Dependency Model Using Posterior Context

Kiyotaka Uchimoto[†] **Masaki Murata**[†]
Satoshi Sekine[‡] **Hitoshi Isahara**[†]

[†]Communications Research Laboratory
588-2, Iwaoka, Iwaoka-cho, Nishi-ku
Kobe, Hyogo, 651-2492, Japan
{uchimoto,murata,isahara}@crl.go.jp

[‡]New York University
715 Broadway, 7th floor
New York, NY 10003, USA
sekine@cs.nyu.edu

Abstract

We describe a new model for dependency structure analysis. This model learns the relationship between two phrasal units called bunsetsus as three categories; 'between', 'dependent', and 'beyond', and estimates the dependency likelihood by considering not only the relationship between two bunsetsus but also the relationship between the left bunsetsu and all of the bunsetsus to its right. We implemented this model based on the maximum entropy model. When using the Kyoto University corpus, the dependency accuracy of our model was 88%, which is about 1% higher than that of the conventional model using exactly the same features.

1 Introduction

Dependency structure analysis is one of the basic techniques used in Japanese sentence analysis. The Japanese dependency structure is usually represented by the relationships between phrasal units called 'bunsetsu.' The analysis is done in two steps. In the first step, a dependency matrix is prepared where each element represents the likelihood of one bunsetsu being dependent on another in a sentence. In the second step, an optimal set of dependencies for the entire sentence is found. In this paper we only discuss the first step, a model for estimating the likelihood of dependency.

In our approach, the value for each element in the dependency matrix is estimated as a probability. We previously developed a statistical model that considers only the relationship between two bunsetsus when estimating the dependency likelihood[1]. We call this model the old model in this work. Here we describe a model that considers not only the relationship between two bunsetsus, but also the relationship between the left bunsetsu and all of the bunsetsus to its right in a sentence. The probability of whole sentence dependencies is calculated as the product of all the dependency probabilities in a sentence. By searching for the dependencies that maximize the probability, we can identify the optimal dependencies in a sentence. The dependencies in a sentence are identified by analyzing it from right to left[2].

2 Dependency Model Using Posterior Context

Given a tokenization of a test corpus, the problem of dependency structure analysis in Japanese can be reduced to the problem of assigning one of two tags to each relationship between two bunsetsus. A relationship can be tagged with a '1' or a '0' to indicate whether or not there is a dependency between the bunsetsus, respectively. Assigning these tags is the usual way to describe a dependency relationship [3, 4, 1]. However, there are two other possibilities when there is not a dependency between two bunsetsus. One is the case where an anterior bunsetsu depends on one between it and the posterior bunsetsu. The other case is where an anterior bunsetsu depends on a bunsetsu beyond the posterior one. We believe there is a big difference between the two cases.

We developed a dependency model to identify this difference. A dependency relationship between two bunsetsus is tagged with a '0,' '1,' or '2' to indicate the three cases, respectively. The anterior bunsetsu can depend on (1) a bunsetsu between the two, (2) the posterior bunsetsu, or (3) a bunsetsu beyond the posterior one. Our new model uses these three categories while the old model uses only two. The dependency probability of two bunsetsus is estimated by using the product of the probabilities of the relationship between the left bunsetsu and those to its right in a sentence.

We show how to estimate dependency probability with this model using the example in Fig. 1. Figure. 1 shows a simulated calculation of the dependency probabilities of a bunsetsu that has five bunsetsus to its right and is represented by the left most circle. The probabilities of the relationship between this bunsetsu and each modifiee candidate are shown in the table in Fig. 1. The dependencies of the five bunsetsus on the right are assumed to have been identified. Each dependency is represented by an arrow with a dotted line. In this example, bunsetsu 3 and 4 cannot be modified by the current bunsetsu because we assume that "dependencies do not cross." For example, the dependency probability between the current bunsetsu and bunsetsu 5 is calculated as shown in the bottom example in Fig. 1. It has a normalized probability of 52.2%.

