Dependency Structure Analysis and Sentence Boundary Detection in Spontaneous Japanese

Kazuya Shitaoka[†], *Kiyotaka Uchimoto*[‡], *Tatsuya Kawahara*[†], *Hitoshi Isahara*[‡]

[†]Graduate School of Informatics, Kyoto University
Kyoto 606-8501, Japan
{shitaoka,kawahara}@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
{uchimoto,isahara}@nict.go.jp

Abstract

This paper addresses automatic detection of dependencies between Japanese phrasal units called *bunsetsu*s, and sentence boundaries in a spontaneous speech corpus. In spontaneous speech, the biggest problem with dependency structure analysis is that sentence boundaries are ambiguous. In this paper, we propose two methods for improving the accuracy of sentence boundary detection in spontaneous Japanese: one based on unsupervised learning and the other based on machine learning. Experimental results show that the sentence boundary detection accuracy of 84.85 in F-measure is achieved by using the proposed methods and the accuracy of dependency structure analysis is also improved by using the information on automatically detected sentence boundaries.

1. Introduction

The "Spontaneous Speech: Corpus and Processing Technology" project has been sponsoring the construction of a large spontaneous Japanese speech corpus, *Corpus of Spontaneous Japanese (CSJ)* [1]. The CSJ is the biggest spontaneous speech corpus in the world, and it is a collection of monologues and dialogues. The CSJ includes transcriptions of the speeches as well as audio recordings. A future goal of the project is higher level of linguistic analysis including syntactic and discourse structures. This paper focuses on methods for automatically detecting sentence boundaries and dependency structures in Japanese spoken text.

In many cases, Japanese dependency structures are defined in terms of the dependency relationships between Japanese phrasal units called *bunsetsu*s. To define dependency relationships between all *bunsetsu*s in spontaneous speech, we need to define not only dependency structure in a sentence but also inter-sentential relationships, namely, discourse relationships between sentences. However, it is difficult to define discourse relationships between sentences, and it is enough to define and detect dependency structure within sentences in actual applications.

Almost all previous works on Japanese dependency structure analysis [2, 3, 4, 5, 6] dealt with dependency structures in written text. Although Matsubara et al.[7] dealt with dependency structures in spontaneous speech, the target speech was dialogue where the utterances were short and sentence boundaries could be easily defined based on turn-taking data. In constract, we investigated dependency structures in spontaneous and long speeches in the CSJ. The biggest problem in dependency structure analysis with spontaneous and long speeches is that sentence boundaries are ambiguous.

In this paper, we focus on sentence boundary detection in spontaneous Japanese, and propose two methods for improving the accuracy. We show that the accuracy of dependency structure analysis is improved by using the results of automatic sentence boundary detection. We also show that both accuracies of dependency structure analysis and sentence boundary detection are improved by interactively utilizing the information obtained in each other's analysis and detection process.

2. Dependency Structure Analysis and Sentence Boundary Detection in Spontaneous Japanese

2.1. Problems with Dependency Structure Analysis

As we consider that the biggest problem in dependency structure analysis of spontaneous and long speech is that **sentence boundaries are ambiguous**, we mainly focus on this problem and describe our solution to it. Other problems are followings:

- Independent *bunsetsu*s
- Crossed dependency
- Self-correction
- Inversion

These problems are left for future works.

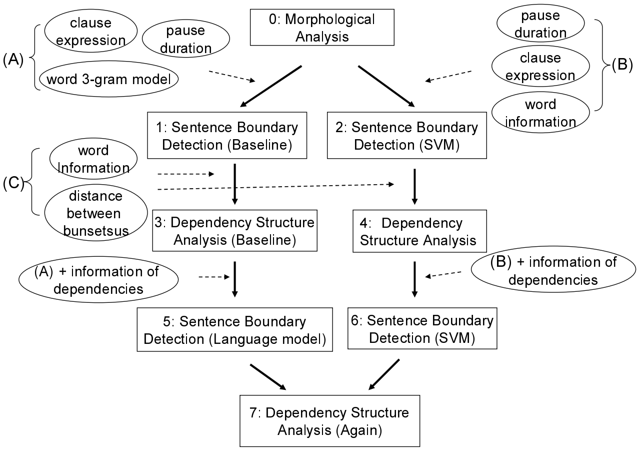
(A)
clause expression
pause duration
word 3-gram model
0: Morphological Analysis
pause duration
clause expression
word information
(B)
1: Sentence Boundary Detection (Baseline)
2: Sentence Boundary Detection (SVM)
(C)
word Information
distance between bunsetsus
3: Dependency Structure Analysis (Baseline)
4: Dependency Structure Analysis
(A) + information of dependencies
(B) + information of dependencies
5: Sentence Boundary Detection (Language model)
6: Sentence Boundary Detection (SVM)
7: Dependency Structure Analysis (Again)

