Exploring ways beyond the simple supervised learning approach for biological event extraction

György Móra[1], Richárd Farkas[1], György Szarvas[2*], Zsolt Molnár[3]
gymora@gmail.com, rfarkas@inf.u-szeged.hu,
szarvas@tk.informatik.tu-darmstadt.de, zsolt@acheuron.hu
[1] Hungarian Academy of Sciences, Research Group on Artificial Intelligence
Aradi vértanuk tere 1., H-6720 Szeged, Hungary
[2] Ubiquitous Knowledge Processing Lab, Technische Universität Darmstadt
Hochschulstraße 10., D-64289 Darmstadt, Germany
[3] Acheuron Hungary Ltd., Chemo-, and Bioinformatics group,
Tiszavirág u. 11., H-6726 Szeged, Hungary

Abstract

Our paper presents the comparison of a machine-learnt and a manually constructed expert-rule-based biological event extraction system and some preliminary experiments to apply a negation and speculation detection system to further classify the extracted events. We report results on the *BioNLP'09 Shared Task on Event Extraction* evaluation datasets, and also on an external dataset for negation and speculation detection.

1 Introduction

When we consider the sizes of publicly available biomedical scientific literature databases for researchers, valuable biological knowledge is accessible today in enormous amounts. The efficient processing of these large text collections is becoming an increasingly important issue in Natural Language Processing. For a survey on techniques used in biological Information Extraction, see (Tuncbag et al., 2009).

The *BioNLP'09 Shared Task* (Kim et al., 2009) involved the recognition of bio-molecular events in scientific abstracts. In this paper we describe our systems submitted to the *event detection and characterization* (Task1) and the *recognition of negations and speculations* (Task3) subtasks. Our experiments can be regarded as case studies on i) how to define a framework for a hybrid human-machine biological information extraction system, ii) how the linguistic scopes of negation/speculation keywords relate to biological event annotations.

2 Event detection

We formulated the event extraction task as a classification problem for each event-trigger-word/protein pair. A domain expert collected 140 keywords which he found meaningful and reliable by manual inspection of the corpus. This set of high-precision keywords covered 69.8% of the event annotations in the training data.

We analysed each occurrence of these keywords in two different approaches. We used C4.5 decision tree classifier to predict one of the event types considered in the shared task or the keyword/protein pair being unrelated; and we also developed a hand-crafted expert system with a biological expert. We observed that the two systems extract markedly different sets of true positive events. Our final submission was thus the union of the events extracted by the expert-rule-based and the statistical systems (we call this *hybrid system* later on).

2.1 The statistical event classifier

The preprocessing of the data was performed using the *UltraCompare* (Kano et al., 2008) repository provided by the organizers of the challenge: *Genia sentence splitter*, *Genia tagger* for POS coding and NER.

The statistical system classified each keyword/protein pair into 9 event and 2 non-event classes. A pair was either labeled according to the predicted event type (the keyword as an event trigger and the protein name as the theme of the event), `non-event` (keyword not an event trigger) or `wrong-protein` (the theme of the event is a different protein). We chose to use two non-event

[*] On leave from RGAI of Hungarian Acad. Sci.

Proceedings of the Workshop on BioNLP: Shared Task, pages 137–140,
Boulder, Colorado, June 2009. ©2009 Association for Computational Linguistics

classes to make the decision tree more human readable (the negative cases being separated). This made the comparison of the statistical model and the rule-based system easier.

The features we used were the following: 1) the words and POS codes in a window (± 3 tokens) around the keyword, preserving position information relative to the keyword; 2) the distances between the keyword and the two nearest annotated proteins (left and right) and the theme candidate as numeric features[1]. The protein annotations were replaced by the term `$protein`, *Genia tagger* annotations by `$genia-protein` (mainly complexes), to enable the classifier to learn the difference between events involved in the shared task, and events out of the scope of the task. Events with protein complexes and families often had the same linguistic structure as events with annotated proteins. As complexes did not form events in the shared task, they sometimes misled our local-context-based classifier. For example ’*the binding of* ***ISGF3***’ was not annotated as an event because the theme is not a ”protein” (as defined by the shared task guidelines), while ’*the binding of* ***TRAF2***’ was (TRAF2 being a protein, and not a complex as in the former example).

We trained a *C4.5* decision tree classifier using *Weka* (Witten and Frank, 2005). The human readable models and fast training time motivated our selection of a learning algorithm which allowed a straightforward comparison with the expert system.

