TU Group at NTCIR9-RITE: Leveraging Diverse Lexical Resources for Recognizing Textual Entailment

Yotaro Watanabe Junta Mizuno Eric Nichols
Katsuma Narisawa Keita Nabeshima Kentaro Inui
Graduate School of Information Sciences
Tohoku University, Japan
{yotaro-w,junta-m,eric,katsuma,nabeshima,inui}@ecei.tohoku.ac.jp

ABSTRACT

This paper describes the TU system that participated in the Entrance Exam Subtask of NTCIR-9 RITE. The system consists of two phases: alignment and entailment relation recognition. In the alignment phase, the system aligns words in the two sentences by exploiting diverse lexical resources such as entailment information, hypernym-hyponym relations and synonyms. Based on the alignments and relations between them, the system recognizes semantic relations between two sentences. Our system achieved an accuracy of 0.672 on the development data, and an accuracy of 0.6493 on the formal run.

Keywords

recognizing textual entailment, lexical resources, alignment

1. INTRODUCTION

Our TU system for the NTCIR-9 RITE [9] Entrance Exam Subtask is based on the semantic relation recognizer developed in the Statement Map project [7]. An important feature of the system is that the system uses diverse lexical resources including predicate semantic relations and hypernym-hyponym relations. In addition, the system uses results obtained from several linguistic analyses for alignment and entailment relation recognition: syntactic dependencies, predicate-argument structures, factuality information and sentiment polarity information. Also, the system performs not only word alignment but also structure-based alignment (it aligns the edges in syntactic and semantic dependencies).

The dataset used in the Entrance Exam Subtask is domain specific. The sentences in examples describe in terms of history or social things, and contain many person names, time expressions, etc. In order to deal with them, we added domain specific lexical knowledge and a temporal expression reasoner to the system.

This paper is organized as follows. We at first describe the details of our system in Section 2, then report the results of the run on the development data and the formal run data in Section 3. Next, we describe the error types of our system in Section 4, and finally conclude in Section 5.

2. SYSTEM DESCRIPTION

This section describes the details of our system.

2.1 Preprocessing

Given the two sentences t_1 and t_2, the system conducts the following preprocessing: morphological analysis [4], dependency parsing [3], predicate-argument structure analysis [11], factuality analysis using the resource provided in [5] and sentiment polarity analysis.

2.2 Alignment

2.2.1 Surface-based Alignment

When all of the content words in a phrase in t_2 are all contained in a phrase in t_1, they are aligned. Even if the number of words in the phrase of t_2 is greater than it of t_1, they are aligned.

2.2.2 Knowledge-based Alignment

We use the following resources to determine semantic similarity.

Ontologies We use the Japanese WordNet [1] to check for hypernymy and synonymy between words. E.g. ⟨ 効果 *kouka* "good effect" - 作用 *sayou* "effect"⟩

In addition, we use Wikipedia to check hypernymy [10]. Wikipedia is also used to check synonymy. The synonym word pairs are extracted automatically from the redirect database in Wikipedia. In Wikipedia, some words are redirected to other more frequently used phrases.

Predicate databases To determine if two predicates are semantically related, we consult a database of predicate relations [6] and a database of predicate entailments [2] using the predicates' default case frames. E.g. ⟨ 維持する *iji-suru* "to preserve" - 守る *mamoru* "to maintain"⟩ and ⟨ 予防する *yobou-suru* "to prevent" - 気をつける *ki-wo-tsukeru* "to be careful"⟩

Paraphrases obtained from parentheses We observe that terms are often followed by paraphrases give in brackets. We exploit this pattern to obtain additional synonym word pairs. This operation is done automatically before all other analysis. Then because bracketed expressions often cause errors in dependency parsing, the bracketed expressions are removed.

(1) t_1 16 世紀に入り、海禁政策が弛緩してアメリカ大陸や日本から多くの 銀（メキシコ銀、日本銀）が中国に流入した。“In the 16th century, a lot of silver (Mexican and Japanese silver) poured into China from America and Japan when the Haijin Policy was relaxed.”

t_2 明代には、中国で日本銀が流通した。 “In the Ming era, Japanese silver circulated throughout China.”

