future, we will consider changing the domain of the evaluation data and the experiment. Moreover, we will propose a new method for the feature of machine learning.

We will also consider to expand WordNet. Shnarch et al. (Shnarch et al., 2009) researched the extraction from Wikipedia of lexical reference rules, identifying references to term meaning triggered by other terms. They evaluated their lexical reference relation for RTE. They improved previous RTE methods. We will use their method for ours in order to expand Japanese WordNet. We believe that this can help us improve our method/results.

References

Michitaka Odani, Tomohide Shibata, Sadao Kurohashi, Takayuki Nakata, Building data of japanese Text Entailment and recognition of inferencing relation based on automatic achieved similar expression. *In Proceeding of 14th Annual Meeting of the Association for Natural Language Processing,* pp. 1140-1143, 2008 (in Japanese)

Diana Pérez and Enrique Alfonseca. Application of the Bleu algorithm for recognising textual entailment. *In Proceedings of the first PASCAL Recognizing Textual Entailment Challenge*, pp. 9-12, 2005

Oren Glickman, Ido Dagan and Moshe Koppel. Web Based Probabilistic Textual Entailment. *In Proceedings of the PASCAL Recognizing Textual Entailment Challenge*, pp. 33-36, 2005

Francis Bond, Hitoshi Isahara, Sanae Fujita, Kiyotaka Uchimoto, Takayuki Kuribayashi and Kyoko Kanzaki. Enhancing the Japanese WordNet. *In the 7th Workshop on Asian Language Resources*, in conjunction with ACL-IJCNLP, pp. 1-8, 2009

Hiroshi Ichikawa, Taiichi Hashimoto, Takenobu Tokunaka and Hodumi Tanaka. New methods to retrieve sentences based on syntactic similarity. *Information Processing Society of Japan SIGNL Note*, pp39-46, 2005(in Japanese)

Kishore Panieni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. *In Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 311-318, 2002

Ido Dagan, Oren Glickman and Bernardo Magnini. The PASCAL Recognizing Textual Entailment Challenge. *In Proceedings of the first PASCAL Recognizing Textual Entailment Challenge*, pp. 1-8,2005

Jesús Herrera, Anselmo Peñas and Felisa Verdejo, Textual Entailment Recognition Based on Dependency Analysis and WordNet. *In Proceedings of the first PASCAL Recognizing Textual Entailment Challenge*, pp. 21-24, 2005

Kaname Kasahara, Hirotoshi Taira and Masaaki Nagata, Consider of the possibility Textual Entailment applied to Reading Comprehension Task consisted of multi documents. *In Proceeding of 14th Annual Meeting of the Association for Natural Language Processing,* pp. 780-783, 2010 (in Japanese)

M. Collins and N. Duffy. Convolution kernel for natural language. *In Advances in Neural Information Proccessing Systems (NIPS)*, volume 16, pages 625–632, 2002.

Prodromos Malakasiotis and Ion Androutsopoulos. Learning Textual Entailment using SVMs and String Similarity Measures. *In Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pp. 42-47, 2007

Vladimir N. Vapnik, The Statisitcal Learning Theory. Springer, 1998.

Church, K. W. & Gale, W. A.. A comparison of the enhanced Good-Turing and deleted estimation methods for estimating probabilities of English bigrams. Computer Speech and Language, volume 5, 19-54.

Rodney D. Nielsen, Wayne Ward and James H. Martin. Toward Dependency Path based Entailment. *In Proceedings of the second PASCAL Recognizing Textual Entailment Challenge*, pp. 44-49, 2006

Eyal Shnarch, Libby barak, Ido Dagan. Extracting Lexical Reference Rules from Wikipedia. *In Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 450-458, 2009