Detection of Quotations and Inserted Clauses and its Application to Dependency Structure Analysis in Spontaneous Japanese

Ryoji Hamabe†, Kiyotaka Uchimoto‡, Tatsuya Kawahara†, Hitoshi Isahara‡

†Graduate School of Informatics, Kyoto University
Yoshida-honmachi, Sakyo-ku, Kyoto 606-8501, Japan
hamabe@ar.media.kyoto-u.ac.jp
‡National Institute of Information and Communications Technology
3-5, Hikari-dai, Seika-cho, Soraku-gun, Kyoto 619-0289, Japan

10.21437/Interspeech.2006-251

Abstract

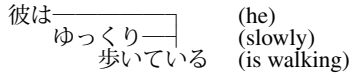
Japanese dependency structure is usually represented by relationships between phrasal units called *bunsetsu*s. One of the biggest problems with dependency structure analysis in spontaneous speech is that clause boundaries are ambiguous. This paper describes a method for detecting the boundaries of quotations and inserted clauses and that for improving the dependency accuracy by applying the detected boundaries to dependency structure analysis. The quotations and inserted clauses are determined by using an SVM-based text chunking method that considers information on morphemes, pauses, fillers, etc. The information on automatically analyzed dependency structure is also used to detect the beginning of the clauses. Our evaluation experiment using *Corpus of Spontaneous Japanese (CSJ)* showed that the automatically estimated boundaries of quotations and inserted clauses helped to improve the accuracy of dependency structure analysis.

Index Terms: spontaneous Japanese, dependency structure analysis, clause boundary, quotation, inserted clause, support vector machine, text chunking

1. Introduction

The "Spontaneous Speech: Corpus and Processing Technology" project sponsored the construction of the *Corpus of Spontaneous Japanese (CSJ)* [1]. The CSJ is the biggest spontaneous speech corpus in the world, consisting of roughly 7M words with the total speech length of 700 hours, and is a collection of monologues such as academic presentations and simulated public speeches. The CSJ includes transcriptions of the speeches as well as audio recordings of them. Approximately one tenth of the speeches in the CSJ were manually annotated with various kinds of information such as morphemes, sentence boundaries, dependency structures, and discourse structures.

In Japanese sentences, word order is rather free, and subjects or objects are often omitted. In Japanese, therefore, the syntactic structure of a sentence is generally represented by the relationships between phrasal units, or *bunsetsu*s in Japanese, based on a dependency grammar, as represented in the Kyoto University text corpus [2]. In the same way, the syntactic structure of a sentence is represented by dependency relationships between *bunsetsu*s in the CSJ. For example, the sentence "彼はゆっくり歩いている" (He is walking slowly) can be divided into three *bunsetsu*s, "彼は, *kare-wa*" (he), "ゆっくり, *yukkuri*" (slowly), and "歩いている, *arui-te-iru*" (is walking). In this sentence, the first and second *bunsetsu*s depend on the third one. The dependency structure is described as follows.



In this paper, we first describe the problems with dependency structure analysis of spontaneous speech. We focus on ambiguous clause boundaries as the biggest problem and present a solution.

2. Problems with Dependency Structure Analysis in Spontaneous Japanese

There are many differences between written text and spontaneous speech, and consequently, problems peculiar to spontaneous speech arise in dependency structure analysis, such as ambiguous clause boundaries, independent *bunsetsu*s, crossed dependencies, self-corrections, and inversions. In this study, we address the problem of ambiguous clause boundaries in dependency structure analysis in spontaneous speech. We treated the other problems in the same way as Shitaoka et al. [3].

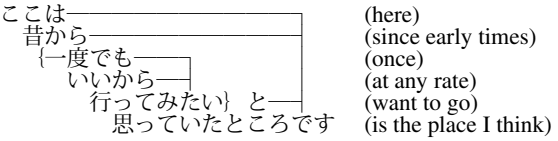
There are several types of clause boundaries such as sentence boundaries, boundaries of quotations and inserted clauses. In the CSJ, clause boundaries were automatically detected by using surface information [4], and sentence boundaries were manually selected from them [5]. Boundaries of quotations and inserted clauses were also defined and detected manually. Dependency relationships between *bunsetsu*s were annotated within sentences. Our definition of clause boundaries follows the definition used in the CSJ.

Shitaoka et al. worked on automatic sentence boundary detection by using SVM-based text chunking. However, quotations and inserted clauses were not considered. In this paper, we focus on these problems in a context of ambiguous clause boundaries.

