

Memory-based biomedical Named-Entity tagging¹

Walter Daelemans, Vincent Van Asch, Roser Morante
CLiPS Computational Linguistics Group, University of Antwerp, Belgium

The annotation of named entities was done by means of a Memory-Based Shallow Parser (MBSP) for biomedical text. The parser was developed in the context of the BioMinT EU project².

Shallow parsing is based on the idea that full (or deep) syntactic structure is not always necessary for information extraction tasks and consequently assigns only a limited amount of syntactic information to natural language sentences. For example, in biomedical information extraction, we are interested in finding concepts (represented by basic nominal and verbal phrases) and relations between them, rather than in recovering an elaborate syntactic analysis. MBSP, as an instance of this approach, combines a number of modules, each trained using memory-based learning from linguistically annotated corpora (Daelemans & Van den Bosch, 2005).³ An MBL system consists of two components: a memory-based learning component and a similarity-based performance component. During training, the learning component adds new training instances to the memory without any abstraction or restructuring. During classification, the classification of the most similar instance in memory is taken as classification for the new test instance.

In our case, the training data was version III of the Wall Street Journal part of the manually annotated Penn Treebank corpus (Mitchell et al, 1993) and the GENIA annotated Medline corpus (Ohta et al., 2002). Tools trained on Penn Treebank data (consisting exclusively of newspaper text) are not very well suited for other domains like biomedical text, and consequently they have to be adapted to the biomedical domain. In order to achieve better accuracy on the analysis of biomedical text, we adapted the shallow parser to the biomedical domain by replacing the tokenizer and the part-of-speech in the original shallow parser by modules developed on and trained on the GENIA corpus, a hand-annotated corpus of medline abstracts. A general named entity recognizer (locations, organizations, persons) was replaced by a named entity recognizer assigning biomedical concepts, trained on the Gernia corpus.

¹ This work was supported through the BOF-GOA project Biograph of the University of Antwerp

² The BioMinT project was funded by the European Commission, contract-no. QLRI-CT-2002-02770 under the RTD programme "Quality of Life and Management of Living Resources".

³ See <http://ilk.uvt.nl/timbl> for the implementation of the machine learning software on which MBSP is based.

The shallow parser incorporates the following modules:

- A regular-expression based tokenizer that splits punctuation from adjoining words and splits documents into sentences.
- A part-of-speech tagger that disambiguates the contextually appropriate morphosyntactic class (e.g. noun, verb, etc.) of each word in each sentence.
- A chunker that combines syntactically related words into non-overlapping phrases (e.g. NP, nominal phrase; VP, verbal phrase; PP, prepositional phrase, etc.) on the basis of the part-of-speech tagged words.
- A relation finder, working on output of the chunker, and deciding on the main relations between the heads of the NPs and VP in each sentence (e.g., subject, object, location, etc.).
- A named entity recognizer, working on output of the previous modules and assigning strings of words to named entity types as used in the Genia ontology and annotated in the Genia corpus.

The latter module was used to process the CALBC data. The named entity recognizer was not optimized for optimal feature construction, e.g. most of the syntactic information provided by the tagger and the chunker was not used, and most features were lexical (context) or word-internal (special characters, prefixes, suffixes). It also doesn't use any gazetteer information (except what is present in the training material). Measured by cross-validation on the Genia corpus, the shallow parser as a whole is fairly accurate for biomedical text (99% for tokenization, 98% for part of speech tagging, 90% f-score for chunking), and reasonably accurate for named entity recognition (65% f-score overall, 75% f-score for proteins). Accuracy here is mentioned as complete identification of the (multi-word) named entities.

References

Daelemans W, van den Bosch A: *Memory-based Language Processing*. Cambridge University Press 2005.

Mitchell PM, Santorini B, Marcinkiewicz M: *Building a Large Annotated Corpus of English: The Penn Treebank*. *Computational Linguistics* 1993, 19(2):313-330.

Ohta T, Tateisi Y, Kim JD: *The GENIA corpus: an annotated research abstract corpus in molecular biology domain*. In *Proceedings of the second international conference on Human Language Technology Research, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. 2002:82-86*.