Title	SVMと３つ組／４つ組法による日本語係り受け解析に関する研究
Author(s)	石村，健二
Citation	
Issue Date	2003-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/1666
Rights	
Description	Supervisor:鳥澤 健太郎，情報科学研究科，修士

Japanese Dependency Structure Analysis Based on Support Vector Machine and triplet/quadruplet model

Kenji Ishimura (110011)

School of Information Science,
Japan Advanced Institute of Science and Technology

February 14, 2003

Keywords: Japanese Dependency Analysis, triplet/quadruplet model, Support Vector Machine, HPSG.

This paper describes a statistical method for Japanese dependency analysis. We try to introduces Support Vector Machine (SVM) [1] into the triplet/quadruplet model [2], which is a statistical model for Japanese dependency analysis, and aims at improving accuracy of dependency analysis. Recent progress in grammar formalisms enables us to develop wide-coverage grammars such as SLUNG [3] that can produce parse trees for most sentences in a wide range of texts, including newspaper and magazine articles. But these grammars likely to have high ambiguities, and just enumerate possible sentence structures in parsing. There have been many attempts to develop a postprocessing module that can select a preferable parse tree from many tress that the grammars have enumerated. Kanayama's triplet/quadruplet model is a statistical model and can be seen as one of such postprocessing modules. His method could selects a parse tree from many parse trees generated by an HPSG-based grammar called SLUNG with a high precision.

A problem in Kanayama's method is that it requires more features than other similar techniques. And improving the accuracy of the triplet/quadruplet model requires a effective smoothing technique. We try to introduce Support Vector Machine (SVM) as such smoothing technique, and aims at further improvement of parsing accuracy. SVM is one of the best smoothing techniques currently available, and many researchers have reported that it works better than the maximum entropy method (ME method), which Kanayams's original triplet/quadruplet method adopted.

The triplet/quadruplet model is one of the methods for dependency structure analysis, which have been recognized as a basic technique in Japanese sentence analysis. Japanese dependency structure is usually defined in terms of the relationship between phrasal units called ‘ bunsetsu ’ segments. Generally, dependency structure analysis consists of two step. At the first step, dependency matrix is constructed, in which each element corresponds to a pair of ‘ bunsetsu ’and represents the probability of a dependency relation between them. The second step is to find the optimal combination of dependencies to form the entire sentence.

The difference between the triplet/quadruplet model and other conventional models is the way of calculating statistical values. The conventional models calculate the probability of a correct dependency between two bunsetsus (phrasal units of Japanese) for each pair of bunsetsus. On the other hand, in the triplet/quadruplet model, the conditional part of the probability consists of information on a modifier bunsetsu and all its modification candidates. This enables the model to capture context dependency to a certain degree. The number of candidates is restricted to three or less by using SLUNG and a heuristic. According to Kanayama, 98.6% of all the correct bunsetsu dependencies exists between a modifier bunsetsu and “ modification candidate bunsetsu nearest to modifier bunsetsu ”, “ modification candidate bunsetu which is secondly near to modifier bunsetsu ” and “ modification candidate bunsetsu which is the farthest from modifier bunsetsu. ” The heuristic rule that restricts the modification candidates to at most three was made on the basis of this observation.

In Kanayama's original model the probability of a correct dependency was estimated by using the ME method This triplet/quadruplet model achieves high accuracy : 88.6% for the analysis of the EDR annotated corpus.

Support Vector Machine(SVM) is a statistical learning technique that was developed recently. This technique take a strategy that maximize the margin between critical examples and the separating hyper-plane for avoiding overfitting. As a result, SVM can enjoy high generalization ability even with training data of a very high dimension. In addition, by optimizing a function called the Kernel function, SVM can deal with non-linear feature spaces, and carry out the training with considering combinations of more than on feature. Kudo et al. has already developed a parsing method that utilizes SBM and reported that their system achieves the accuracy of 89.09% even with small training data(7958 sentences). in experiments on parsing of Kyoto University corpus

The goal of this study is introducing SVM that has desirable properties mentioned above into the triplet/quadruplet model. One major obstacle to achieve this goal is that SVM is binary classifier, while triplet/quadruplet model requires three value classifier. The triplet/quadruplet model must choose one from three values " modification candidate bunsetsu nearest to modifier bunsetsu ", " modification candidate bunsetu which is secondly near to modifier bunsetsu " and " modification candidate bunsetsu which is the farthest from modifier bunsetsu ". Thus, Existing SVM must be extended to multi-value classifier. We solved this problem by applying one vs. rest method. More precisely, we prepared three SVM and combine them to obtain three valued classifiers.

Another important point is that Kanayama's original triplet/quadruplet model could not use all the words appeared in the training data as features since introducing such many features caused over-fitting and the parsing performance was reduced. On the other hand, since SVM outperforms the ME in generalization ability in many situations, we can expect that introducing words to our model as features causes improvement of parsing performance. We conducted the experiments of parsing with the model to which words were introduced as features given to SVM.

We implanted our method by using TinySVM, which is an implementation of SVM developed by Kudo [4], and LiLFeS, which is a programming language particularly developed for describing parsing and grammars in the framework of HPSGs and was developed by Makino et al., [5]. We also used the grammar SLUNG developed by Mitsuishi et al. [3]

Finally, parsing experiments using the statistics model described above were conducted. We used EDR annotated corpus as learning data and test data. As features given to SVM, we basically used the feature set that

Kanayama used. But we eliminated combinations of more than one feature from the feature set that Kanayama used. In addition, as mentioned above, all the words appeared in training data were also added to our feature set. As training data, we used about 10^5 sentences in the EDR corpus, while test data consisted of 2,603 sentences. The achieved maximum accuracy was 87.7%, which was not better than Kaneyama's original triplet/quadruplet model unfortunately. One possible reason is that we could use only a half of the training data Kanayama used as our training data. He used about 192,778 sentences in the training. The learning process requried too long time when the same amount of the training data that Kanayama used. In addition, our parsing performance is considerably lower than Kudo's results although both frameworks uses SVM. A hypothesis that may explain this phenomenon is that the feature space in the triplet/quadruplet model is far more complex than that of Kudo's method and finer tuning of feature sets and kernel functions may be necessary.

References

[1] C.Cortes and V.Vapnik. Support Vector Networks. *Machine Learning, Vol.20, pp.273-297, 1995.*

[2] 金山博，鳥澤健太郎，光石豊，辻井潤一．３つ以下の候補からかかり先を選択する係り受け解析モデル．自然言語処理，*Vol.5 No.5 pp71-91. Nov,2000.*

[3] Mitsuishi, Y., Torisawa, K., and Tsujii, J. (1998). “ HPSG-Style Underspecified Japanese Grammar with Wide Coverage. ”In Proc. COLING-ACL ’98, pp.876-880.

[4] 工藤拓，松本裕治．*Support Vector Machine* による日本語係り受け解析．*SIG-NL-128. 2000.*

[5] Takaki Makino, Kentaro Torisawa, and Jun-ichi Tsujii. LiLFeS - practical programming language for typed feature structures. In Proceedings of the 4th Natural Language Processing Pacific Rim Symposium, pp.239-244, Phuket, Thailand, 1997