(1) Quotations

In written text, quotations are often bracketed by "「」" (angle brackets), but no brackets are annotated in spontaneous speech.

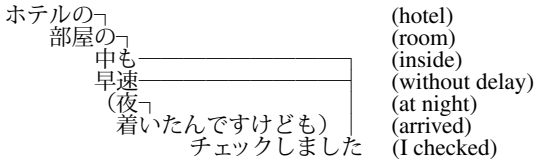
ex) "一度でもいいから行ってみたい" (I want to go there at any rate) is a quotation. In the CSJ, quotations were manually annotated as follows.



(2) Inserted Clauses

In spontaneous speech, speakers insert clauses in the middle of other clauses. This occurs when speakers change their speech plans while producing utterances, which results in supplements, annotations, or paraphrases of main clauses.

ex) "夜着いたんですけども" (where I arrived at night) is an inserted clause.



Dependency relationships are closed within a quotation or an inserted clause. Therefore, dependencies except the rightmost *bunsetsu* in each clause do not cross boundaries of the same clause, meaning no dependency exists between the *bunsetsu* inside a clause and that outside the clause. However, automatically detected dependencies often cross clause boundaries erroneously because sentences including quotations or inserted clauses can have complicated clause structures. This is one of the reasons dependency structure analysis of spontaneous speech has more errors than that of written texts. We propose a method for improving dependency structure analysis based on automatic detection of quotations and inserted clauses.

3. Dependency Structure Analysis and Detection of Quotations and Inserted Clauses

The outline of the proposed processes is shown in Figure 1. Here, we use "clause" to describe a quotation and an inserted clause.

3.1. Dependency Structure Analysis

In this research, we use the method proposed by Uchimoto et al. [6] to analyze dependency structures. This method is a two-step procedure, and the first step is preparation of a dependency matrix in which each element represents the likelihood that one *bunsetsu* depends on another. The second step of the analysis is finding an optimal set of dependencies for the entire sentence. The likelihood of dependency is represented by a probability, using a dependency probability model. The model in this study [6] takes into account not only the relationship between two *bunsetsu*s but also the relationship between the left *bunsetsu* and all the *bunsetsu* to its right.

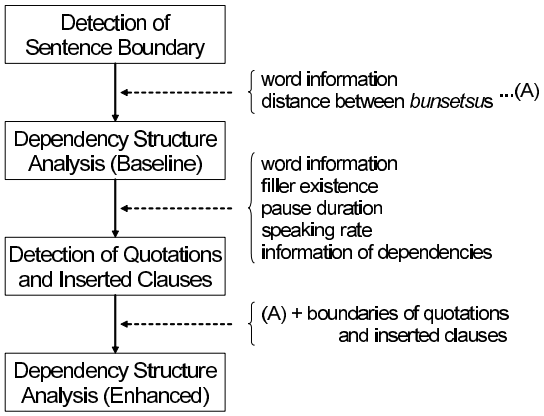


Figure 1: Outline of proposed processes

Table 1: Tag categories used for chunking

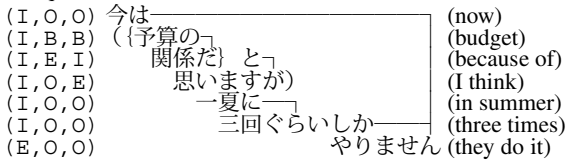
Tag	Explanation of tag
B	Beginning of a clause
E	End of a clause
I	Interior of a clause (except B and E)
O	Exterior of a clause
S	Clause consisting of one *bunsetsu*

We implemented this model within a maximum entropy modeling framework. The features used in the model were basically attributes related to the target two *bunsetsu*s: attributes of a *bunsetsu* itself, such as character strings, parts of speech, and inflection types of a *bunsetsu* together with attributes between *bunsetsu*s, such as the distance between *bunsetsu*s, etc. Combinations of these features were also used. In this work, we added to the features whether there is a boundary of quotations or inserted clauses between the target *bunsetsu*s. If there is, the probability that the left *bunsetsu* depends on the right *bunsetsu* is estimated to be low.

