Experiments for NTCIR-9 RITE Task at Shibaura Institute of Technology

Toru Sugimoto
Department of Information Science and Engineering, Shibaura Institute of Technology
3-7-5 Toyosu, Koto-ku, Tokyo 135-8548, Japan
sugimoto@shibaura-it.ac.jp

ABSTRACT

This paper reports the evaluation results of our textual entailment system at NTCIR-9 RITE task. We participated in the Japanese Binary-Class (BC) subtask. In our system, the meaning of a text is represented as a set of dependency triples consisting of two words and their relation. Comparing two sets of dependency triples with respect to conceptual and character-based similarity, a subsumption score is calculated and used to identify textual entailment. This paper provides a description of our algorithm, the evaluation results and discussion on the results.

Keywords

textual entailment, dependency triple, conceptual similarity

1. INTRODUCTION

The meaning of a sentence can typically be considered as a set of *facts* expressed in it. In this view, textual entailment between two texts can be seen as a subsumption relation between sets of facts expressed in them. In a sentence, facts are often expressed by pairs of words, or *bunsetsu* segments in Japanese, one of which depends on the other. We introduce a notion of a *dependency triple* that consists of two dependent words and their relation type, which is often represented by postpositional particles in the sentences. Therefore, sentence meaning is modeled as a set of dependency triples, and textual entailment as identifying how much of dependency triples in t_2 are subsumed in t_1. By taking this approach, we can deal nicely with word dependency as well as word transposition and adnominal clauses.

In order to determine to what extent a triple is subsumed in a text, we need to compare two triples. We define similarity measures between two triples considering both conceptual and character-based aspects. With these measures, a subsumption score is calculated, and textual entailment is identified depending on whether the subsumption score is greater than a threshold value predetermined through an experiment with the training data.

2. ALGORITHM

2.1 Overview

Processing flow of our system is illustrated in Figure 1. When a text pair (t_1, t_2) is given, we first conduct linguistic analysis of each text. Then we extract dependency triples from each text, and obtain two sets of dependency triples corresponding to t_1 and t_2. Comparing these sets, we calculate a subsumption score, and identify whether t_1 entails t_2 or not based on this score.

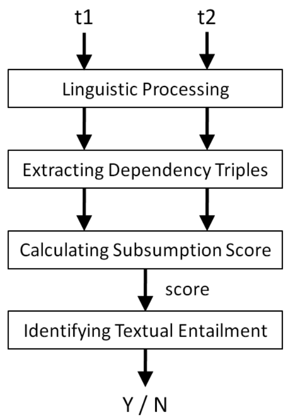

Figure 1: Processing Flow

2.2 Linguistic Processing

First, we conduct the morphological analysis using MeCab [1] and the dependency analysis using CaboCha [2] for input Japanese texts t_1 and t_2. Then, referring to the EDR Japanese word dictionary [3], concept IDs are attached to the head word of each bunsetsu segment in t_1 and t_2.

2.3 Extracting Dependency Triples

In this phase, a set of dependency triples is created from each text. A dependency triple (w_1, p, w_2) consists of two words w_1 and w_2, where w_1 depends on w_2, and the relation type p between w_1 and w_2. For each pair of dependent words contained in the dependency analysis result of a text, we create a dependency triple, in which the relation type is typically identified with a postpositional particle of the dependent word. Exceptions are adnominal clauses and binding particles such as "*wa*" and "*mo*"; in these contexts, appropriate case particles are inferred using heuristic rules.

For various reasons, similar facts are not necessarily expressed by the same dependency structure in different texts. Therefore, we need to expand the set of dependency triples in order to enable more flexible matching and achieve higher recall. We apply the expansion rules shown in Table 1 to the triple set for t_1.

Table 1: Rules for Expanding a Triple Set

Condition	Added Triple
(w_1, p, w_2) where p is a coordinating particle (CP) such as "*to*" and "*ya*"	(w_2, p, w_1)
$(w_1, p_1, w_2), (w_2, p_2, w_3)$ where p_1 is either a CP or adnominal particle "*no*"	(w_1, p_2, w_3)
$(w_1, p_1, w_2), (w_2, p_2, w_3)$ where p_1 is neither a CP nor "*no*"	(w_1, p_1, w_3)

2.4 Calculating Subsumption Score

In order to identify how much of dependency triples in t_2 are subsumed in t_1, we search for the most similar t_1 triple to each t_2 triple, and calculate a subsumption score by averaging the similarity values. Assume the set of dependency triples created from t_1 is $\{u_1, u_2, \ldots, u_n\}$, and the set created from t_2 is $\{v_1, v_2, \ldots, v_m\}$. Then the subsumption score is calculated as follows:

$$\text{subsumption score} = \frac{1}{m}\sum_{j=1}^{m}\left(\max_{1\le i\le n}(sim(u_i, v_j))\right)$$

where $sim(u_i, v_j)$ is a similarity value between two triples u_i and v_j.

