A Hybrid Machine Learning Approach to Information Extraction

Günter Neumann*

German Research Center for Artificial Intelligence (DFKI)
LT–Lab, DFKI Saarbrücken, D-66123 Saarbrücken, Germany

Abstract. A central issue of Information Extraction (IE) research is the adaptation of an IE–system to a new domain. Since IE–systems are by definition domain–specific in order to achieve the desired efficiency and robustness, the mapping of linguistic structures to domain specific structures, is specified in an explicit and direct way. Because of the very idosyncratic nature of these mappings, they are usually not re–useable in other domains or for other text styles. However, it has been shown that a manual specification of such mapping rules are very expensive and that it is very hard to keep the mappings up to date. Hence, recently Machine Learning (ML) methods for automatic acquisition of such mappings are exploited and systematically evaluated. So far, a number of statistical and symbolic-based methods have been explored and evaluated, however mainly in non-hybrid environments. Therefore, an interesting question is whether the combination of a stochastic and a symbolic ML-method can improve the performance of an IE–system. As basis for our statistical-based learner we have chosen the Maximum Entropy Modelling (MEM) framework. The symbolic learner is based on our work on *data–driven extraction of lexicalized tree grammars*. The core idea is to generate trees from the information obtained from the shallow parser applied on the annotated training data. These trees are further generalized by cutting of irrelevant subtrees. Both learning methods are applied independently of each other during the training phase. The application phase is realized as an iterative tag–insertion algorithm, where the tags are actually determined by the learned mappings. The envisaged hybrid learning behavior is achieved through a voting mechanism, which is applied in each iteration on the tagging results of all active mappings. We have systematically evaluated our approach following the MUC7 guide lines on a manually annotated small corpus of *German* newspaper articles about company turnover (75 documents with a total of 5878 tokens; 60 documents are used for training, 15 for testing). For the template element task, we obtained 85.18% F-measure using the hybrid approach, compared to 79.27% for MEM and 51.85% for the symbolic learner when running them in isolation. The overall result is competitive with related work described in IE literature mainly for English using larger document sets than we had available for German, e.g., Chieu & Ng (2002).

References

CHIEU & NG (2002): A Maximum Entropy Approach to Information Extraction from Semi–Structured and Free Text. In Proceedings of AAAI-2002.

* Thanks to my student Volker Morbach for his great help during the implementation and evaluation phase of the project.