3.2. Detection of Quotations and Inserted Clauses

We regard the problem of clause boundary detection as a text chunking task. We used YamCha [7] as a text chunker, which is based on Support Vector Machine (SVM). We used the chunk labels consisting of three tags which correspond to sentence boundaries, boundaries of quotations, and boundaries of inserted clauses, respectively. The tag for sentence boundaries can be either E (the rightmost *bunsetsu* in a sentence) or I (the others). The tags for the boundaries of quotations and inserted clauses are shown in Table 1. An example of chunk labels assigned to each *bunsetsu* in a sentence is as follows.

ex) "予算の関係だ" (It is because of the budget) is a quotation, and "予算の関係だと思いますが" (which I think is because of the budget) is an inserted clause. For a chunk label, for example, the *bunsetsu* that the chunk label (I, B, B) is assigned to means that it is not related to a sentence boundary but is related to the beginning of a quotation and an inserted clause.



We used the following parameters for YamCha.

- Degree of polynomial kernel: 3rd
- Analysis direction: Right to left
- Dynamic features: Following three chunk labels
- Multi-class method: Pairwise

The chunk label is estimated for each *bunsetsu*. The features used to estimate the chunk labels are as follows.

(1) word information We used word information such as character strings, pronunciation, part of speech, inflection type, and inflection form. Specific expressions are often used at the ends of quotations and inserted clauses. For instance, "と思う, *to-omou*" (think) and "って言う, *tte-iu*" (say) are used at the ends of quotations. Expressions such as "ですが, *desu-ga*" and "けれども, *keredo-mo*" are used at the ends of inserted clauses.

(2) fillers and pauses Fillers and pauses are often inserted just before or after quotations and inserted clauses. Pause duration is normalized in a talk with its mean and variance.

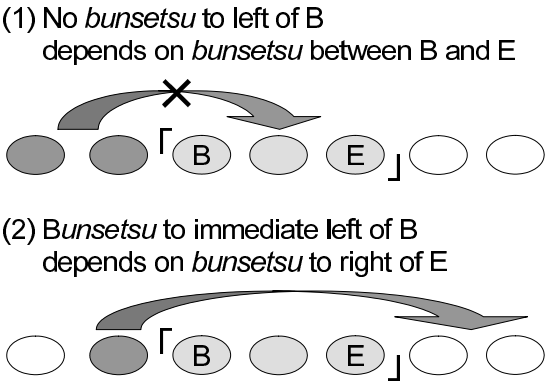


Figure 2: Dependency structures of *bunsetsu*s to left of beginning of quotations or inserted clauses

(3) speaking rate Inside inserted clauses, speakers tend to speak fast. The speaking rate is also normalized in a talk.

Detecting the ends of clauses appears easy because specific expressions are frequently used at the ends of clauses as previously mentioned. However, determining the beginnings of clauses is difficult in a single process because all features mentioned above are local information. Therefore, the global information is also used to detect the beginning of the clauses. If the end of a clause is given, the *bunsetsu*s to the left of the clause should satisfy the two conditions described in Figure 2. Our method uses the constraint as global information. They are considered as additional features based on dependency probabilities estimated for the *bunsetsu*s to the left of the clause. Thus, our chunking method has two steps. First, clause boundaries are detected based on the three types of features itemized above. Second, the beginnings of clauses are determined after adding to the features the following probabilities obtained by automatic dependency structure analysis.

(4) probability that *bunsetsu* to left of target depends on *bunsetsu* inside clause
(5) probability that *bunsetsu* to immediate left of target depends on *bunsetsu* to right of clause

Figure 2 shows that the target *bunsetsu* is likely to be the beginning of the clause if probability (4) is low and probability (5) is high. For instance, the following example sentence has an inserted clause. In the first chunking step, the *bunsetsu* "話なんですけど" (which is a story) is found to be the end of the inserted clause.

ex) "父から聞いた話なんですけど" (which is a story that I heard from my father) is an inserted clause.

この	(this)
辺りは	(area)
（父から	(from my father)
聞いた	(heard)
話なんですけど）	(story)
昔	(in the old days)
たんぼだったんです	(was a rice field)

The three *bunsetsu*s "辺りは, *atari-wa*", "聞いた, *kii-ta*", and "話なんですけど, *hanashi-na-ndesu-kedo*" are less likely to be the beginning of the inserted clause because in the three cases the *bunsetsu* to the immediate left depends on the target *bunsetsu*. On the other hand, the *bunsetsu* "父から, *chichi-kara*" is the most likely to be the beginning since the *bunsetsu* to its immediate left "辺りは, *atari-wa*" depends on the *bunsetsu* to the right of the inserted clause "田んぼだったんです, *tanbo-datta-ndesu*".