2.2 Expert-rule-based system

The expert system was constructed by a biologist who had over 4 years of experience in similar tasks. The main idea was to define rules – which have a very high precision – in order to compare them with the learnt decision trees and to increase the coverage of the final system by adding these annotations to the output of the statistical system. We only managed to prepare expert rules for the *Phosphorylation* and *Gene_expression* classes due to time constraints (a total of 46 patterns). The expert was asked to construct high-precision rules (they were tested on the train set to keep the false positive rate near zero) in order to gain insight into the structure of reliable rules.

Here each rule is bound to a specific keyword. Every rule is a sequence of ”word patterns” (with or without a suffix). A word pattern can match a protein, an arbitrary word, an exact word or the keyword. Every pattern can have a *Regular Expression* style suffix:

Table 1: Word pattern types and suffixes

`<keyword>`	matching the keyword of the event
`"word"`	matching regular words
`_`	matching any token
`$protein`	matching any annotated protein
`?`	zero or one of the word pattern
`*`	zero or more of the word pattern
`+`	one or more of the word pattern
`{a,b}`	definite number of word patterns

For example the `'<expression> _? "of" _? $protein'` pattern recognizes an event with the keyword *expression*, followed by an arbitrary word and then the word *of*, or immediately by *of* and then a protein (or immediately by the protein name).

An obvious drawback of this system is that negation is not allowed, so the expert was unable to define a word pattern like `!"of"` to match any token besides *of*. This extension would have been a straightforward way of improving the system.

2.3 Experimental results

We expected the recall of the hybrid system to be near the sum of the recalls of the individual systems, meaning that they had recognized different events, as the pattern matching was mainly based on the order of the tokens, while the statistical classifier learned position-oriented contextual clues. Thanks to the high precision of the rule-based system, the overall precision also increased. The two event classes which were included in the expert system had a significantly better precision score. The coverage of the *Phosphorylation* class was lower than that for the *Gene_expression* class because its patterns were still incomplete[2].

[1] More information on the features and parameters used can be found at `www.inf.u-szeged.hu/rgai/BioEventExtraction`

[2] A discussion on comparing the contribution of the two approaches and individual rules can be found at `www.inf.u-szeged.hu/rgai/BioEventExtraction`

Table 2: Results of rule based-system compared to the statistical and combined systems (R/P/fscore)

	All Event	Gene_exp.	Phosph.
stat.	16 / 31 / 21	36 / 41 / 38	73 / 37 / 49
rule	5 / 80 / 10	20 / 85 / 33	17 / 58 / 26
hybrid	22 / 37 / 27	56 / 51 / 54	81 / 40 / 53

3 Recognition of negations and speculations

For negation and speculation detection, we applied a model trained on a different dataset (Vincze et al., 2008) of scientific abstracts, which had been specially annotated for negative and uncertain keywords and their linguistic scope. Due to time constraints we used our model to produce annotations for Task3 without any sort of fine tuning to the shared task gold standard annotations.

The only exception here was a subclass of speculative annotations that were not triggered by a word used to express uncertainty, but were judged to be speculative because the sentence itself reported on some experiments performed, the focus of the investigations described in the article, etc. That is, it was not the meaning of the text that was uncertain, but – as saying that something has been examined does not mean it actually exists – the sentence implicitly contained uncertain information. Since such sentences were not covered by our corpus, for these cases we collected the most reliable text cues from the shared task training data and applied a dictionary-lookup-based approach. We did this so as to get a comprehensive model for the Genia negation and speculation task.

As for the explicit uncertain and negative statements, we applied a more sophisticated approach that exploited the annotations of the *BioScope* corpus (Vincze et al., 2008). For each frequent and ambiguous keyword found in the approximately 1200 abstracts annotated in *BioScope*, we trained a separate classifier to discriminate keyword/non-keyword uses of each term, using local contextual patterns (neighbouring lemmas, their POS codes, etc.) as features. In others words, for the most common uncertain and negative keywords, we attempted a context-based disambiguation, instead of a simple keyword lookup. Having the keywords, we predicted their scope using simple heuristics ('*to the end of the sentence*', '*to the next punctation mark in both directions*', etc.). In the shared task we examined each extracted event and they were said to be negated or hedged when some of their arguments (trigger word, theme or clause) were within a linguistic scope.

3.1 Experimental results

First we evaluated our negation and speculation keyword/non-keyword classification models on the *BioScope* corpus by 5-fold cross-validation. We trained models for 15 negation and 41 speculative keywords. We considered different word forms of the same lemma to be different keywords because they may be used in a different meaning/context. For instance, different keyword/non-keyword decision rules must be used for *appear*, *appears* and *appeared*. We trained a *C4.5* decision tree using word uni- and bigram features and POS codes to discriminate keyword/non-keyword uses and compared the results with the most frequent class (MFC) baseline.

