Building Japanese Textual Entailment Specialized Data Sets for Inference of Basic Sentence Relations

Kimi Kaneko † **Yusuke Miyao** ‡ **Daisuke Bekki** †
† Ochanomizu University, Tokyo, Japan
‡ National Institute of Informatics, Tokyo, Japan
† {kaneko.kimi | bekki}@is.ocha.ac.jp
‡ yusuke@nii.ac.jp

Abstract

This paper proposes a methodology for generating specialized Japanese data sets for textual entailment, which consists of pairs decomposed into basic sentence relations. We experimented with our methodology over a number of pairs taken from the RITE-2 data set. We compared our methodology with existing studies in terms of agreement, frequencies and times, and we evaluated its validity by investigating recognition accuracy.

1 Introduction

In recognizing textual entailment (RTE), automated systems assess whether a human reader would consider that, given a snippet of text t1 and some unspecified (but restricted) world knowledge, a second snippet of text t2 is true. An example is given below.

Ex. 1) Example of a sentence pair for RTE
- Label: Y
- t1: Shakespeare wrote *Hamlet* and *Macbeth*.
- t2: Shakespeare is the author of *Hamlet*.

"Label" on line 1 shows whether textual entailment (TE) holds between t1 and t2. The pair is labeled 'Y' if the pair exhibits TE and 'N' otherwise.

It is difficult for computers to make such assessments because pairs have multiple interrelated basic sentence relations (BSRs, for detailed information on BSRs, see section 3). Recognizing each BSRs in pairs exactly is difficult for computers. Therefore, we should generate specialized data sets consisting of t1-t2 pairs decomposed into BSRs and a methodology for generating such data sets since such data and methodologies for Japanese are unavailable at present.

This paper proposes a methodology for generating specialized Japanese data sets for TE that consist of *monothematic* t1-t2 pairs (i.e., pairs in which only one BSR relevant to the entailment relation is highlighted and isolated). In addition, we compare our methodology with existing studies and analyze its validity.

2 Existing Studies

Sammons *et al.*(2010) point out that it is necessary to establish a methodology for decomposing pairs into chains of BSRs, and that establishing such methodology will enable understanding of how other existing studies can be combined to solve problems in natural language processing and identification of currently unsolvable problems. Sammons *et al.* experimented with their methodology over the RTE-5 data set and showed that the recognition accuracy of a system trained with their specialized data set was higher than that of the system trained with the original data set. In addition, Bentivogli *et al.*(2010) proposed a methodology for classifying more details than was possible in the study by Sammons *et al.*.

However, these studies were based on only English data sets. In this regard, the word-order rules and the grammar of many languages (such as Japanese) are different from those of English. We thus cannot assess the validity of methodologies for any Japanese data set because each language has different usages. Therefore, it is necessary to assess the validity of such methodologies with specialized Japanese data sets.

Kotani *et al.* (2008) generated specialized Japanese data sets for RTE that were designed such that each pair included only one BSR. However, in that approach the data set is generated artificially, and BSRs between pairs of real world texts cannot be analyzed.

We develop our methodology by generating specialized data sets from a collection of pairs from RITE-2[1] binary class (BC) subtask data sets containing sentences from Wikipedia. RITE-2 is

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 273–277,
Sofia, Bulgaria, August 4-9 2013.

an evaluation-based workshop focusing on RTE. Four subtasks are available in RITE-2, one of which is the BC subtask whereby systems assess whether there is TE between t1 and t2. The reason why we apply our methodology to part of the RITE-2 BC subtask data set is that we can consider the validity of the methodology in view of the recognition accuracy by using the data sets generated in RITE-2 tasks, and that we can analyze BSRs in real texts by using sentence pairs extracted from Wikipedia.

3 Methodology

In this study, we extended and refined the methodology defined in Bentivogli *et al.*(2010) and developed a methodology for generating Japanese data sets broken down into BSRs and non-BSRs as defined below.

Basic sentence relations (BSRs):
- *Lexical:* Synonymy, Hypernymy, Entailment, Meronymy;
- *Phrasal:* Synonymy, Hypernymy, Entailment, Meronymy, Nominalization, Corference;
- *Syntactic:* Scrambling, Case alteration, Modifier, Transparent head, Clause, List, Apposition, Relative clause;
- *Reasoning:* Temporal, Spatial, Quantity, Implicit relation, Inference;

Non-basic sentence relations (non-BSRs):
- *Disagreement:* Lexical, Phrasal, Modal, Modifier, Temporal, Spatial, Quantity;

Mainly, we used relations defined in Bentivogli *et al.*(2010) and divided **Synonymy**, **Hypernymy**, **Entailment** and **Meronymy** into *Lexical* and *Phrasal*. The differences between our study and Bentivogli *et al.*(2010) are as follows. **Demonymy** and **Statements** in Bentivogli *et al.*(2010) were not considered in our study because they were not necessary for Japanese data sets. In addition, **Scrambling**, **Entailment**, **Disagreement: temporal**, **Disagreement: spatial** and **Disagreement: quantity** were newly added in our study. **Scrambling** is a rule for changing the order of phrases and clauses. **Entailment** is a rule whereby the latter sentence is true whenever the former is true (e.g., "divorce" → "marry"). **Entailment** is a rule different from **Synonymy**, **Hypernymy** and **Meronymy**.