Figure 1: Outline of dependency structure analysis and sentence boundary detection.

2.2. Problems with Sentence Boundary Detection

In the CSJ, sentence boundaries were defined[8] based on clauses whose boundaries are detected using surface information[9]. They were manually tagged. Clause boundaries can be classified into the following three groups.

Absolute boundaries correspond to sentence boundaries in the usual meaning.

Strong boundaries are the points that can be regarded as major breaks in utterances and proper points for segmentation.

Weak boundaries are not regarded as proper points for segmentation because they are strongly dependent on other clauses.

Among these clause boundaries, absolute boundaries and strong boundaries are basically defined as sentence boundaries.

In this paper, we propose two methods for improving the accuracy of sentence boundary detection in spontaneous Japanese and compare them with the conventional method as described in Section 3.2[10]. One is a method combining the conventional method with dependency information. The other is a method based on machine learning.

3. Approach of Dependency Structure Analysis and Sentence Boundary Detection

The outline of the processes is shown in Figure 1.

3.1. Dependency Structure Analysis

The Japanese dependency structure is usually represented by the dependency relationships between phrasal units called *bunsetsu*. In statistical dependency structure analysis of Japanese speech, the likelihood of dependency is represented by a probability estimated by a dependency probability model.

The conventional statistical model [2, 3, 11, 4] learns the relationship between two *bunsetsu*s as either “dependent” or “not dependent”, whereas the model [5] that we employ in this paper learns it as “between”, “dependent”, or “beyond”, and estimates dependency likelihood by considering not only the relationship between two *bunsetsu*s but also the relationship between the left *bunsetsu* and all of the *bunsetsu*s to its right.

We implemented this model within a maximum entropy modeling framework. The features used in the model are basically attributes of *bunsetsu*s, such as character strings, parts of speech, and types of inflection, as well as those that describe the relationships between *bunsetsu*s, such as the distance between *bunsetsu*s. Combinations of these features are also used.

3.2. Sentence Boundary Detection based on Statistical Machine Translation (Conventional method)

The problem of sentence boundary detection can be reduced to the problem of translating a sequence of words X that do not include periods but include pause duration into a sequence of words Y that include periods[10]. We adopt a framework of statistical machine translation for this problem. Specifically, at every position where a pause can be converted into a period, whether a period should be inserted or not is determined by comparing language model scores.

We used a model that uses pause duration and surface expressions around pauses as a translation model $P(X|Y)$. Specifically, a pause following the expressions such as “$H(*to*)”, “$J$$(*nai*)”, and “$?(*ta*)”, and a pause preceding the expressions such as “$G(*de*)”, can be converted when the duration is longer than average. A pause preceding or following the other expressions can be converted even if the duration is short. To calculate $P(Y)$, we use a word 3-gram model trained with transcriptions of the CSJ.

3.3. Sentence Boundary Detection Using Dependency Information (Method 1)

There are three conditions that should be satisfied by the rightmost *bunsetsu* in every sentence, which is refered to as a target *bunsetsu*.

(1) There is no *bunsetsu* **that depend on a** *bunsetsu* **beyond the target** *bunsetsu***.**

Every *bunsetsu* depends on a *bunsetsu* in the same sentence.

(2) One or more *bunsetsu***s depend on the target** *bunsetsu***.**

Since every *bunsetsu* depends on another *bunsetsu* in the same sentence, the second rightmost *bunsetsu* always depends on the rightmost *bunsetsu* in a sentence.

(3) Dependency probability of the target *bunsetsu* is low.

The target *bunsetsu* does not depend on any *bunsetsu*.