The similarity value between triples $u_i = (w_1, p, w_2)$ and $v_j = (w'_1, p', w'_2)$ is calculated as a product $sim(w_1, w'_1) \cdot sim(p, p') \cdot sim(w_2, w'_2)$. In our experiment, the similarity between words w and w' is calculated by three methods. The first method calculates the conceptual similarity based on the EDR concept classification dictionary. We use a variation of the path length based method [4], which considers the proportion between depths of concepts of both words and the most specific concept that subsumes them. The second method calculates the surface character-based similarity, that is, the ratio of the number of characters commonly contained in w and w' divided by the number of characters in w'. The third method is a conjunction of the first and the second methods, that adopts the maximum value of them. The similarity between the relation types p and p' is 1 if $p = p'$, and α otherwise, where α is a constant between 0 and 1.

2.5 Identifying Textual Entailment

We conclude that t_1 entails t_2 if the calculated subsumption score is greater than a threshold value predetermined through a preliminary experiment using the training data. Figure 2 shows an example of the relationship between threshold value and accuracy taken from the preliminary experiment results. In this example, we only consider the conceptual similarity between words in calculation. In this case, we determine the threshold value as 0.3, which maximize accuracy to 0.580.

2.6 Example

In this section, I illustrate our algorithm with the following simple example from the training data.

t1: アゴヒゲアザラシは北極圏や亜北極圏に生息する。
(Bearded seals live in the Arctic and Subarctic Circle.)

t2: アゴヒゲアザラシは北極海に生息する。
(Bearded seals live in the Arctic Ocean.)

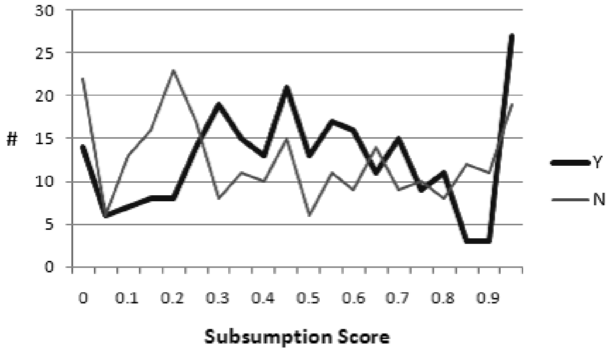

Figure 2: Plot of Threshold and Accuracy

The dependency triples extracted from each text and the most similar pairs are shown in Figure 3.

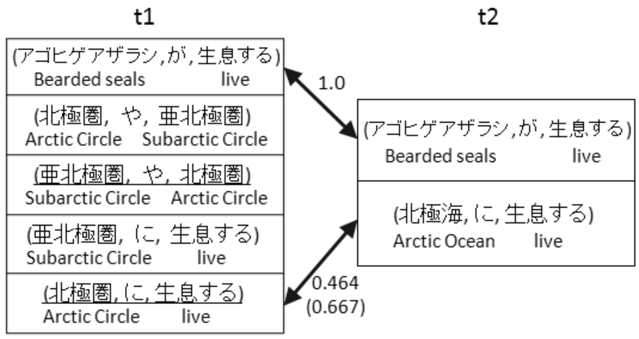

Figure 3: Subsumption of Dependency Triples

Underlined triples in the figure are obtained by the expansion rules in Table 1. The conceptual similarity between the Arctic Circle and the Arctic Ocean is 0.464, and the character-based similarity between them is 0.667 (two common characters out of three). Therefore, the subsumption score between t_1 and t_2 is $(1.0 + 0.464)/2 = 0.732$ if we use the conceptual similarity, and $(0.1 + 0.667)/2 = 0.833$ if we use the character-based similarity.

3. EXPERIMENT

3.1 Formal Run Results

For the formal run, we submitted three runs, which are summarized in Table 2.

Table 2: Submitted Runs

Run ID	Similarity Calculation			Threshold
	Concept	Char	Rel	
SITLP-JA-BC-01	√			0.30
SITLP-JA-BC-02	√		√	0.25
SITLP-JA-BC-03	√	√	√	0.30

Three runs differ in the method to calculate similarity between dependency triples, which we explained in Section

Table 3: Result of Submitted Runs

SITLP-JA-BC-01		Answer Y	Answer N	Total
System	Y	181	173	354
	N	69	77	146
Total		250	250	500

Accuracy = 0.516

SITLP-JA-BC-02		Answer Y	Answer N	Total
System	Y	163	157	320
	N	87	93	180
Total		250	250	500

Accuracy = 0.512

SITLP-JA-BC-03		Answer Y	Answer N	Total
System	Y	168	171	339
	N	82	79	161
Total		250	250	500

Accuracy = 0.494

2.4. "Concept" and "Char" mean that the conceptual similarity and the character-based similarity are used respectively. In SITLP-JA-BC-03, the maximum of two similarity values is used in calculation. In SITLP-JA-BC-01, the similarity and difference of the relation type is not considered, that is, $\alpha = 1$. In the other runs, α is set to 0.5.