Overall, our context-based classification method outperformed the baseline algorithm by 3.7% (giving an error reduction of 46%) and 3.1% (giving an error reduction of 27%) on the negation and speculation keywords, respectively. The learnt models were typically very small decision trees i.e. they represented very simple rules indicating collocations (like '*hypothesis* is a keyword if and only if followed by *that*, etc.). More complex rules (e.g. '*clear* is a keyword if and only if *not* is in ±3 environment') were learnt just in a few cases.

Our second set of experiments focused on Task3 of the shared task (Kim et al., 2009). As the official evaluation process of Task3 was built upon the detected events of Task1, it did not provide any useful feedback about our negation and speculation detection approach. Thus instead of our Task1 output, we evaluated our model on the gold standard Task1 annotation of the training and the development datasets. The statistical parts of the system were learnt on the *BioScope* corpus, thus the train set was kept blind as well. Table 3 summarises the results obtained by the explicit negation, speculation and by the full speculation (both explicit and implicit keywords) detection methods.

Analysing the errors of the system, we found that

Table 3: Negation and speculation detection results

	Train (R/P/F)	Dev. (R/P/F)
negation	46.9 / 61.3 / 52.8	42.8 / 57.9 / 49.2
exp. spec.	15.4 / 39.5 / 23.6	15.4 / 32.6 / 20.1
full spec.	25.5 / 71.1 / 37.5	27.9 / 65.3 / 39.1

most of the false positives came from the different approaches of the *BioScope* and the *Genia* annotations (see below for a detailed discussion). Most of the false negative predictions were a consequence of the incompleteness of our keyword list.

3.2 Discussion

We applied this negation and speculation detection model more as a case study to assess the usability of the *BioScope* corpus. This means that we did not fine-tune the system to the *Genia* annotations. Our experiments revealed some fundamental and interesting differences between the Genia-interpretation of negation and speculation, and the corpus used by us. The chief difference is that the *BioScope* corpus was constructed following more linguistic-oriented principles than the Genia negation and speculation annotation did, which sought to extract biological information. These differences taken together explain the relatively poor results we got for the shared task.

There are significant differences in the interpretation of both at the keyword level (i.e. what triggers negation/uncertainty and what does not) and in the definition of the scope of keywords. For example, in a sentence like '*have **NO** effect on the inducibility of the IL-2 promoter*', Genia annotation just considers the *effect* to be negated. This means that the *inducibility* of *IL-2* is regarded as an assertive event here. In *BioScope*, the complements of *effect* are also placed within the scope of *no*, thus it would also be annotated as a negative one. We argue here that the above example is not a regular sentence to express the fact: *IL-2* is *inducible*. We rather think that if the paper has some result (evidence) regarding this event, it should be stated elsewhere in the text, and we should not retrieve this information as a fact just based on the above sentence. Thus we argue that more sophisticated guidelines are needed for the consistent annotation and efficient handling of negation and uncertainty in biomedical text mining.

4 Conclusions

We described preliminary experiments on two different approaches which take us beyond the "take-goldstandard-data, extract-some-features, train-a-classifier" approach for biomedical event extraction from scientific texts (incorporating rule-based systems and linguistic negation/uncertainty detection). The systems introduced here participated in the Genia Event annotation shared task. They achieved relatively poor results on this dataset, mainly due to 1) the special annotation guidelines of the shared task (like disregarding events with protein complex or family arguments, and treating subevents as assertive information) and 2) the limited resources we had to allocate for the task during the challenge timeline. We consider that the lessons learnt here are still useful and we also plan to improve our system in the near future.

5 Acknowledgements

The authors would like to thank the organizers of the shared task for their efforts. This work was supported in part by the NKTH grant of the Hungarian government (codename BELAMI).

References

Y. Kano, N. Nguyen, R. Saetre, K. Yoshida, Y. Miyao, Y. Tsuruoka, Y. Matsubayashi, S. Ananiadou, and J. Tsujii. 2008. Filling the gaps between tools and users: a tool comparator, using protein-protein interaction as an example. *Pac Symp Biocomput*.

J-D. Kim, T. Ohta, S. Pyysalo, Y. Kano, and J. Tsujii. 2009. Overview of bionlp'09 shared task on event extraction. In *Proceedings of Natural Language Processing in Biomedicine (BioNLP) NAACL 2009 Workshop*. To appear.

N. Tuncbag, G. Kar, O. Keskin, A. Gursoy, and R. Nussinov. 2009. A survey of available tools and web servers for analysis of protein-protein interactions and interfaces. *Briefings in Bioinformatics*.

V. Vincze, Gy. Szarvas, R. Farkas, Gy. Móra, and J. Csirik. 2008. The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9(Suppl 11):S9.

I. H. Witten and E. Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition*. Morgan Kaufmann.