For example, in (1), ⟨ 銀 *gin* “silver”⟩ has a bracket. According to our strategy, ⟨ メキシコ銀 *Mekishiko-gin* “Mexican silver”⟩ and ⟨ 日本銀 [“Japanese silver”Nihongin]⟩ are synonyms of ⟨ 銀 *gin* “silver”⟩. In the alignment phrase, after removing the bracketed phrase, ⟨ 銀 *gin* “silver”⟩ of t_1 and ⟨ 日本銀 *Nihon-gin* “Japanese silver”⟩ of t_2 are aligned by this method.

During the alignment phase, when a pair of phrases, one from t_1 and the other from t_2, is found in one of the above resources, the phrases are aligned. Phrases are matched against the resources using a word-level bi-gram cosine-based similarity measure [8].

2.2.3 *Structure-based Alignment*

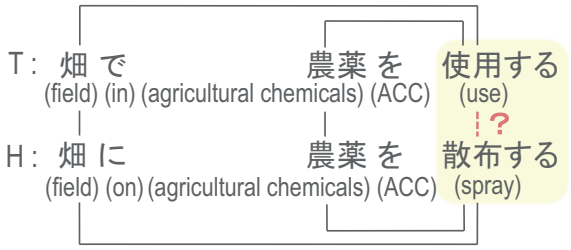

Figure 1: Determining the compatibility of semantic structures

We compare the predicate-argument structure of the query to that of the text and determine if the argument structures are compatible. This process is illustrated in Figure 1 where the T(ext) “Agricultural chemicals are used in the field.” is aligned with the H(ypothesis) “Agricultural chemicals are sprayed on the field.” Although the verbs “used” and “sprayed” are not directly semantically related, they are aligned because they share the same argument structures. In this way, we can align predicates which we lack lexical semantic resources for.

2.3 Entailment Relation Recognition

Our approach to entailment relation recognition consists of two phases: (1) relevance recognition and (2) semantic relation recognition. Given a pair of sentences, the system at first determines relevance using a set of alignments (1). A pair is classified as “relevant” if all of the phrases in t_2 are aligned to phrases in t_1, and “irrelevant” otherwise. However we made exceptions in the above condition. Phrases in t_2 are allowed to be unaligned if the headwords of the phrases contain *light verbs*. If the pair is classified as “irrelevant” then the system outputs “non-entailment”. Otherwise, the system classifies the semantic relation (“entailment” or “contradiction”) of relevant pairs (2). Contradiction relations are determined by considering the semantic relation of an alignment (e.g. if the aligned predicates have an antonym relation), factualities (e.g. factive - counter-factive), and sentiment polarities.

3. RESULTS

We entered three settings **TU1**,**TU2** and **TU3** in the formal run. In **TU1**, the system performs the three steps described above and classifies all of the examples in the dataset. The threshold of cosine similarity used in the alignment phase was set to 0.6. In the two settings **TU2** and **TU3**, performances of the system are evaluated with the examples in which t_2 is a simple sentence, because our system is constructed based on the system of Statement Map [7] in which hypotheses are assumed to be simple sentences. **TU3** is the same as **TU2** except that the system uses only structure-based alignments in entailment relation recognition.

The results on the development data and formal run data are shown in Table 1. TU2 and TU3 achieved significant improvements of performance especially on recall compared to TU1. This results suggest that our system performs well to the examples in which hypothesis has a simple syntactic structure. Although TU3 achieved a slightly higher precision compared to TU2, the performances are the same on accuracy, therefore, the structural alignment approach is less effective on this dataset.

Table 2 shows the number of phrase alignments on the development data for each alignment method, including different lexical resources, employed by our system. The resource making the greatest contribution was Wikpedia since there are many named entities including person names, locations and countries in the dataset. Also, Japanese WordNet and the database of predicate entailments were effective. Note that all of the alignments except for structure-based alignments may have overlaps with multiple resources. Also, there are many false positives in structure-based alignments.

4. ERROR ANALYSIS

Most of the errors are due to false negatives of alignments. We show major error types with examples[1] in the following.

The majority of errors are caused by lack of lexical, paraphrase, and verb entailment knowledge. The following examples are misclassified as N (Y is the correct answer) due to lack of lexical knowledge: ⟨ 征服する *seifuku-suru* “conquer” - 滅ぼす *horobosu* “destroy”⟩ and ⟨ 管轄する *kankatsu-suru* “have jurisdiction” - 統括する *toukatsu-suru* “unify”⟩.