*Bunsetsu*s that satisfy all conditions of (1)-(3) are extracted as rightmost *bunsetsu* candidates in a sentence. Then, for every point following the extracted *bunsetsu*s and every pause preceding or following the expressions described in Section 3.2, a decision is made regarding whether a period should be inserted or not. In condition (2), *bunsetsu*s that depend on a *bunsetsu* beyond more than 50 *bunsetsu*s are ignored because no such long-distance dependencies were found in our data used in the experiments. *Bunsetsu*s whose dependency probability is very low are also ignored. Let this threshold probability be p, and let the threshold probability in condition (3) be q. The optimal parameters (p, q) are determined by using held-out data.

3.4. Sentence Boundary Detection based on Machine Learning (Method 2)

As another approach, we introduce a machine learning model of Support Vector Machine (SVM).

We consider the problem of sentence boundary detection as a text chunking task. As a text chunker, we used YamCha [12], which is based on SVM with polynomial kernel functions. To estimate an appropriate chunk label assigned to the current word, YamCha uses information of preceding and subsequent words as baseline features, and it also uses chunk labels that are dynamically assigned to the two preceding or subsequent words as dynamical features.

4. Experiments and Discussion

We used the transcriptions of 178 talks in the CSJ for training and 10 talks for testing. Dependency accuracy is a percentage of correct dependencies out of all dependencies, Sentence boundary detection results are evaluated in terms of F-measure. In Tables 1 to 3, we show the results for "closed" and "open" cases.

The method described in Section 3.2 was used as the baseline method for sentence boundary detection (Process 1 in Figure 1). The method described in Section 3.1 was used as the baseline method for dependency structure analysis (Process 3 in Figure 1).

4.1. Sentence Boundary Detection Results Obtained by Method 1

We evaluated the results obtained by the method described in Section 3.3 (Process 5 in Figure 1).

First, we investigated the optimal values of parameters (p, q) described in Section 3.3 by using held-out data, which differs from the test data and consists of 15 talks. The optimal values of p and q were, respectively, 0 and 0.9 for open-test data, and were 0 and 0.8 for closed-test data. These values were used in the following experiments.

Table 1: Sentence boundary detection results obtained by using dependency information.

	recall	precision	F
With dependency information (open)	74.09% (835/1,127)	82.51% (835/1,012)	78.01
With dependency information (closed)	74.18% (836/1,127)	83.52% (836/1,001)	78.57
baseline	64.51% (727/1,127)	94.17% (727/772)	76.57

The obtained results are shown in Table 1. When dependency information was used, the F-measure increased by approximately 1.4 for the open-test data and by 2.0 for the closed-test data, respectively. We found that errors were caused at noun final clauses, clauses where the rightmost constituents were adjectives or verbs such as " $H;W$&(*to-omou*, think)" or "$OFq$7$$(*wa-muzukashii*, difficult)", and clauses where the rightmost constituents were "$H$$$&$N$O(*to-iu-no-wa*, becuase)" and "$H$7$F $O(*to-si-te-wa*, as)", and so on. Some of these errors, except for those in noun final clauses, could have been correctly detected if we had had more training data.

We also found that periods were sometimes erroneously inserted when preceding expressions were "$, (*ga*, but)", "$^$7$F(*mashite*, and)", and "$1$l$I$b (*keredomo*, but)", which are typically the rightmost constituents of a sentence, and "$F(*te*, and)", which is typically not the rightmost constituent of a sentence. The language models were not good at discriminating between subtle differences.

4.2. Sentence Boundary Detection Results Obtained by Method 2

We evaluated the results obtained by the method described in Section 3.4 (Process 6 in Figure 1). The features used in SVMs are the following:

1. Morphological information of the three preceding and subsequent words such as character strings, pronunciation, part of speech, inflection type, and inflection form
2. Pause duration that is normalized in a talk to a z-score
3. Clause boundary information
4. Dependency probability of the target *bunsetsu*
5. The number of *bunsetsu*s that depend on the target *bunsetsu* and their dependency probabilities

The results are shown in Table 2. The F-measure was about 6.8 points higher than that described in Section 4.1. The results show that the machine learning approach is more effective than the unsupervised learning

Table 2: Sentence boundary detection results obtained by using SVMs

	recall	precision	F
With dependency information (open)	80.04% (902/1,127)	90.29% (902/999)	84.85
With dependency information (closed)	79.86% (900/1,127)	90.54% (900/994)	84.87
Without dependency information	79.33% (894/1,127)	90.12% (894/992)	84.38

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

	open	closed
With results in Section 4.1	75.78%	81.20%
With results in Section 4.2	77.15%	82.53%
Baseline	75.21%	80.74%

approach. The results also show that the accuracy of sentence boundary detection can be increased by using dependency information in the Method 2.