The rules for decomposition are schematized as follows:

- Break down pairs into BSRs in order to bring t1 close to t2 gradually, as the interpretation of the converted sentence becomes wider
- Label each pair of BSRs or non-BSRs such that each pair is decomposed to ensure that there are not multiple BSRs

An example is shown below, where the underlined parts represent the revised points.

t1：	シェイクスピアは ハムレット や マクベスを 書いた。
	Shakespeare$_{nom}$ Hamlet $_{com}$ Macbeth$_{acc}$ write$_{past}$
	'Shakespeare wrote Hamlet and Macbeth.'
[List]	シェイクスピアは ハムレットを 書いた。
	Shakespeare$_{nom}$ Hamlet$_{acc}$ write$_{past}$
	'Shakespeare wrote Hamlet.'
t2：**[Synonymy]**	シェイクスピアは ハムレットの 作者 である。
：*phrasal*	Shakespeare$_{nom}$ Hamlet$_{gen}$ author$_{comp}$ be$_{cop}$
	'Shakespeare is the author of Hamlet.'

Table 1: Example of a pair with TE

An example of a pair without TE is shown below.

t1：	ブルガリアは ユーラシア大陸に ある。
	Bulgaria$_{nom}$ Eurasia.continent$_{dat}$ be$_{cop}$
	'Bulgaria is on the Eurasian continent.'
[Entailment]	ブルガリアは 大陸国家 である。
：*phrasal*	Bulgaria$_{nom}$ continental.state$_{comp}$ be$_{cop}$
	'Bulgaria is a continental state.'
t2：**[Disagreement]**	ブルガリアは 島国 である。
：*lexical*	Bulgaria$_{nom}$ island.country$_{comp}$ be$_{cop}$
	'Bulgaria is an island country.'

Table 2: Example of a pair without TE (Part 1)

To facilitate TE assessments like Table 3, non-BSR labels were used in decomposing pairs. In addition, we allowed labels to be used several times when some BSRs in a pair are related to 'N' assessments.

t1：	ブルガリアは ユーラシア大陸に ある。
	Bulgaria$_{nom}$ Eurasia.continent$_{dat}$ be$_{cop}$
	'Bulgaria is on the Eurasian continent.'
[Disagreement]	ブルガリアは ユーラシア大陸に ない。
：*modal*	Bulgaria$_{nom}$ Eurasia.continent$_{dat}$ be$_{cop-neg}$
	'Bulgaria is not on the Eurasian continent.'
t2：**[Synonymy]**	ブルガリアは ヨーロッパに 属さない。
：*lexical*	Bulgaria$_{nom}$ Europe$_{dat}$ belong$_{cop-neg}$
	'Bulgaria does not belong to Europe.'

Table 3: Example of a pair without TE (Part 2)

As mentioned above, the idea here is to decompose pairs in order to bring t1 closer to t2, the latter of which in principle has a wider semantic scope. We prohibited the conversion of t2 because it was possible to decompose the pairs such that they could be true even if there was no TE. Nevertheless, since it is sometimes easier to convert t2,

[1] http://www.cl.ecei.tohoku.ac.jp/rite2/doku.php

we allowed the conversion of t2 in only the case that t1 contradicted t2 and the scope of t2 did not overlap with that of t1 even if t2 was converted and TE would be unchanged. An example in case that we allowed to convert t2 is shown below. Boldfaced types in Table 4 shows that it becomes easy to compare t1 with t2 by converting to t2.

t1：	トムは 今日、朝食を 食べなかった。
	Tom_{nom} today breakfast_{acc} $\text{eat}_{past-neg}$
	'Tom didn't eat breakfast today.'
[Scrambling]	今日、 トムは 朝食を **食べなかった**。
	today Tom_{nom} breakfast_{acc} $\text{eat}_{past-neg}$
	'Today, Tom **didn't eat** breakfast.'
t2：	今朝、 トムは パンを 食べた。
	this.morning Tom_{nom} bread_{acc} eat_{past}
	'This morning, Tom ate bread and salad.'
[Entailment]	今日、 トムは 朝食を 食べた。
：*phrasal*	today Tom_{nom} breakfast_{acc} eat_{past}
	'Today, Tom ate breakfast.'
[Disagreement]	今日、トムは朝食を**食べた**。
：*modal*	'Today, Tom **ate** breakfast.'