4. Experiments and Discussion

For experimental evaluation, we used the transcriptions of 188 talks in the CSJ, which contain 6,255 quotations and 818 inserted clauses. We used 20 talks for testing. For training, we used 168 talks excluding the test data to conduct the open test and all the 188 talks to conduct the closed test.

First, we detected sentence boundaries by using the method [3] and analyzed the dependency structure of each sentence by the method described in Section 3.1 without using information on quotations and inserted clauses. We obtained an F-measure of 85.9 for the sentence boundary detection, and the baseline accuracy of the dependency structure analysis was 77.7% for the open test and 86.5% for the closed test.

(a) Results of clause boundary detection

The results obtained by the method described in Section 3.2 are shown in Table 2. The table shows five kinds of results:

- results obtained without dependency structure (in the first chunking step)
- results obtained with dependency structure analyzed for the open test (in the second chunking step)
- results obtained with dependency structure analyzed for the closed test (in the second chunking step)
- results obtained with manually annotated dependency structure (in the second chunking step)
- the rate that the ends of clauses are detected correctly

These results indicate that around 90% of quotations were detected correctly, and the boundary detection accuracy of quotations was improved by using automatically analyzed dependency structure. We found that features (4) and (5) in Section 3.2 obtained from automatically analyzed dependency structure contributed to the improvement. In the following example, a part of the quotation "自分のいい長所じゃないか" (my good virtue) was erroneously detected as a quotation in the first chunking step. But, in the second chunking step, automatically analyzed dependency structure contributed to detection of the correct part "これは自分のいい長所じゃないか" (this is my good virtue) as a quotation.

{これは	(this)
自分の	(my)
いい	(good)
長所じゃないか} と	(virtue)
私は	(I)
思います	(think)

We also found that the boundary detection accuracy of quotations was significantly improved by using manually annotated dependency structure. This indicates that the boundary detection accuracy of quotations improves as the accuracy of dependency structure analysis improves.

By contrast, only a few inserted clauses were detected even if dependency structures were used. Most of the ends of the inserted clauses were detected incorrectly as sentence boundaries. The main reason for this is our method could not distinguish between the ends of the inserted clauses and those of the sentences, since the same words often appeared at the ends of both, and it was difficult to learn the difference between them even though our method used features based on acoustic information.

(b) Dependency structure analysis results

We investigated the accuracies of dependency structure analysis obtained when detected boundaries of quotations and inserted clauses were used. The results are shown in Table 3. Although the accuracy of detecting the boundaries of quotations and inserted

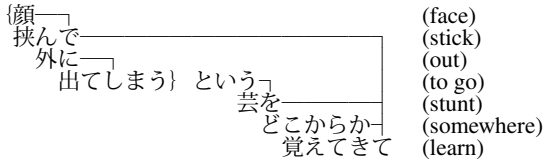
Table 2: Clause boundary detection results (sentence boundaries automatically detected)

	Quotations			Inserted clauses		
	recall	precision	F	recall	precision	F
Without dependency information	41.1% (264/643)	44.3% (264/596)	42.6	1.3% (1/76)	20.0% (1/5)	2.5
With dependency information (open)	42.1% (271/643)	45.5% (271/596)	43.7	2.6% (2/76)	40.0% (2/5)	4.9
With dependency information (closed)	50.9% (327/643)	54.9% (327/596)	52.8	2.6% (2/76)	40.0% (2/5)	4.9
With dependency information (correct)	74.2% (477/643)	80.0% (477/596)	77.0	2.6% (2/76)	33.3% (2/6)	4.9
Correct end of clauses	89.1% (573/643)	96.1% (573/596)	92.5	2.6% (2/76)	40.0% (2/5)	4.9

Table 3: Dependency structure analysis results obtained with clause boundaries (sentence boundaries automatically detected)

	open	closed
Without boundaries of quotations and inserted clauses	77.7%	86.5%
With boundaries of clauses (automatically detected)	78.5%	86.6%
With boundaries of clauses (correct)	79.4%	87.4%

clauses using automatically analyzed dependency structure was not high, the accuracy of dependency structure analysis was improved by 0.7% absolute for the open test. This shows that the model for dependency structure analysis could robustly learn useful information on clause boundaries even if errors were included in the results of clause boundary detection. In the following example, for instance, "顔挟んで外に出てしまう" (to go out with its face stuck) was correctly detected as a quotation in the first chunking step. Then, the initial inappropriate modifiee "覚えてきて, *oboe-te-ki-te*" (learn) of the *bunsetsu* inside the quotation "挟んで, *hasan-de*" (stick) was correctly modified to the *bunsetsu* inside the quotation "出てしまうという, *de-te-shimau-to-iu*" (to go) by using the automatically detected boundary of the quotation.



