Japanese Discourse Structure Analysis Based on Automatically Acquired Large-Scale Knowledge

Qinghan Bu, Daisuke Kawahara and Sadao Kurohashi
Graduate School of Informatics, Kyoto University
Yoshida-honmachi, Sakyo-ku, Kyoto, 606-8501, Japan
bu@nlp.ist.i.kyoto-u.ac.jp
{dk,kuro}@i.kyoto-u.ac.jp

Abstract

In this paper, we propose a novel approach to identify coherence relations of Japanese discourses based on automatically acquired large-scale knowledge. In contrast to previous work, we only consider coherence of sentences in the discourse and make use of rich knowledge in the task of discourse structure analysis. Empirical evaluations over our corpus demonstrate that automatically acquired rich knowledge is effective in discourse structure analysis.

1 Introduction

A piece of text is often not to be understood individually, but understood by linking it with other text units from its context. These units can be surrounding clauses, sentences, or even paragraphs. It is important to develop discourse structure analysis, which identifies the links between such text units, to further accelerate the study of natural language understanding.

In this paper, we present a method for Japanese discourse structure analysis by taking advantage of rich knowledge that is automatically acquired from a large corpus. To illustrate our task and idea, consider the following text that consists of three sentences.

(1) a. 昨日、勉強するために図書館に行った。
(Yesterday, I went to the library to study.)
b. 天気はずっと雨だったので、遊びに行かずによかった。
(It rained all day, so it was good that I didn't go out.)
c. ついでに予約していた本も借りてきた。
(I also borrowed the book that I had reserved.)

In this example, the second sentence is a digress sentence and it is possible to remove this sentence for the understanding of this discourse. The third sentence has a stronger relation to the first sentence than the second sentence. Our task is to analyze such sentence-level connections in text. To accurately solve this task, wide-coverage knowledge is indispensable. For instance, the two words "図書館" (library) and "本" (book) in the above example often occur at the same time in a corpus. Our idea is that such cooccurrence knowledge is effective in judging the coherence of discourse. In this case, this cooccurrence knowledge can link the third sentence to the first sentence.

Our empirical evaluations indicate that our knowledge-rich approach outperforms baseline methods without knowledge.

2 Our Discourse-annotated Corpus

In this paper, we focus on texts that consist of three sentences. This is because three sentences is the minimum volume of text that can have a structural ambiguity if a sentence is considered to be a unit. We suppose that it is the first step for general discourse structure analysis to analyze the discourse of three sentences accurately.

2.1 3-sentence Structure

3-sentence discourse structure can be classified into two types: 3-1 type and 3-2 type. Example (1) is an example of 3-sentence structure of the 3-1 type, where the third sentence (1c) has a coherence relation to the first sentence (1a). This 3-1 type also indicates that the second sentence in 3-sentence discourse is a digress sentence or a parenthesis, and can be removed. In contrast, the 3-2 type means that the third sentence has a coherence relation to the second sentence. This type also indicates that the second sentence serves as a link or plays a pivot role between the first sentence and the third sentence, and cannot be removed from the discourse.

2.2 3-sentence Annotated Corpus

We created manual annotations for the 3-sentence web corpus proposed by (Hangyo et al., 2012). We made use of crowd sourcing for this annotation work by using three workers for each 3-sentence text. An annotated 3-sentence text was added into our corpus if all the workers agree with the type. We obtained 14,152 texts in total, which consist of 6,488 texts of the 3-1 type (45.8%) and 7,664 texts of the 3-2 type (54.2%).

3 Knowledge Sources for Discourse Structure Analysis

We propose a method that works on content words and extracts information as knowledge features from different automatically acquired large-scale knowledge sources.

3.1 Case Frames

Case frames represent the relations between a predicate and its arguments. We suppose that if we use large-scale and wide-coverage case frames as a knowledge source, collocation and coherence information between sentences can be extracted from them, which would be useful for discourse structure analysis. We employ wide-coverage Japanese case frames consisting of 150 thousand predicates that were automatically compiled from a web corpus of seven billion sentences (Kawahara and Kurohashi, 2006).