Also, due to lack of paraphrase and entailment relation knowledge, the aligner provided false negatives: ⟨ インフレーション対策として “as a counter-inflation measures” - 物価上昇を/抑制する/ため “to curb price increases”⟩ ⟨ 自発性を/重んじる *jihatsusei-wo / omonjiru* “respect for initiative” - 自主性を/最大限に/発揮させる “exercise their autonomy in their own best”⟩.

The dataset used in the Entrance Exam Subtask contains various types of time expressions. As the time expression reasoner of the system has limited rules, it provided many false negatives: e.g. ⟨16 世紀 *16-seiki* “16th century” - 明代 *mindai* “the Ming era”⟩. Also, if there are modifiers on time expressions (e.g. beginning of), it provides 1-to-n alignments

[1]Some examples used in this section are slightly modified for ease of explanation.

	devel							formal run
	Prec. (Y)	Rec. (Y)	F1 (Y)	Prec. (N)	Rec. (N)	F1 (N)	Acc.	Acc.
TU1	0.733 (63/86)	0.310 (63/203)	0.436	0.659 (271/411)	0.922 (271/294)	0.769	0.672	0.649 (284/442)
TU2	0.750 (24/32)	0.667 (24/36)	0.706	0.797 (47/59)	0.855 (47/55)	0.825	0.780	0.718 (50/71)
TU3	0.767 (23/30)	0.639 (23/36)	0.697	0.787 (48/61)	0.873 (48/55)	0.828	0.780	0.718 (50/71)

Table 1: Results on the development data and the formal run data.

	WN	predicate relations	predicate entailments	Wikipedia	parenthesis	struct-based align.
#	817	81	414	1810	20	758

Table 2: The number of phrase alignments per resource/approach on the development data.

e.g. ⟨902 年 "in 902" → 10 世紀/初め "in the beginning of 10th century"⟩. Since the modifier is not aligned to any phrases in t_1, it causes incorrect entailment relation recognition.

A few examples are incorrectly classified as "entailment" due to misclassifications of factuality information.

(2) t_1 永住資格を持つ在日外国人に選挙権を付与する法案は、廃案となった。"A proposal to grant the right to vote to foreigners with permanent residency status in Japan was rejected."

t_2 永住資格を有する在日外国人も選挙権を持つ。"Foreigners with permanent residency status in Japan have the right to vote."

In this case, the factuality of the event ⟨ 選挙権を付与 "grant the right to vote"⟩ must be "counter-fact" , however, our factuality analyzer mistakenly labeled "fact" to the event.

(3) t_1 総務省が消防職員への団結権付与について検討することを決めた。"The Ministry of Internal Affairs and Communications decided to examine granting the right to organize to workers in fire departments."

t_2 消防職員には団結権が保証されていない。"Workers in fire departments are not guaranteed the right to organize."

In (3), ⟨ 検討する "examine"⟩ *presupposes* that the event ⟨ 付与 "grant"⟩ is "counter-fact", however, the system also misclassified the factuality of this event as "fact".

The following examples are instances of "entailment" that are misclassified by our system because t_2 contains a specific information not included in t_1:

(4) t_1 鎌倉幕府は 1192 年に始まったとされていたが，現在では実質的な成立は 1185 年であるとする説が支配的である。"The Kamakura Shogunate had been considered to be establised in 1192, however currently the dominant theory is that it was actually established in 1185."

t_2 12 世紀に 日本では鎌倉幕府が開かれた。"The Kamakura Shogunate was established in the 12nd century in Japan."

(5) t_1 デイヴィッド・リヴィングストンはヨーロッパ人で初めてアフリカ大陸を横断し、現地の状況を詳細に報告した。"David Livingstone was the first European to cross Africa. He gave a detailed report of the area."

t_2 19 世紀、リヴィングストンはアフリカ内陸部の探検を行った。"In the 19th century, Livingstone explored inner Africa."

These examples require additional knowledge to infer entailment relations: in (4), Kamakura shogunate was established in Japan, and in (5), David Livingstone lived from 1813 to 1873 i.e. during the 19th century.

Some examples requires more complex inference to determine the correct entailment relation.