(c) Results obtained when correct sentence boundaries are given

We investigated the clause boundary detection accuracy of quotations and inserted clauses and the dependency accuracy when correct sentence boundaries were given manually. The results are shown in Tables 4 and 5, respectively.

When correct sentence boundaries were given, the accuracy of clause detection and dependency structure analysis was improved significantly. Table 4 shows that the boundary detection accuracy of inserted clauses as well as that of quotations was significantly improved by using information of dependencies. Table 5 indicates that when using automatically detected clause boundaries, the accuracy of dependency structure analysis was improved by 0.7% for the open test, and it was further improved by using correct clause boundaries.

These experimental results show that detecting the boundaries of quotations and inserted clauses as well as sentence boundaries is sensitive to the accuracy of dependency structure analysis and the improvements of the boundary detection of quotations and inserted clauses contribute to improvement of dependency structure analysis.

Table 4: Clause boundary detection results (sentence boundaries given)

	Quotations			Inserted clauses		
	recall	precision	F	recall	precision	F
Without dependency information	46.0% (296/643)	50.8% (296/583)	48.3	22.4% (17/76)	23.6% (17/72)	23.0
With dependency information (open)	46.7% (300/643)	53.3% (300/563)	49.8	30.3% (23/76)	38.3% (23/60)	33.8
With dependency information (closed)	55.1% (354/643)	62.9% (354/563)	58.7	30.3% (23/76)	39.0% (23/59)	34.1
With dependency information (correct)	75.3% (484/643)	86.0% (484/563)	80.3	46.1% (35/76)	60.3% (35/58)	52.2
Correct end of clauses	86.5% (556/643)	95.4% (556/583)	90.7	64.5% (49/76)	68.1% (49/72)	66.2

Table 5: Dependency structure analysis results obtained with clause boundaries (sentence boundaries given)

	open	closed
Without boundaries of quotations and inserted clauses	81.0%	90.3%
With boundaries of clauses (automatically detected)	81.7%	90.3%
With boundaries of clauses (correct)	82.8%	91.3%

5. Conclusion

This paper described the method for detecting the boundaries of quotations and inserted clauses and that for applying it to dependency structure analysis. The experiment results showed that the automatically estimated boundaries of quotations and inserted clauses contributed to improvement of dependency structure analysis. In the future, we plan to solve the problems found in the experiments and investigate the robustness of our method when the results of automatic speech recognition are given as the inputs. We will also study use of information on quotations and inserted clauses to text formatting, such as text summarization.

6. References

[1] Kikuo Maekawa, Hanae Koiso, Sadaoki Furui, and Hitoshi Isahara. 2000. Spontaneous Speech Corpus of Japanese. In *Proceedings of the LREC2000*, pages 947–952.

[2] Sadao Kurohashi and Makoto Nagao. 1997. Building a Japanese Parsed Corpus while Improving the Parsing System. In *Proceedings of the NLPRS*, pages 451–456.

[3] Kazuya Shitaoka, Kiyotaka Uchimoto, Tatsuya Kawahara, and Hitoshi Isahara. 2004. Dependency Structure Analysis and Sentence Boundary Detection in Spontaneous Japanese. In *Proceedings of the COLING2004*, pages 1107–1113.

[4] Takehiko Maruyama, Hideki Kashioka, Tadashi Kumano, and Hideki. tanaka. 2003. Rules for Automatic Clause Boundary Detection and Their Evaluation. In *Proceedings of the Nineth Annual Meeting of the Association for Natural Language proceeding*, pages 517–520. (in Japanese).

[5] Katsuya Takanashi, Takehiko Maruyama, Kiyotaka Uchimoto, and Hitoshi Isahara. 2003. Identification of "Sentences" in Spontaneous Japanese — Detection and Modification of Clause Boundaries —. In *Proceedings of the ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pages 183–186.

[6] Kiyotaka Uchimoto, Masaki Murata, Satoshi Sekine, and Hitoshi Isahara. 2000. Dependency Model Using Posterior Context. In *Proceedings of the IWPT*, pages 321–322.

[7] Taku Kudo and Yuji Matsumoto. 2001. Chunking with support vector machines. In *Proceedings of the NAACL*.