# Automatically Adapting an NLP Core Engine to the Biology Domain

Ekaterina Buyko\*1, Joachim Wermter1, Michael Poprat1 and Udo Hahn1

<sup>1</sup>Jena University Language and Information Engineering (JULIE) Lab, Friedrich-Schiller Universität Jena, Fürstengraben 30, D-07743 Jena, Germany

Email: {buyko|wermter|poprat|hahn}@coling-uni-jena.de;

\*Corresponding author

#### **Abstract**

**Background:** Rather than specifying rules, constraints and lexicons for NLP systems manually, we advocate a procedure for automatically acquiring linguistic knowledge using machine learning (ML) methods. In order to demonstrate how feasible this approach is, we automatically adapt OPENNLP, an open source ML-based NLP tool suite, to the sublanguage domain of biology.

**Results:** In the first evaluation ever of a ML-based ensemble of core NLP components in the biology domain, which are all based on the maximum entropy method, we demonstrate that the performance of OPENNLP's sentence splitter, tokenizer, part-of-speech tagger, chunker and parser on bio texts matches up with state-of-the-art performance figures from the newspaper domain.

**Conclusions:** Core BioNLP systems can automatically be derived from non-bio domain settings by (re-)training rather than by manual rule re-writing and lexical re-specification.

## **Background**

The prevailing approach to build large-scale BioNLP systems relies heavily on human efforts and expertise, i.e., manual rule, constraint and lexicon specification and maintenance (e.g., [1,2]). It is often deplored that

such an approach is time-costly, error-prone and heavily iterative (lots of test/modify cycles). Within the NLP community proper, this approach is almost completely abandoned and, as an alternative, procedures for automatically acquiring linguistic knowledge, i.e., machine learning (ML) techniques, dominate the field (cf., e.g., various CoNLL activities<sup>1</sup>). On the flip side of this success story, (semi-)supervised ML methods at least, require annotated corpora to be available as a training resource. Fortunately, in the molecular biology domain two richly annotated language resources (GENIA<sup>2</sup> and PENNBIOIE<sup>3</sup>) already exist. In the following, we stipulate that the exchange of domain (e.g., from newspapers to biological articles), merely boils down to re-training NLP tools on annotated corpora rather than laborious manual re-writing and changing of rule-based systems and lexicons. For this purpose, we chose OPENNLP TOOLS<sup>4</sup> as an open-source tool suite which contains a variety of JAVA-based NLP components. Our focus is here on five "core" NLP components, viz. sentence detection, tokenization, POS tagging, chunking and parsing. OpenNLP is a homogeneous package based on a single machine learning approach, viz. maximum entropy (ME) [3,4]. The rationale behind ME for any collection of facts is to choose a model which is consistent with all the facts but otherwise as uniform as possible. Each OPENNLP tool requires an ME model that contains statistics about the component's default features combining diverse contextual information such as words around end-of-sentence boundaries for the sentence splitter or word/tag combinations in five-word/tag-window for the chunker. The components are partly based on publications like the chunking model described by Sha & Pereira [5] as well as the part-of-speech tagger and the parser described by Ratnaparkhi [6]. For training and testing of all OPENNLP components we considered, the above-mentioned two major biomedically annotated corpora, viz. GENIA and PENNBIOIE, were employed which currently contain the most elaborate annotations relevant for syntactic analysis in the bio field. We performed ten-fold cross validation for all OPENNLP tools on both corpora.

## Results

The sentence splitter achieved an accuracy of approximatively 99.0% on Genia and 97.4% on Pennbiole (see also Table 1).<sup>5</sup> False positives [7] which constitute the majority of errors (67% in Genia) usually do not occur in biological terms but rather occur in abbreviations of names in literature citations. In Pennbiole false negatives are mainly due to section headings ending with ":" such as "OBJECTIVE:".

<sup>&</sup>lt;sup>1</sup>http://ilps.science.uva.nl/~erikt/signll/conll

 $<sup>^2 \</sup>rm http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/$ 

<sup>&</sup>lt;sup>3</sup>http://bioie.ldc.upenn.edu

<sup>&</sup>lt;sup>4</sup>http://opennlp.sourceforge.net

<sup>&</sup>lt;sup>5</sup>By comparison, the accuracy of the sentence splitter on newspaper data in [6] is 98.9%.

	Sentence Splitter	Tokenizer	PoS Tagger	Chunker	Parser
	(Accuracy)	(Accuracy)	(Accuracy)	(F-Score)	(F-Score)
GENIA	99,0%	$99,\!6\%$	98,9%	$93,\!6\%$	87,0%
PENNBIOIE	97,4%	99,0%	98,9%	89,5%	85,2%

Table 1: Evaluation Results of OpenNLP Components

The tokenizer achieved an accuracy of 99.6% on Genia and 99.0% on PennBioIE (Table 1). On both corpora the tokenization of commas and parentheses also occurring in a lot of chemical names does not seem to be problematic. The lower accuracy of the tokenizer with regard to "-" and "/" in PennBioIE, however, reveals the fact that in PennBioIE (though not in Genia) expressions containing these symbols are meticulously token-annotated and thus raise the bar for the tokenizer.