3.2 Related Events

Event relations represent strongly-related event pairs, such as temporal relations, causality and so on. This kind of knowledge would be useful to capture coherence relations in discourse. One example of such event relations for Japanese is the work by (Shibata and Kurohashi, 2011), which automatically acquired strongly-related events from a large corpus using predicate-argument co-occurring statistics and case frames. They represent an event as a predicate-argument structure. We employ 340 thousand strongly-related events that were automatically extracted from 1.6 billion web sentences.

3.3 Related Words

We also use related words as a knowledge source, which represent strongly related nouns to each noun. Related words are collected automatically from approximately 400 million Japanese web sentences and 1.2 million Japanese Wikipedia entries. For a non-ambiguous word, its co-occurring words in the same sentence are first extracted from the web sentences, and a word whose mutual information is high is regarded as a related word. For an ambiguous word, the related words are obtained for each sense in Wikipedia. The definition sentence for each sense is extracted from its disambiguation page, and the related words for each sense are regarded as the related words of nouns in the definition sentence.

4 Our Method

We classify each 3-sentence discourse between the 3-1 type and the 3-2 type using Support Vector Machines (SVMs). In our method, the first step is to segment a sentence into words and parse it. We use the Japanese Morphological Analyzer JUMAN[1] and Japanese dependency parser KNP[2] for this step. Then, we extract features for the classifier. We divide these features into two types: baseline features and knowledge features, which are described below.

4.1 Baseline Features

We extract baseline features for sentence pairs, i.e., pairs of second-third sentences and first-third sentences, respectively. We first enumerate words except whose part-of-speech (POS) is particle, auxiliary verb, copula, special and other meaningless mark. We also use noun and verb features by considering only the words whose POS is noun and verb.[3] Furthermore, we judge if there are overlapping words between these two sentences.

The length of each sentence could influence the result, and thus we also consider the sentence length. Discourse cues, such as conjunctions and demonstratives, can be syntactic features, which frequently indicate the presence of discourse relations. We check the first word of the second sentence and the third sentence, and add conjunction and demonstrative features to our baseline features.

4.2 Knowledge Features

For each type of knowledge feature, we calculate the feature values for sentence pairs, i.e., the second-third sentence pair and the first-third sentence pair, respectively. We also adopt the difference of these values as knowledge features.

4.2.1 Case Frame Features

Case frames are used to capture the relations between a predicate a noun in a sentence pair. For each predicate in a sentence pair, we make couples of (predicate, noun), in which a noun is extracted from the nouns in the sentence that is the other sentence containing the predicate. We use the following score as a knowledge feature:

$$Cf\_feature = \sum_{v} \sum_{n} \mathrm{PMI}_{vn}, \tag{1}$$

where v represents a predicate in a sentence pair, n represents a noun in the sentence that is not the sentence containing the verb, and PMI_{vn} represents a value of pointwise mutual information (PMI) between v and n. This PMI value is calculated from a case frame of v, which is identified by the KNP parser, by using the method proposed by (Sasano and Kurohashi, 2011).

We also count the number of predicate-noun pairs that appear in a case frame. We use the following score as another case frame feature:

$$Cf\_case\_feature = \sum_{v} \sum_{n} \begin{cases} +1 & \mathrm{PMI}_{vn} \text{ exists} \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

4.2.2 Related Event Features

Related events are used to capture the relations between predicates in a sentence pair. Each event pair, i.e., a pair of predicate-argument structures, has a score of strength between the two events. This score is the lift value calculated by the association mining method used in We use the following score as a knowledge feature:

$$Re\_feature = \sum_{i=1}^{n} lift_i / lift_{max}, \tag{3}$$

[1] http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN

[2] http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP

[3] We exclude the verb "する" (do) from the verb feature.

