Using Phrasal Patterns to Identify Discourse Relations

Manami Saito
Nagaoka University of Technology
Niigata, JP 9402188
saito@nlp.nagaokaut.ac.jp

Kazuhide Yamamoto
Nagaoka University of Technology
Niigata, JP 9402188
yamamoto@fw.ipsj.or.jp

Satoshi Sekine
New York University
New York, NY 10003
sekine@cs.nyu.edu

Abstract

This paper describes a system which identifies discourse relations between two successive sentences in Japanese. On top of the lexical information previously proposed, we used phrasal pattern information. Adding phrasal information improves the system's accuracy 12%, from 53% to 65%.

1 Introduction

Identifying discourse relations is important for many applications, such as text/conversation understanding, single/multi-document summarization and question answering. (Marcu and Echihabi 2002) proposed a method to identify discourse relations between text segments using Naïve Bayes classifiers trained on a huge corpus. They showed that lexical pair information extracted from massive amounts of data can have a major impact.

We developed a system which identifies the discourse relation between two successive sentences in Japanese. On top of the lexical information previously proposed, we added phrasal pattern information. A phrasal pattern includes at least three phrases (bunsetsu segments) from two sentences, where function words are mandatory and content words are optional. For example, if the first sentence is "X should have done Y" and the second sentence is "A did B", then we found it very likely that the discourse relation is CONTRAST (89% in our Japanese corpus).

2 Discourse Relation Definitions

There have been many definitions of discourse relation, for example (Wolf 2005) and (Ichikawa 1987) in Japanese. We basically used Ichikawa's classes and categorized 167 cue phrases in the ChaSen dictionary (IPADIC, Ver.2.7.0), as shown in Table 1. Ambiguous cue phrases were categorized into multiple classes. There are 7 classes, but the OTHER class will be ignored in the following experiment, as its frequency is very small.

Table 1. Discourse relations

Discourse relation	Examples of cue phrase (English translation)	Freq. in corpus [%]
ELABORATION	and, also, then, moreover	43.0
CONTRAST	although, but, while	32.2
CAUSE-EFFECT	because, and so, thus, therefore	12.1
EQUIVALENCE	in fact, alternatively, similarly	6.0
CHANGE-TOPIC	by the way, incidentally, and now, meanwhile, well	5.1
EXAMPLE	for example, for instance	1.5
OTHER	most of all, in general	0.2

3 Identification using Lexical Information

The system has two components; one is to identify the discourse relation using lexical information, described in this section, and the other is to identify it using phrasal patterns, described in the next section.

A pair of words in two consecutive sentences can be a clue to identify the discourse relation of those sentences. For example, the CONTRAST relation may hold between two sentences which

Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 133–136,
New York, June 2006. ©2006 Association for Computational Linguistics

have antonyms, such as “*ideal*” and “*reality*” in Example 1. Also, the EXAMPLE relation may hold when the second sentence has hyponyms of a word in the first sentence. For example, “*gift shop*”, “*department store*”, and “*supermarket*” are hyponyms of “*store*” in Example 2.

Ex1)

a. It is *ideal* that people all over the world accept independence and associate on an equal footing with each other.

b. (However,) *Reality* is not that simple.

Ex2)

a. Every town has many *stores*.

b. (For example,) *Gift shops*, *department stores*, and *supermarkets* are the main stores.

In our experiment, we used a corpus from the Web (about 20G of text) and 38 years of newspapers. We extracted pairs of sentences in which an unambiguous discourse cue phrase appears at the beginning of the second sentence. We extracted about 1,300,000 sentence pairs from the Web and about 150,000 pairs from newspapers. 300 pairs (50 of each discourse relation) were set aside as a test corpus.

3.1 Extracting Word Pairs

Word pairs are extracted from two sentences; i.e. one word from each sentence. In order to reduce noise, the words are restricted to *common nouns*, *verbal nouns*, *verbs*, and *adjectives*. Also, the word pairs are restricted to particular kinds of POS combinations in order to reduce the impact of word pairs which are not expected to be useful in discourse relation identification. We confined the combinations to the pairs involving the same part of speech and those between *verb* and *adjective*, and between *verb* and *verbal noun*.

All of the extracted word pairs are used in base form. In addition, each word is annotated with a positive or negative label. If a phrase segment includes negative words like “not”, the words in the same segment are annotated with a negative label. Otherwise, words are annotated with a positive label. We don’t consider double negatives. In Example 1-b, “simple” is annotated with a negative, as it includes “not” in the same segment.