The part-of-speech tagger achieved an accuracy of 98.9% on Genia and 98.9% on PennBioIE (Table 1).<sup>6</sup> For these runs, we also used the tagger option to include an automatically generated tag-dictionary from the training data enumerating a fixed set of possible tags for each word to delimit the decision space. Without the dictionaries, the tagging accuracy drops to 98.2% on Genia and 98.3% on PennBioIE. With the exception of "binding" the tagging mistakes in the top 10 are not biomedical terms. The erroneous taggings seem to replicate the tagging mistakes discussed already in [6].<sup>7</sup>

The chunker achieved an overall F-score of 93.6% on Genia and 89.5% on Pennbiole (Table 1). At a first glance, it seems that the chunker performs markedly better when trained on Genia, in particular, concerning the important recognition rate for NPs (92.3% on Genia vs. 85.1% on Pennbiole). However, these results must be treated with caution because of some inadequacies of the Chunklink script<sup>8</sup> in converting treebank anntotations into IOB-notation, especially when Pennbiole is concerned. The overall F-score of 93.6% on Genia is a state-of-the-art figure (in comparison with performance figures from Conll 2000) and even better than when trained on newspaper data. In particular, the good recognition rates for NPs, PPs (96.9% in Genia) and VPs (95.9% Genia) are essential for deeper linguistic analysis. We evaluated the parser along standard parameters such as Bracketing recall and Bracketing precision [8,9] on all sentences and on those with a length less than 40 words (about 90% of the sentences in Genia and 88,6% in Pennbiole). The parser F-score on Genia 87.0% is about two percentage points

 $<sup>^6\</sup>mathrm{The}$  ME-based POS tagger in [6] achieves 96.6% accuracy on new spaper data.

<sup>&</sup>lt;sup>7</sup>For example, the word "that" is confused being used as a relative pronoun (WDT), a subordinating conjunction (IN) or a determiner (DT); similarly, the word "both" by its use as a determiner (DT) and a conjunction (CC).

<sup>8</sup>http://ilk.kub.nl/~sabine/chunklink/

 $<sup>^9\</sup>mathrm{The}$  OpenNLP chunker obtained an overall F-score of 92.4 on the CoNLL 2000 data (own evaluation).

<sup>&</sup>lt;sup>10</sup>These proportions are similar to those in the Penn Treebank [9].

better than on PennBioIE 85.2% (Table 1). It should also be noted that the number of parsing failures is quite low. For Genia, only one out of 4370 sentences could not be parsed (0.02%), and for PennBioIE, this number amounts to eight out of 6296 (0.1%).<sup>11</sup> Particularly interesting from the perspective of deeper linguistic analysis of biological literature is the parser's capacity to predict semantic function tags of constituents. For these, the parser's F-scores only decrease about one point on each corpus.

### **Conclusion**

There are two major results of this study. We have the first evaluation of an ensemble of NLP components all of which are based on the same ML paradigm (maximum entropy) within the biology domain. Second, our results from the biology domain match up to state-of-the-art performance figures from the newspaper domain. As OPENNLP TOOLS stand for an ML-based approach to acquire linguistic regularities (which is quite unusual in the biology domain, up until now) the state-of-the-art performance figures we determined provide ample evidence for challenging common wisdom. We might no longer create and maintain rule systems and lexicons manually but simply (re-)train NLP components on properly annotated data. Our current efforts are directed at exploring an ML paradigm which helps minimize the amount of annotated data required for (still) effective training of NLP tools.

Acknowledgments. This research was funded by the EC's 6th Framework Programme (4th call) and conducted within the BOOTStrep consortium under grant FP6-028099.

#### References

- 1. Friedman C, Kra P, Yu H, Krauthammer M, Rzhetsky A: Genies: A natural language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics* 2001, 17(Suppl. 1):74–82.
- 2. Gaizauskas R, Demetriou G, Artymiuk P, Willett P: **Protein structures and information extraction from biological texts: The Pasta system**. *Bioinformatics* 2003, **19**:135–143.
- 3. Berger A, Della Pietra S, Della Pietra V: A maximum entropy approach to natural language processing. *Computational Linguistics* 1996, **22**:39–71.
- 4. Rosenfeld R: A maximum entropy approach to adaptive statistical language modeling. Computer, Speech and Language 1996, 10:187–228.
- 5. Sha F, Pereira F: Shallow Parsing with Conditional Random Fields. In *Proceedings of the HLT-NAACL* 2003 2003.
- 6. Ratnaparkhi A: Maximum Entropy Models for Natural Language Ambiguity Resolution. *PhD thesis*, University of Pennsylvania 1998.
- Palmer DD, Hearst MA: Adaptive Multilingual Sentence Boundary Disambiguation. Computational Linguistics 1997, 23(2):241–267.
- 8. Collins M: Three generative, lexicalised models for statistical parsing. In *Proceedings of the ACL'97/EACL'97* 1997:16–23.
- 9. Clegg AB, Shepherd AJ: Evaluating and integrating Treebank parsers on a biomedical corpus. In Proceedings of the ACL 2005 Workshop on Software 2005.

<sup>&</sup>lt;sup>11</sup>A flat chunk phrase structure is returned in such a case.