(6) t_1 日本・イギリス・アメリカなどは、ロシア革命に対してシベリア出兵を行い、日本軍は最後までシベリアに残っていた。"The countries including Japan, UK and USA sent troops into Siberia in response to the Russia Revolution, and only Japan remained until the end."

t_2 日本は、ロシア革命に対してイギリスなど他の国よりも長期にわたって介入を継続した。"Japan intervened in the Russia Revolution for a longer period than all the other countries."

In order to infer that ⟨ 日本は最後まで残っていた "only Japan remained until the end"⟩ implies ⟨ 他の国よりも長期にわたって介入 "intervene for a longer period than all the other countries"⟩, systems are required to recognize ⟨ イギリスなどの他の国 "other countries including UK"⟩ corresponds to ⟨ イギリス・アメリカ "UK and USA"⟩ and deal with the comparative expression ⟨ イギリスなど他の国 よりも "compared to the other countries including UK"⟩.

In the following example, it is difficult to obtain the correct alignment since t_1 describes multiple and more specific events which correspond to one predicate in t_2.

(7) t_1 グスタフ・シュトレーゼマン首相はインフレ沈静化のため、ドイツ・レンテン銀行を設立し、レンテンマルクを発行した。"In order to reduce inflation, the prime minister Gustav Stresemann founded the Deutsche Rentenbank and issued the Rentenmark currency."

t_2 シュトレーゼマンがインフレーション対策のために改革を行った。"Stresemann made reforms to reduce inflation."

⟨ 改革を/行った "made reforms"⟩ in t_2 corresponds to multiple events, and these describes more specific level compared to t_1. How to deal with these kinds of examples is an open problem.

5. CONCLUSION

In this paper we described the TU system for the NTCIR-9 RITE Entrance Exam Subtask. The results of the experiments and the error analysis suggest that majority of the errors still result from lack of lexical knowledge. We are planning to construct massive lexical and verb entailment knowledge and exploit them for recognizing textual entailment.

6. REFERENCES

[1] F. Bond, H. Isahara, S. Fujita, K. Uchimoto, T. Kuribayashi, and K. Kanzaki. Enhancing the Japanese WordNet. In *Proc. of ACL-IJCNLP*, 2009.

[2] C. Hashimoto, K. Torisawa, K. Kuroda, M. Murata, and J. Kazama. Large-Scale Verb Entailment Acquisition from the Web. In *Proc. of EMNLP*, pages 1172–1181, 2009.

[3] T. Kudo and Y. Matsumoto. Japanese dependency analysis using cascaded chunking. In *Proc. of CoNLL*, pages 63–69, 2002.

[4] T. Kudo, K. Yamamoto, and Y. Matsumoto. Appliying conditional random fields to japanese morphological analysis. In *In Proc. of EMNLP*, 2004.

[5] S. Matsuyoshi, M. Eguchi, C. Sao, K. Murakami, K. Inui, and Y. Matsumoto. Annotating event mentions in text with modality, focus, and source information. In *Proc. of LREC*, 2010.

[6] S. Matsuyoshi, K. Murakami, Y. Matsumoto, and K. Inui. A database of relations between predicate argument structures for recognizing textual entailment and contradiction. In *Proc. of ISUC*, pages 366–373, 2008.

[7] K. Murakami, E. Nichols, J. Mizuno, Y. Watanabe, S. Masuda, H. Goto, M. Ohki, C. Sao, S. Matsuyoshi, K. Inui, and Y. Matsumoto. Statement map: Reducing web information credibility noise through opinion classification. In *Proceedings of the Fourth Workshop on Analytics for Noisy Unstructured Text Data (AND)*, pages 59–66, 2010.

[8] N. Okazaki and J. Tsujii. Simple and efficient algorithm for approximate dictionary matching. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 851–859, Beijing, China, August 2010.

[9] 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 *NTCIR-9 Proceedings (to appear)*, 2011.

[10] A. Sumida, N. Yoshinaga, and K. Torisawa. Boosting precision and recall of hyponymy relation acquisition from hierarchical layouts in Wikipedia. In *Proc. of LREC*, 2008.

[11] Y. Watanabe, M. Asahara, and Y. Matsumoto. A structured model for joint learning of argument roles and predicate senses. In *Proceedings of the 48th Annual Meeting of the Association of Computational Linguistics*, pages 98–102, 2010.