Method	Accuracy
Majority baseline	54.15%
Benchmark	68.73%
Baseline	67.33%
+Cf	68.82%
+Re	67.29%
+Rw	71.19%
+Cf +Re +Rw	72.29%

Table 1: Accuracy by using different methods. '+' represents 'baseline plus.' 'Cf' represents 'Case frames,' 'Re' represents 'Related events,' and 'Rw' represents 'Related words.'

Baseline +	Re	Rw	Cf
Re	67.29%	71.30%	68.97%
Rw		71.19%	72.24%
Cf			68.82%

Table 2: Combination matrix for baseline features plus knowledge features.

where n represents the number of pairs of predicate-argument structures in a sentence pair, $lift_i$ represents the lift value of the the i-th event pair in the knowledge source, and $lift_{max}$ represents the maximum lift value.

4.2.3 Related Word Features

Related words are used to capture the relations between nouns in a sentence pair. Each noun is represented as a vector whose dimension corresponds to its related words. The similarity between two nouns is computed as the cosine similarity between these two vectors. We use the following score as a knowledge feature:

$$Rw\_feature = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} cos\_sim(w_i, w_j)}{m * n}, \quad (4)$$

where m and n represent the number of nouns in each of the two sentences and $cos\_sim(w_i, w_j)$ represents the cosine similarity between the two nouns, i.e., w_i and w_j.

5 Experiments

We conducted experiments using five-fold cross-validation on our Japanese discourse-annotated corpus. We use LIBLINEAR[4] as an implementation of SVMs. We employ a rule-based method for Japanese discourse structure analysis proposed by (Shibata and Kurohashi, 2005) as a benchmark.

5.1 Experimental Results

The results of using each knowledge source and using all knowledge sources are listed in Table 1. This table also contains the results of the majority baseline (the 3-2 type), the benchmark and the systems based on only the baseline features. Table 2 shows the combination matrix of knowledge features. These results show that the best performing system uses all the baseline and knowledge features. Table 3 lists the coverage for each knowledge source. We can see that our automatically acquired knowledge, in particular related words and case frames, has a wide coverage.

[4]http://www.csie.ntu.edu.tw/˜cjlin/liblinear/

Knowledge	Coverage
All	95.6%
Related words	90.3%
Case frames	84.3%
Related events	7.4%

Table 3: Coverage of knowledge sources.

5.2 Discussions

It is not surprising that case frames and related words worked well, both of which have a high coverage rate. Related events only covered a few sets of text and made a negative contribution if they are used solely, but they made a positive effect if used with other knowledge sources.

(2) a. 魚介も野菜も肉もたっぷりと調理されて皿に盛られている。(Plenty of fish, vegetables, and meat were prepared and piled up on plates.)
b. 男二人とは言え、到底食べきれるとは思えなかった。(Despite being two men, we never thought we could eat it all.)
c. バイキングのようだと思いながら、「いただきます」と手を合わせた手島は、手近にあった皿から取って一口食べる。(Tejima, thinking of it as a buffet, took a bite of food from a nearly plate.)

Example (2)(3-1 type) shows an improved example. The words ' 魚介'(fish), ' 野菜'(vegetables), ' 肉'(meat) in the first sentence are strongly relevant to ' 食べる'(eat) in case frames and ' 一口'(a bite) in related words in the third sentence. ' 調理'(cook/prepared) and ' 食べる'(eat) are also related events.

From the experimental results, we found two main types of error. The first type of error was that values of some word pairs were too high and dominated the results but their meaning was misunderstood especially in the predicate-argument related knowledge sources. The other type of error was that some words function as key words on a certain topic and had many related words.

(3) a. スタッフさんたちの温かい雰囲気で和みます。(I felt relaxed by the staff's warm attitude.)
b. インドエステした日は体がポカポカし心地よいです。(My body felt warm and relaxed on the day of my Indian salon visit.)

c. セルライトがほぐれ血液循環良くなって行く気がして嬉しいです。(I was happy that my cellulite went away and that my blood circulation improved.)