3.2 Score Calculation

All possible word pairs are extracted from the sentence pairs and the frequencies of pairs are counted for each discourse relation. For a new (test) sentence pair, two types of score are calculated for each discourse relation based on all of the word pairs found in the two sentences. The scores are given by formulas (1) and (2). Here *Freq(dr, wp)* is the frequency of word pair (*wp*) in the discourse relation (*dr*). Score$_1$ is the fraction of the given discourse relation among all the word pairs in the sentences. Score$_2$ incorporates an adjustment based on the rate ($Rate_{DR}$) of the discourse relation in the corpus, i.e. the third column in Table 1. The score actually compares the ratio of a discourse relation in the particular word pairs against the ratio in the entire corpus. It helps the low frequency discourse relations get better scores.

$$Score_1(DR) = \frac{\sum_{wp} Freq(DR, wp)}{\sum_{dr,wp} Freq(dr, wp)} \quad (1)$$

$$Score_2(DR) = \frac{\sum_{wp} Freq(DR, wp)}{\sum_{dr,wp} Freq(dr, wp) \times Rate_{DR}} \quad (2)$$

4 Identification using Phrasal Pattern

We can sometimes identify the discourse relation between two sentences from fragments of the two sentences. For example, the CONTRAST relation is likely to hold between the pair of fragments “*… should have done ….*” and “*… did ….*”, and the EXAMPLE relation is likely to hold between the pair of fragments “*There is…*” and “*Those are … and so on.*”. Here “…” represents any sequence of words. The above examples indicate that the discourse relation between two sentences can be recognized using fragments of the sentences even if there are no clues based on the sort of content words involved in the word pairs. Accumulating such fragments in Japanese, we observe that these fragments actually form a phrasal pattern. A phrase (bunsetsu) in Japanese is a basic component of sentences, and consists of one or more content words and zero or more function words. We specify that a phrasal pattern contain at least three subphrases, with at least one from each sentence. Each subphrase contains the function words of the phrase, and may also include accompanying content words. We describe the method to create patterns in three steps using an example sentence pair (Example 3) which actually has the CONTRAST relation.

Ex3)

a. “kanojo-no kokoro-ni donna omoi-ga at-ta-ka-ha wakara-nai.” (No one knows what feeling she had in her mind.)

b. “sore-ha totemo yuuki-ga iru koto-dat-ta-ni-chigai-nai.” (I think that she must have needed courage.)

1) Deleting unnecessary phrases

Noun modifiers using “no” (a typical particle for a noun modifier) are excised from the sentences, as they are generally not useful to identify a discourse relation. For example, in the compound phrase “kanozyo-no (her) kokoro (mind)” in Example 3, the first phrase (her), which just modifies a noun (mind), is excised. Also, all of the phrases which modify excised phrases, and all but the last phrase in a conjunctive clause are excised.

2) Restricting phrasal pattern

In order to avoid meaningless phrases, we restrict the phrase participants to components matching the following regular expression pattern. Here, *noun-x* means all types of nouns except common nouns, i.e. verbal nouns, proper nouns, pronouns, etc.

“(*noun-x* | *verb* | *adjective*)? (*particle* | *auxiliary verb* | *period*)+$”, or “*adverb*$”

3) Combining phrases and selecting words in a phrase

All possible combinations of phrases including at least one phrase from each sentence and at least three phrases in total are extracted from a pair of sentences in order to build up phrasal patterns. For each phrase which satisfies the regular expression in 2), the subphrases to be used in phrasal patterns are selected based on the following four criteria (A to D). In each criterion, a sample of the result pattern (using all the phrases in Example 3) is expressed in bold face. Note that it is quite difficult to translate those patterns into English as many function words in Japanese are encoded as a position in English. We hope readers understand the procedure intuitively.

A) Use all components in each phrase
kanojo-no kokoro-**ni** donna omoi-**ga at-ta-ka-ha wakara-nai**.
sore-ha totemo yuuki-**ga** iru **koto-dat-ta-ni-chigai-nai**.

B) Remove *verbal noun* and *proper noun*
kanojo-no kokoro-**ni** donna omoi-**ga at-ta-ka-ha wakara-nai**.
sore-ha totemo yuuki-**ga** iru **koto-dat-ta-ni-chigai-nai**.

C) In addition, remove *verb* and *adjective*
kanojo-no kokoro-**ni** donna omoi-**ga** at-**ta-ka-ha** wakara-**nai**.
sore-ha totemo yuuki-**ga** iru **koto-dat-ta-ni-chigai-nai**.

D) In addition, remove *adverb* and remaining *noun*
kanojo-no kokoro-**ni** donna omoi-**ga** at-**ta-ka-ha** wakara-**nai**.
sore-**ha** totemo yuuki-**ga** iru koto-**dat-ta-ni-chigai-nai**.

