

Noun Phrase Recognition by System Combination

Erik F. Tjong Kim Sang
CNTS - Language Technology Group
UIA - University of Antwerp
Universiteitsplein 1, 2610 Antwerpen, Belgium
erikt@uia.ua.ac.be

1 Introduction

The performance of machine learning algorithms can be improved by combining the output of different systems. In the domain of natural language learning, this idea has been applied to word class tagging [4]. In this paper we will use system combination for recognizing groups of words of one syntactical category: noun phrases. We will apply a single memory-based learning technique to data that has been represented in different ways. We compare various combination techniques on a part of a standard corpus for English, the Penn Treebank, and use the best method on three standard noun phrase data sets.

2 Experiments and Results

A noun phrase (NP) is a consecutive group of words which can be replaced by a single noun without influences on the syntactical correctness of its context. The example sentence ‘In (early trading) in (Hong Kong) (Monday) , (gold) was quoted at ((\$ 366.50) (an ounce)) .’ contains seven NPs (the phrases between brackets). NP recognition can be divided in two tasks: finding so-called baseNPs, NPs which do not contain other NPs, and identifying arbitrary NPs.

In our example sentence, baseNPs have been represented with bracket structures. Alternatively, the baseNP structure of a sentence can be represented by a list of word-related chunk tags. Ramshaw and Marcus [2] have suggested to use an I for words inside baseNPs, a B for baseNP words immediately after another baseNP and an O for words outside of the baseNPs. The chunk tag list for our example sentence is O I I O I I B O I O O O I I B I O. We have used this chunk tag representation together with four tag variants defined by Tjong Kim Sang and Veenstra [3] to build five NP models with the memory-based classifier IB1IG [1].

Like Van Halteren et.al [4], we have compared nine methods for combining the output of the five NP models. The most simple one examined the chunk tags generated for a particular word and selected the most frequent tag. Four more advanced techniques computed weights for the models based on their performance

¹The complete version of this paper has been published in the *Proceedings of ANLP-NAACL 2000*, Seattle, WA, USA, Morgan Kaufman.

on held-out training data. Additionally, we have trained two classification algorithms on this held-out data and applied them to the output of the models, both with and without extra context information.

We have performed a ten-fold cross-validation experiment on the training data of a standard baseNP data set. With `IB1IG` we generated five different baseNP models. We used the nine combination methods for combining their output. Each combined model performed significantly better than the best individual model ($p < 0.001$, measured on baseNP start and end accuracies) but there was no significant difference between the performances of the nine combination models. We have selected the most simple method, majority voting, for the remaining work.

We have applied `IB1IG` in combination with majority voting to the two standard baseNP data sets put forward by Ramshaw and Marcus [2]. For the small data set we obtained an $F_{\beta=1}$ score of 93.26 which reduces the error of the best reported result for this data set (92.8) with 6%. For the large data set our system reached $F_{\beta=1}=94.90$, an 18% error reduction of the best published result for this data set (93.81). We have also used this approach for processing a standard data set for arbitrary NPs. Here we started with finding baseNPs, then we did the NPs at level 1, then those at level 2 and so on. This broke down the problem to a sequence of baseNP identification processes. For this more difficult problem we obtained $F_{\beta=1}=83.79$, 5% less error than the best reported result for this data set (82.98).

3 Concluding remarks

We have put forward a method for recognizing noun phrases by combining the results of a memory-based classifier applied to different representations of the data. We have examined different combination techniques and each of them performed significantly better than the best individual classifier. We applied majority voting to three standard NP data sets and for all of them it managed to improve the best result for that data set known to us.

References

- [1] Walter Daelemans, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch. *TiMBL: Tilburg Memory Based Learner, version 2.0, Reference Guide*. ILK Technical Report 99-01., 1999. <http://ilk.kub.nl/>.
- [2] Lance A. Ramshaw and Mitchell P. Marcus. Text chunking using transformation-based learning. In *Proceedings of the Third ACL Workshop on Very Large Corpora*. Cambridge, MA, USA, 1995.
- [3] Erik F. Tjong Kim Sang and Jorn Veenstra. Representing text chunks. In *Proceedings of EACL'99*. Bergen, Norway, 1999.
- [4] Hans van Halteren, Jakub Zavrel, and Walter Daelemans. Improving data driven wordclass tagging by system combination. In *Proceedings of COLING-ACL '98*. Montreal, Canada, 1998.