Tackling the BioCreative2 Gene Mention task with Conditional Random Fields and Syntactic Parsing

Andreas Vlachos[1] av308@cl.cam.ac.uk

[1] Computer Laboratory, University of Cambridge, CB3 0FD, Cambridge, UK

Abstract

This paper presents an approach to Gene Mention tagging using Conditional Random Fields (CRFs) and syntactic parsing, by taking advantage of the flexibility of the former in order to add features from the output of the latter. We did not use any material or information other than the training data provided in order to maintain the domain independence of the system. Nevertheless, the resulting system achieved 82.84% F-score, which places it in the second performance quartile of the competition.

Keywords: CRFs, syntactic parsing, gene mention tagging

1 Introduction

In this paper we describe our participation in the BioCreative Gene Mention tagging task. The main components used were the Conditional Random Fields implementation (CRFs) [2] from MALLET [3] and the RASP tokenizer, part-of-speech (POS) tagger, lemmatizer and syntactic parser [1]. CRFs were chosen due to the recent success in similar named entity recognition (NER) tasks [4], as well as their flexibility in adding features. The latter aspect we intend to take advantage of in our system, by adding linguistic features from the output of the various components of the RASP toolkit. No other resources were used, therefore the system presented here could be used for other NER tasks. Our expectation is that the combination of deep linguistic analysis and a state-of-the-art statistical model should be able to achieve competitive performance without using domain-specific resources.

2 Methods

As a first step we created tokenized training data from the materials provided, which were a list of sentences with two sets of annotations. We used only the first set of annotations (from the GENE.eval file) in order to annotate the sentences. Then we tokenized the text using RASP's domain independent tokenizer, adding as token boundaries the gene mention boundaries from the annotations. We used the BIEWO scheme for labelling the resulting tokens – the first token of a multitoken mention is tagged as B, the last token as E, the inner ones as I, single token mentions as W and tokens outside an entity as O. In our experiments we found that we obtained better performance with this scheme than with the standard IOB format, possibly due to the large number of multi-token gene mentions and their overlap with common English words or biomedical terms. For each token we extracted the simple orthographic features listed in Table 1.

Then we pass each tokenized sentence to RASP's syntactic parser. We parameterized RASP to pass multiple POS tags per token to the parser to ameliorate unknown word errors and used the grammatical relations (GRs) output from the top-ranked parse. The output of RASP (without the XML tags for brevity) looks like this:

Table 1: Simple orthographic features

the token itself	if it contains digit(s)
if it is alphanumeric	if it contains only digits
if it is alphabetic	if it contains dash(es)
if it is titlecase	if it contains dot(s)
if it is lowercase	if it contains any punctuation marks
if it is uppercase	if it contains punctuation marks and digits
if it is mixed case	2 and 3 letter prefixes and suffixes

```
("No" "post-operative" "haemorrhages" "from" "the"
"prostheses" "were" "observed" ".")

(|ncsubj| |observe+ed:8_VVN| |haemorrhage+s:3_NN2| _)
(|aux| |observe+ed:8_VVN| |be+ed:7_VBDR|)
(|passive| |observe+ed:8_VVN|)
(|det| |haemorrhage+s:3_NN2| |No:1_AT|)
(|ncmod| _ |haemorrhage+s:3_NN2| |from:4_II|)
(|dobj| |from:4_II| |prosthesis+s:6_NN2|)
(|det| |prosthesis+s:6_NN2| |the:5_AT|)
(|ncmod| _ |haemorrhage+s:3_NN2| |post-operative:2_JJ|)
```

The features extracted from RASP's output for each token are listed in Table 2. It must be noted at this point that the features added from the output of RASP may contain noise, since syntactic parsing is a very complicated task.

Table 2: Features extracted from the output of RASP

the lemma and the POS tag(s) associated with the token
the lemmas for the previous two and the following two tokens
the lemmas of the verbs to which this token is subject (*ncsubj* relation)
the lemmas of the verbs to which this token is object (*dobj* relation)
the lemmas of the nouns to which this token acts as modifier (*ncmod* relation)
the lemmas of the modifiers of this token (*ncmod* relation)

3 Results and analysis

For each of the experiments we used the CRF implementation of MALLET and trained the model until convergence. During testing, we followed the same preprocessing and feature extraction procedure, with the exception that we didn't use the boundaries of the gene mentions for tokenization since they were unknown. The results for the three submitted runs appear in Table 3. For our first run, we trained a 3rd order CRF model on the standard RASP tokenizer's output. For the second run, we altered the tokenization step in order to include dashes and slashes as token separators, since, according to the annotation scheme, in cases such as *"p65-selected"*, only *"p65"* should be returned as a gene mention. This improved the performance substantially. For the third run, we kept the tokenization from the second run, but we reduced the CRF order to second order, since we would

like to reduce the training time of the system. There was a slight increase in performance, probably because the lower order CRF looks for simpler patterns which resulted in better recall.

Table 3: Evaluation of the submitted runs

	Precision	Recall	F
Run1	85.37	74.11	79.34
Run2	86.59	79.15	82.70
Run3	86.28	79.66	82.84

We also wanted to explore how beneficial was the use of linguistic features. Therefore, using Run3 as the basis (2nd order CRF with adapted tokenization), we ran experiments with subsets of the features extracted from the output of RASP. The results of Table 4 suggest that lemmas appear to be the most useful features, while POS tags and syntactic features improve performance less. One should take into account though that, apart from the noise introduced during parsing, specific syntactic features are only useful in sentences that exhibit them. For example, in the sentence *"For the P transcript from phage with the G(-) orientation..."*, *"P transcript"* is a gene mention but the lemmas *"transcript"* and *"p"* are not strong enough cues since they can be found outside of gene mentions. As a result, the model without syntactic features fails to recongize it as such. However, when the fact that *"p"* is a modifier of *"transcript"* is added as a feature from the syntactic analysis of RASP, then it is recongized correctly. In order to demonstrate the usefulness of the syntactic features more clearly, there needs to be an evaluation on an appropriate test set that contains more cases that need such features. Also, consistent annotation of the test set is important for quantitative assessment. In order to demonstrate this point, we measured our performance using only the first set of annotations (GENE.eval). As column F-strict of Table 4 shows, while the scores are lower, the gains in performance obtained by adding more features are larger than those observed when evaluating using both sets of annotations.

Table 4: Evaluation of the features

features	Precision	Recall	F	F-strict
simple_features	82.97	76.64	79.68	66.55
simple_features+lemmas	86.13	79.56	82.72	70.85
simple_features+lemmas+pos	85.82	79.91	82.76	71.03
simple_features+lemmas+pos+syntax	86.28	79.66	82.84	71.55

References

[1] Briscoe, E., Carroll, J. and Watson, R., The Second Release of the RASP System, *In Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, 2006.

[2] Lafferty, J. D., McCallum, A. and Pereira, F. C. N., Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, *Proceedings of ICML 2001*, 282–289, 2001.

[3] McCallum A. K., MALLET: A Machine Learning for Language Toolkit, `http://mallet.cs.umass.edu`

[4] Settles B., Biomedical Named Entity Recognition Using Conditional Random Fields and Novel Feature Sets, *Proceedings of the JNLPBA*, 2004.