For instance, in example (3)(3-2 type), although ' 血液'(blood), ' 循環'(circulation) are related to ' 体'(body) in related words, ' 雰囲気'(atmoshpere) is related to ' 解れる'(relieve/go away) and ' 良くなる'(improve) strongly and ' 和む' (calm down/relax) and ' 嬉しい'(happy) are related events between the first and third sentences, which made the result incorrect. For the first error, we can use refined strongly-related event knowledge with a predicate and their relevant arguments to limit the word pair recognition. For the second error, we can try to limit the number of word pairs by narrowing the window size and set a threshold for word pairs extracted. In addition, as mentioned in the previous section, different knowledge may have tendencies of affect differently on our discourse structure analyzer. A way to improve accuracy is to incorporate more knowledge sources into our system.

6 Related Work

There are several well-known discourse-annotated corpora, such as the RST Discourse Treebank (Carlson et al., 2001) and the Penn Discourse Treebank (Prasad et al., 2008). These corpora were developed on newspaper articles. Our aim is to establish discourse structure analysis no matter if the target text is well-structured or not. Hence, we are interested in working with usual discourses around us and it is essential to build an annotated corpus that includes diverse-domain documents. We built a Japanese annotated corpus that consists of various genres and each discourse consists of three sentences.

Some work on discourse structure analysis, like (Ghosh et al., 2011), empirically showed that there is a strong correlation between syntax and discourse. In our method, we extracted features from multiple preformed knowledge sources as well as syntactic features. This is a main different point from previous studies. As for statistical approaches to discourse structure analysis, Hernault and his colleagues (Hernault et al., 2010) proposed HILDA, which is a fully-implemented feature-based discourse parser that works at the full text-level rather than individual sentences. (Joty et al., 2013) recently presented a two-stage document-level discourse parser. Their parser significantly outperformed state-of-the-art methods, but they mentioned that they need to take advantage of richer semantic knowledge to further improve their work, which is similar to our idea.

7 Conclusion

We have presented an approach using automatically acquired large-scale knowledge to analyze coherent relations in Japanese 3-sentence discourses. Our experimental results indicated that our proposed knowledge-rich method outperformed baseline systems without knowledge. This result showed that knowledge is useful to discourse structure analysis. Our approach also brings us a inspiration that using a variety of knowledge sources may help improve not only discourse structure analysis but also other analyses, such as parsing and zero anaphora resolution.

References

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2001. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*.

Sucheta Ghosh, Sara Tonelli, Giuseppe Riccardi, and Richard Johansson. 2011. End-to-end discourse parser evaluation. In *Fifth IEEE International Conference on Semantic Computing (ICSC)*.

Masatsugu Hangyo, Daisuke Kawahara, and Sadao Kurohashi. 2012. Building a diverse document leads corpus annotated with semantic relations. In *Proceedings of 26th Pacific Asia Conference on Language Information and Computing*.

Hugo Hernault, Helmut Prendinger, David duVerle, and Mitsuru Ishizuka. 2010. Hilda: A discourse parser using support vector machine classification. *Dialogue & Discourse*.

Shafiq Joty, Giuseppe Carenini, Raymond Ng, and Yashar Mehdad. 2013. Combining intra- and multi-sentential rhetorical parsing for document-level discourse analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*.

Daisuke Kawahara and Sadao Kurohashi. 2006. Case frame compilation from the web using high-performance computing. In *Proceedings of LREC2006*.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn discourse treebank 2.0. In *In Proceedings of LREC2008*.

Ryohei Sasano and Sadao Kurohashi. 2011. A discriminative approach to Japanese zero anaphora resolution with large-scale lexicalized case frames. In *Proceedings of 5th International Joint Conference on Natural Language Processing*.

Tomohide Shibata and Sadao Kurohashi. 2005. Automatic slide generation based on discourse structure analysis. In *Proceedings of Second International Joint Conference on Natural Language Processing*.

Tomohide Shibata and Sadao Kurohashi. 2011. Acquiring strongly-related events using predicate-argument co-occurring statistics and case frames. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*.