4.1 Score Calculation

By taking combinations of 3 or more subphrases produced as described above, 348 distinct patterns can be created for the sentences in Example 3; all of them are counted with frequency 1 for the CONTRAST relation. Like the score calculation using lexical information, we count the frequency of patterns for each discourse relation over the entire corpus. Patterns appearing more than 1000 times are not used, as those are found not useful to distinguish discourse relations.

The scores are calculated replacing *Freq(dr, wp)* in formulas (1) and (2) by *Freq(dr, pp)*. Here, *pp* is a phrasal pattern and *Freq(dr, pp)* is the number of times discourse relation *dr* connects sentences for which phrasal pattern *pp* is matched. These scores will be called Score$_3$ and Score$_4$, respectively.

5 Evaluation

The system identifies one of six discourse relations, described in Table 1, for a test sentence pair. Using the 300 sentence pairs set aside earlier (50 of each discourse relation type), we ran two experiments for comparison purposes: one using only lexical information, the other using phrasal patterns as well. In the experiment using only lexical information, the system selects the relation maximizing Score$_2$ (this did better than Score$_1$). In the other, the system chooses a relation as follows: if one relation maximizes both Score$_1$ and Score$_2$, choose that relation; else, if one relation maximizes both Score_3 and Score_4, choose that relation; else choose the relation maximizing Score_2.

Table 2 shows the result. For all discourse relations, the results using phrasal patterns are better or the same. When we consider the frequency of discourse relations, i.e. 43% for ELABORATION, 32% for CONTRAST etc., the weighted accuracy was 53% using only lexical information, which is comparable to the similar experiment by (Marcu and Echihabi 2002) of 49.7%. Using phrasal patterns, the accuracy improves 12% to 65%. Note that the baseline accuracy (by always selecting the most frequent relation) is 43%, so the improvement is significant.

Table 2. The result

Discourse relation	Lexical info. Only	With phrasal pattern
ELABORATION	44% (22/50)	52% (26/50)
CONTRAST	62% (31/50)	86% (43/50)
CAUSE-EFFECT	56% (28/50)	56% (28/50)
EQUIVALENCE	58% (29/50)	58% (29/50)
CHANGE-TOPIC	66% (33/50)	72% (36/50)
EXAMPLE	56% (28/50)	60% (30/50)
Total	57% (171/300)	64% (192/300)
Weighted accuracy	53%	65%

Since they are more frequent in the corpus, ELABORATION and CONTRAST are more likely to be selected by Score_1 or Score_3. But adjusting the influence of rate bias using Score_2 and Score_4, it sometimes identifies the other relations.

The system makes many mistakes, but people also may not be able to identify a discourse relation just using the two sentences if the cue phrase is deleted. We asked three human subjects (two of them are not authors of this paper) to do the same task. The total (un-weighted) accuracies are 63, 54 and 48%, which are about the same or even lower than the system performance. Note that the subjects are allowed to annotate more than one relation (Actually, they did it for 3% to 30% of the data). If the correct relation is included among their *N* choices, then *1/N* is credited to the accuracy count. We measured inter annotator agreements. The average of the inter-annotator agreements is 69%. We also measured the system performance on the data where all three subjects identified the correct relation, or two of them identified the correct relation and so on (Table 3). We can see the correlation between the number of subjects who answered correctly and the system accuracy. In short, we can observe from the result and the analyses that the system works as well as a human does under the condition that only two sentences can be read.

Table 3. Accuracy for different agreements

# of subjects correct	3	2	1	0
System accuracy	71%	63%	60%	47%

.

6 Conclusion

In this paper, we proposed a system which identifies discourse relations between two successive sentences in Japanese. On top of the lexical information previously proposed, we used phrasal pattern information. Using phrasal information improves accuracy 12%, from 53% to 65%. The accuracy is comparable to human performance. There are many future directions, which include 1) applying other machine learning methods, 2) analyzing discourse relation categorization strategy, and 3) including a longer context beyond two sentences.

Acknowledgements

This research was partially supported by the National Science Foundation under Grant IIS-00325657. This paper does not necessarily reflect the position of the U.S. Government. We would like to thank Prof. Ralph Grishman, New York University, who provided useful suggestions and discussions.

References

Daniel Marcu and Abdessamad Echihabi. 2002. An Unsupervised Approach to Recognizing Discourse Relations, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics,* 368-375.

Florian Wolf and Edward Gibson. 2005. Representing Discourse Coherence: A Corpus-Based Study, *Computational Linguistics*, 31(2):249-287.

Takashi Ichikawa. 1978. Syntactic Overview for Japanese Education, Kyo-iku publishing, 65-67 (in Japanese).