Table 4: Example of conversion of t2

4 Results

4.1 Comparison with Existing Studies

We applied our methodology to 173 pairs from the RITE-2 BC subtask data set. The pairs were decomposed by one annotator, and the decomposed pairs were assigned labels by two annotators. During labeling, we used the labels presented in Section 3 and "unknown" in cases where pairs could not be labeled. Our methodology was developed based on 112 pairs, and by using the other 61 pairs, we evaluated the inter-annotator agreement as well as the frequencies and times of decomposition.

The agreement for 241 monothematic pairs generated from 61 pairs amounted to 0.83 and was computed as follows. The kappa coefficient for them amounted 0.81.

$$Agreement = \text{"}Agreed\text{"}\ labels/Total\ ^{2}$$

Bentivogli *et al.* (2010) reported an agreement rate of 0.78, although they computed the agreement by using the Dice coefficient (Dice, 1945), and therefore the results are not directly comparable to ours. Nevertheless, the close values suggest that our methodology is comparable to that in Bentivogli's study in terms of agreement.

Table 5 shows the distribution of monothematic pairs with respect to original Y/N pairs.

Original pairs		Monothematic pairs		
		Y	N	Total
	Y (32)	116	–	116
	N (29)	96	29	125
	Total (61)	212	29	241

Table 5: Distribution of monothematic pairs with respect to original Y/N pairs

When the methodology was applied to 61 pairs, a total of 241 and an average of 3.95 monothematic pairs were derived. The average was slightly greater than the 2.98 reported in (Bentivogli *et al.*, 2010). For pairs originally labeled 'Y' and 'N', an average of 3.62 and 3.31 monothematic pairs were derived, respectively. Both average values were slightly higher than the values of 3.03 and 2.80 reported in (Bentivogli *et al.*, 2010). On the basis of the small differences between the average values in our study and those in (Bentivogli *et al.*, 2010), we are justified in saying that our methodology is valid.

Table 6 [3] shows the distribution of BSRs in t1-t2 pairs in an existing study and the present study. We can see from Table 6 that **Corference** was seen more frequently in Bentivogli's study than in our study, while **Entailment** and **Scrambling** were seen more frequently in our study. This demonstrates that differences between languages are relevant to the distribution and classification of BSRs.

An average of 5 and 4 original pairs were decomposed per hour in our study and Bentivogli's study, respectively. This indicates that the complexity of our methodology is not much different from that in Bentivogli *et al.*(2010).

4.2 Evaluation of Accuracy in BSR

In the RITE-2 formal run[4], 15 teams used our specialized data set for the evaluation of their systems. Table 7 shows the average of F_1 scores[5] for each BSR.

Scrambling and **Modifier** yielded high scores (close to 90%). The score of **List** was also

[2]Because the "Agreed" pairs were clear to be classified as "Agreed", where "Total" is the number of pairs labeled "Agreed" subtracted from the number of labeled pairs. "Agreed" labels is the number of pairs labeled "Agreed" subtract from the number of pairs with the same label assigned by the two annotators.

[3]Because "lexical" and "phrasal" are classified together in Bentivogli *et al.*(2010), they are not shown separately in Table 6.

[4]In RITE-2, data generated by our methodology were released as "unit test data".

[5]The traditional F_1 score is the harmonic mean of precision and recall.

BSR	Monothematic pairs					
	Bentivogli *et al.*			Present study		
	Total	Y	N	Total	Y	N
Synonymy	25	22	3	45	45	0
Hypernymy	5	3	2	5	5	0
Entailment	-	-	-	44	44	0
Meronymy	7	4	3	1	1	0
Nominalization	9	9	0	1	1	0
Corference	49	48	1	3	3	0
Scrambling	-	-	-	15	15	0
Case alteration	7	5	2	7	7	0
Modifier	25	15	10	42	42	0
Transparent head	6	6	0	1	1	0
Clause	5	4	1	14	14	0
List	1	1	0	3	3	0
Apposition	3	2	1	1	1	0
Relative clause	1	1	0	8	8	0
Temporal	2	1	1	1	1	0
Spatial	1	1	0	1	1	0
Quantity	6	0	6	0	0	0
Implicit relation	7	7	0	18	18	0
Inference	40	26	14	2	2	0
Disagreement: lexical/phrasal	3	0	3	27	0	27
Disagreement: modal	1	0	1	1	0	1
Disagreement: temporal	-	-	-	1	0	1
Disagreement: spatial	-	-	-	0	0	0
Disagreement: quantity	-	-	-	0	0	0
Demonymy	1	1	0	-	-	-
Statements	1	1	0	-	-	-
total	**205**	**157**	**48**	**241**	**212**	**29**