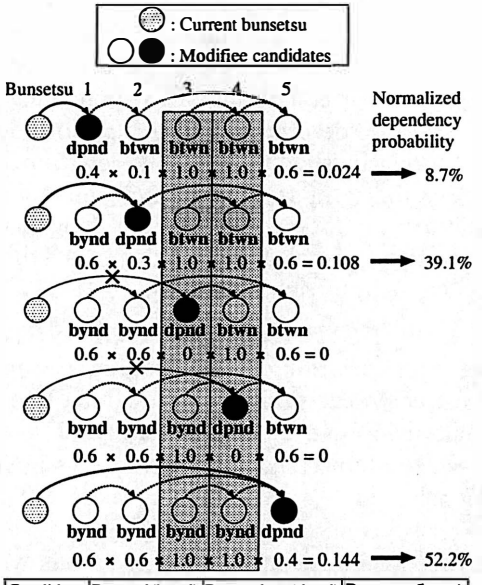
3 Experimental Result

We implemented our model based on the maximum entropy model. We used in our experiments the same features as in Ref. [1]. Those features were basically some attributes of a bunsetsu itself or those between bunsetsus. We used the Kyoto University text corpus (Version 2) [5], a tagged corpus of the Mainichi newspaper. For training we used 7,958 sentences from newspaper articles appearing from January 1st to January 8th in 1995, and for testing we used 1,246 sentences from articles appearing on January 9th. We assumed that the input sentences were morphologically analyzed and their bunsetsus were identified correctly.

The results of our experiment are shown in Table 1. The dependency accuracy means the percentage of correct dependencies out of the total analyzed. The sentence accuracy means the percentage of sentences in which all the dependencies were analyzed correctly. The first and the second lines in Table 1 compare the accuracy of our new model and the old model. The bottom line in Table 1 shows the accuracy when we assumed that every bunsetsu depended on the next one. The dependency accuracy of the new model was about 1% better than that of the old model and there was a 3% improvement in sentence accuracy. Our Investigation of the relationships between sentence length (number of bunsetsus) and dependency accuracy, and between the amount of training data (number of sentences) and the accuracy of the model found that the accuracy of the new model was almost always better than that of the old one for any sentence length. And we found that the accuracy of the new model was about 1% higher than that of the old model for any size of training data used.

Table 1: Results of dependency analysis.

Model	Dependency accuracy	Sentence accuracy
New	87.93% (9,904/11,263)	43.58%(540/1,239)
Old	87.02% (9,801/11,263)	40.68%(504/1,239)
Baseline	64.09% (7,219/11,263)	6.38% (79/1,239)



Candidate	Beyond(bynd)	Dependent(dpnd)	Between(btwn)
1	0.6	0.4	0
2	0.6	0.3	0.1
3	0.3	0.5	0.2
4	0.1	0.5	0.4
5	0	0.4	0.6

Figure 1: Simulated calculation of dependency probability.

References

[1] Kiyotaka Uchimoto, Satoshi Sekine, and Hitoshi Isahara. Japanese dependency structure analysis based on maximum entropy models. In *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 196–203, 1999.

[2] Satoshi Sekine, Kiyotaka Uchimoto, and Hitoshi Isahara. Statistical dependency analysis using backward beam search. *Journal of Natural Language Processing*, 6(3):59–73, 1999. (in Japanese).

[3] Masahiko Haruno, Satoshi Shirai, and Yoshifumi Ooyama. Using decision trees to construct a practical parser. *Proceedings of the COLING-ACL '98*, 1998.

[4] Masakazu Fujio and Yuji Matsumoto. Japanese dependency structure analysis based on lexicalized statistics. *Proceedings of Third Conference on Empirical Methods in Natural Language Processing*, pages 87–96, 1998.

[5] Sadao Kurohashi and Makoto Nagao. Kyoto university text corpus project. In *3rd Annual Meeting of the Association for Natural Language Processing*, pages 115–118, 1997. (in Japanese).