4.3. Dependency Structure Analysis Results

We evaluated the dependency structure analysis obtained when sentence boundaries detected automatically by the two methods described above were used ad inputs (Process 7 in Figure 1). The results are shown in Table 3. The accuracy of dependency structure analysis improved by about 2% when the most accurate and automatically detected sentence boundaries were used as inputs. This is because more sentence boundaries were detected correctly, and the number of *bunsetsu*s that depended on those in other sentences decreased.

We also investigated the accuracy of dependency structure analysis when 100% accurate sentence boundaries were used as inputs. The accuracy was 80.59% for the open-test data, and 86.12% for the closed-test data. The accuracy of dependency structure analysis for spoken text was about 8% lower than that for written text (newspapers). However, the result in Table 3 is much closer to the ideal case.

5. Conclusion

This paper addresses automatic detection of dependencies between *bunsetsu*s and sentence boundaries in a spontaneous speech corpus. We proposed two methods for improving the accuracy of sentence boundary detection in spontaneous Japanese speech. Using these methods, we obtained an F-measure of 84.85 for the accuracy of sentence boundary detection. The accuracy of dependency structure analysis was also improved from 75.21% to 77.15% by using automatically detected sentence boundaries. The accuracies of dependency structure analysis and that of sentence boundary detection were improved by interactively using automatically detected dependency information and sentence boundaries.

6. References

[1] K. Maekawa, H. Koiso, S. Furui, and H. Isahara. Spontaneous Speech Corpus of Japanese. In *Proceedings of the LREC2000*, pp. 947–952, 2000.

[2] M. Fujio and Y. Matsumoto. Japanese Dependency Structure Analysis based on Lexicalized Statistics. In *Proceedings of the EMNLP*, pp. 87–96, 1998.

[3] M. Haruno, S. Shirai, and Y. Ooyama. Using Decision Trees to Construct a Practical Parser. In *Proceedings of the COLING-ACL*, pp. 505–511, 1998.

[4] K. Uchimoto, S. Sekine, and H. Isahara. Japanese Dependency Structure Analysis Based on Maximum Entropy Models. In *Proceedings of the EACL*, pp. 196–203, 1999.

[5] K. Uchimoto, M. Murata, S. Sekine, and H. Isahara. Dependency Model Using Posterior Context. In *Proceedings of the IWPT*, pp. 321–322, 2000.

[6] T. Kudo and Y. Matsumoto. Japanese Dependency Structure Analysis Based on Support Vector Machines. In *Proceedings of the EMLNP*, pp. 18–25, 2000.

[7] S. Matsubara, T. Murase, N. Kawaguchi, and Y. Inagaki. Stochastic Dependency Parsing of Spontaneous Japanese Spoken Language. In *Proceedings of the COLING2002*, pp. 640–645, 2002.

[8] K. Takanashi, T. Maruyama, K. Uchimoto, and H. Isahara. 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*, pp. 183–186, 2003.

[9] T. Maruyama, H. Kashioka, H. Kumano, and H. Tanaka. Rules for Automatic Clause Boundary Detection and Their Evaluation. In *Proceedings of the Nineth Annual Meeting of the Association for Natural Language Proceeding*, pp. 517–520, 2003. (in Japanese).

[10] K. Shitaoka, T. Kawahara, and H. G. Okuno. Automatic Transformation of Lecture Transcription into Document Style using Statistical Framework. In *IPSJ–WGSLP SLP-41-3*, pp. 17–24, 2002. (in Japanese).

[11] M. Collins. A New Statistical Parser Based on Bigram Lexical Dependencies. In *Proceedings of the ACL*, pp. 184–191, 1996.

[12] T. Kudo and Y. Matsumoto. Chunking with support vector machines. In *Proceedings of the NAACL*, 2001.