Table 6: Distribution of BSRs in t1-t2 pairs in an existing study and in the present study using our methodology

BSR	F_1(%)	Monothematic Pairs	Miss
Scrambling	89.6	15	4
Modifier	88.8	42	0
List	88.6	3	0
Temporal	85.7	1	1
Relative clause	85.4	8	2
Clause	85.0	14	2
Hypernymy: lexical	85.0	5	1
Disagreement: phrasal	80.1	25	0
Case alteration	79.9	7	2
Synonymy: lexical	79.7	9	6
Transparent head	78.6	1	2
Implicit relation	75.7	18	2
Synonymy: phrasal	73.6	36	9
Corference	70.9	3	1
Entailment: phrasal	70.2	44	7
Disagreement: lexical	69.0	2	0
Meronymy: lexical	64.3	1	1
Nominalization	64.3	1	0
Apposition	50.0	1	1
Spatial	50.0	1	1
Inference	40.5	2	2
Disagreement: modal	35.7	1	0
Disagreement: temporal	28.6	1	1
Total	-	**241**	**41**

Table 7: Average F_1 scores in BSR and frequencies of misclassifications by annotators

nearly 90%, although the data sets included only 3 instances. These scores were high because pairs with these BSRs are easily recognized in terms of syntactic structure. By contrast, **Disagreement: temporal**, **Disagreement: modal**, **Inference**, **Spatial** and **Apposition** yielded low scores (less than 50%). The scores of **Disagreement: lexical**, **Nominalization** and **Disagreement: Meronymy** were about 50-70%. BSRs that yielded scores of less than 70% occurred less than 3 times, and those that yielded scores of not more than 70% occurred 3 times or more, except for **Temporal** and **Transparent head**. Therefore, the frequencies of BSRs are related to F_1 scores, and we should consider how to build systems that recognize infrequent BSRs accurately. In addition, F_1 scores in **Synonymy: phrasal** and **Entailment: phrasal** are low, although these are labeled frequently. This is one possible direction of future work.

Table 7 also shows the number of pairs in BSR to which the two annotators assigned different labels. For example, one annotator labeled t2 **[Apposition]** while the other labeled t2 **[Spatial]** in the following pair:

Ex. 2) Example of a pair for RTE
- t1: Tokyo, the capital of Japan, is in Asia.
- t2: The capital of Japan is in Asia.

We can see from Table 7 that the F_1 scores for BSRs, which are often assessed as different by different people, are generally low, except for several labels, such as **Synonymy: lexical** and **Scrambling**. For this reason, we can conjecture that cases in which computers experience difficulty determining the correct labels are correlated with cases in which humans also experience such difficulty.

5 Conclusions

This paper presented a methodology for generating Japanese data sets broken down into BSRs and Non-BSRs, and we conducted experiments in which we applied our methodology to 61 pairs extracted from the RITE-2 BC subtask data set. We compared our method with that of Bentivogli *et al.*(2010) in terms of agreement as well as frequencies and times of decomposition, and we obtained similar results. This demonstrated that our methodology is as feasible as Bentivogli *et al.*(2010) and that differences between languages emerge only as the different sets of labels and the different distributions of BSRs. In addition, 241 monothematic pairs were recognized by computers, and we showed that both the frequencies of BSRs and the rate of misclassification by humans are relevant to F_1 scores.

Decomposition patterns were not empirically compared in the present study and will be investigated in future work. We will also develop an RTE inference system by using our specialized data set.

References

Bentivogli, L., Cabrio, E., Dagan, I, Giampiccolo, D., Leggio, M. L., Magnini,B. 2010. *Building Textual Entailment Specialized Data Sets: a Methodology for Isolating Linguistic Phenomena Relevant to Inference*. In Proceedings of LREC 2010, Valletta, Malta.

Dagan, I, Glickman, O., Magnini, B. 2005. *Recognizing Textual Entailment Challenge*. In Proc. of the First PASCAL Challenges Workshop on RTE. Southampton, U.K.

Kotani, M., Shibata, T., Nakata, T, Kurohashi, S. 2008. *Building Textual Entailment Japanese Data Sets and Recognizing Reasoning Relations Based on Synonymy Acquired Automatically*. In Proceedings of the 14th Annual Meeting of the Association for Natural Language Processing, Tokyo, Japan.

Magnini, B., Cabrio, E. 2009. *Combining Specializedd Entailment Engines*. In Proceedings of LTC '09. Poznan, Poland.

Dice, L. R. 1945. *Measures of the amount of ecologic association between species*. Ecology, 26(3):297-302.

Mark Sammons, V.G.Vinod Vydiswaran, Dan Roth. 2010. *"Ask not what textual entailment can do for you..."*. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden, pp. 1199-1208.