The results of the formal runs are shown in Table 3. The highest accuracy is 0.516 in SITLP-JA-BC-01.

3.2 Post-Submission Experiment

We conducted an additional experiment to see whether a more appropriate threshold value exists. This experiment is similar to the preliminary one explained in Section 2.5 except that it uses the formal run data. Figure 4 shows the relationship between threshold value and accuracy taken from the experiment result. The similarity calculation method used here is the same as one in Figure 2.

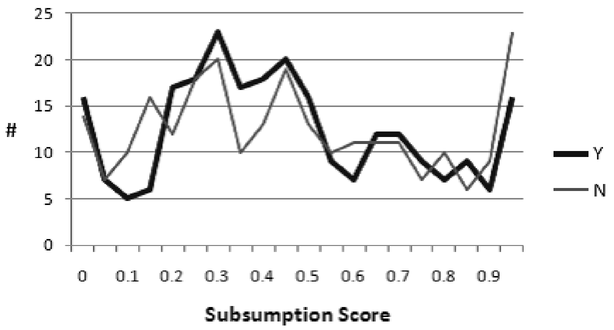

Figure 4: Plot of Threshold and Accuracy

The threshold value that maximize the accuracy is 0.2, and in that case, the accuracy is 0.526.

4. DISCUSSION

Figures 2 and 4 show us that the subsumption score calculated by our algorithm is inadequate to deal with various patterns of textual entailment occur in the training and the formal run data. In this section, we analyze the reasons for the mismatch between our subsumption score and the correct answer.

First, we analyze cases in which the subsumption score is high but the correct answer is N, that is, t_1 does not entail t_2. From the training and the formal run data, we extract 62 cases in which the score is higher than 0.9 and the answer is N, and classify reasons for the mismatch (see Table 4).

"Non-monotonicity" means that t_1 is obtained from t_2 by adding some words (in our words, t_1 subsumes t_2), but t_1 does not entail t_2. The following text pair is an example.

Table 4: Reasons for Entailment Misrecognition

Reason	Count
Small but essential difference	19
Non-monotonicity	10
Similar compound nouns	8
Negation	7
Different case structures	5
Complement clause	4
Superfluous matching	3
Others	6

t1: 東京オリンピックは五輪史上初めて海外に衛星中継された。
(The Tokyo Olympics were broadcasted overseas by satellite for the first time in Olympic.)

t2: 東京オリンピックは初めて衛星中継された。 (The Tokyo Olympics were broadcasted by satellite for the first time.)

Phrases such as "*for the first time*" are called to create *downward-monotone* contexts, and the natural logic approach deals with this type of inference [5].

Second, we analyze cases in which the subsumption score is low but the correct answer is Y. From the training and the formal run data, we extracted 69 cases in which the score is lower than 0.2 and the answer is Y, and classify reasons for the mismatch (see Table 5).

Table 5: Reasons for Recognition Failure

Reason	Count
Commonsense reasoning	22
Paraphrase (between noun and predicate)	17
Paraphrase (others)	16
Different syntactic structures	12
Parenthesis in apposition	9
Quotation	8
Numerical reasoning	4
More than one sentences	3
Others	10

Here, more than one reason is possibly attached to each case. The following is an example of paraphrase between a noun phrase and a verb phrase.

t1: モロッコ周辺海域は良好な漁場だ。
(Water near Morocco is a good fishing area.)

t2: モロッコ周辺は魚がよく取れる。
(You can easily catch fish near Morocco.)

In order to deal with this type of reasoning as well as commonsense reasoning, we need to use other lexical and commonsense knowledge resources.

5. CONCLUSION

In this paper, we reported the evaluation results of our textual entailment system at NTCIR-9 RITE task. We described our algorithm and the evaluation results, and discussed on the reasons for the mismatch between the calculated score and the correct answer.

6. REFERENCES

[1] Taku Kudo, Kaoru Yamamoto and Yuji Matsumoto: Applying Conditional Random Fields to Japanese Morphological Analysis, Proceedings of EMNLP 2004.

[2] Taku Kudo and Yuji Matsumoto: Japanese Dependency Analysis using Cascaded Chunking, Proceedings of 6th CoNLL, 2002.

[3] EDR: EDR Electronic Dictionary Technical Guide, Japan Electronic Dictionary Research Institute, Ltd., 1993.

[4] Zhibiao Wu and Martha Palmer: Verb Semantics and Lexical Selection, Proceedings of 32nd ACL, 1994.

[5] Bill MacCartney and Christopher D. Manning: Natural Logic for Textual Inference, Proceedings of WTEP 2007.