Deterministic Japanese Word Segmentation by Decision List Method

Hiroyuki Shinnou

Ibaraki University, 4-12-1 Nakanarusawa Hitachi Ibaraki 316-8511, Japan,
shinnou@dse.ibaraki.ac.jp

In Japanese natural language processing, morphological analysis is a very important technique, and many methods for it have been proposed. The task of Japanese morphological analysis is essentially word segmentation. In this study, we propose a new method of Japanese word segmentation. Our method regards word segmentation as the classification problem and solves it by the decision list method. The advantage of our method is that it avoids the unknown word problem because it is a kind of character based method. Another advantage is that it is deterministic, and the time taken for deterministic analysis is proportional to the length of the sentence. Moreover, our approach can use various features to solve the classification problem, and various machine learning methods.

The biggest problem of Japanese word segmentation is to cope with unknown words. To overcome this problem, character-based Hidden Markov Model (HMM) has been proposed. In HMM, the state transfer probabilities and the output symbol probabilities judge whether a word boundary exists between two characters or not. These probabilities are learned from bi-gram and tri-gram of the training corpus. However, character-based HMM needs further resources because n-gram alone cannot give high precision. In this study, we regard word segmentation as the problem judging whether a word boundary exists between two characters or not, that is, the classification problem. Therefore, we can conduct word segmentation by various machine learning methods for the classification problem. Moreover, we can use various resources besides n-gram as features to achieve the high precision.

In this study, we used the decision list method[1] as the machine learning method. Moreover, we used the kinds of Japanese character and the information of parentheses besides n-gram resources as features of the decision list.

In experiments, we constructed the decision list by using training data of one year of newspaper articles. With the constructed decision list, we conducted word segmentation for 1,000 sentences and compared our results with those of Chasen, which is the de facto standard system of Japanese morphological analysis. As a result, our method was slightly superior to the Chasen system.

References

1. Yarowsky, D. : "Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French", 32th Annual Meeting of the Association for Computational Linguistics, pp. 88–95 (1994).

R. Mizoguchi and J. Slaney (Eds.): PRICAI 2000, LNAI 1886, p. 822, 2000.