AKANE System: Protein-Protein Interaction Pairs in the BioCreAtIvE2 Challenge, PPI-IPS subtask

Rune Sætre[1] satre@is.s.u-tokyo.ac.jp
Kazuhiro Yoshida[1] kyoshida@is.s.u-tokyo.ac.jp
Akane Yakushiji[2] yakushiji.akane@jp.fujitsu.com
Yusuke Miyao[1] yusuke@is.s.u-tokyo.ac.jp
Yuichiro Matsubayashi[1] y-matsu@is.s.u-tokyo.ac.jp
Tomoko Ohta[1] okap@is.s.u-tokyo.ac.jp

[1] Department of Computer Science, University of Tokyo, Hongo 7-3-1, Bunkyo-ku, Tokyo 113-0033, Japan
[2] FUJITSU LABORATORIES LTD. 4-1-1 Kamikodanaka, Nakahara-ku, Kawasaki-shi, Kanagawa 211-8588, Japan

Abstract

This report summarizes the participation of the Tsujii-lab group in the 2006 BioCreative2 text mining challenge[1]. It describes the systems used, the results attained, and the lessons learned. The basic idea was to see how well the AKANE system could perform on a full-text Protein-Protein Interaction (PPI) Information Extraction (IE) task. AKANE system is a recently developed, sentence-level PPI system that achieved a 57.3 F-score on the AImed corpus. In order to use the AKANE system for the BioCreative task, the given training data had to be preprocessed. The BioCreative training data contained just a list of interacting protein pair identifiers for each given full-text article, while the expected input for the AKANE system is annotated sentences like in the AImed corpus. In order to transform the full-text articles into AImed sentence-level annotations, the text was first stripped of all HTML coding to get a plain text representation. Then, each mention of protein names were tagged by a Named Entity Recognizer (NER), and all interacting and co-occurring pairs in single sentences were used for training. A pipeline architecture was made to deal with each of these challenges. Some postprocessing was also necessary, in order to transform the results from the AKANE system into the expected format for the BioCreative2 challenge. The postprocessing included filtering and ranking the results, and balancing precision and recall to maximize the F-score.

Keywords: bionlp, protein-protein interaction, natural language processing

1 Introduction and Methods

Our system implements a pipeline architecture, where the modules deal with Sentence Detection (SD), Named Entity Recognition (NER), Parsing, and Protein-Protein-Interaction (PPI) extraction. All the modules use machine learning to maximize the performance on small manually annotated biological training corpora. A separate system was made for transforming the article level BioCreative training data into a sentence level AImed PPI-style format (See section 1.3.1). Each module is briefly described below.

1.1 Sentence Detection

The sentence splitter for biomedical text was trained by a maximum entropy (MaxEnt) method [1], and it employs the GENIA corpus for training [4]. First, the sentence splitter detects candidate

[1] http://biocreative.sourceforge.net/biocreative_2.html

positions for splitting using selected delimiters: periods, commas, single/double quotation marks, right parentheses, etc. Then, it classifies whether the positions really split sentences or not. Features used by the classifier are: Delimiters, Previous/Next words, and info about special characters, numbers and capitalization. Some transformations of the words, like removing commas, parentheses, etc. and making lower-case versions were also used. The classifier achieved an F-score of 99.7 on 200 unseen GENIA abstracts. However, there seem to be slightly more errors on the BioCreative data set, mainly because of full text HTML encoding and figure explanation texts.

1.2 Named Entity Recognition

The named entity identifier takes sentence-split, POS-tagged sentences as input. It first applies a statistical named entity recognizer to the input. The statistical recognizer was trained on the data provided by the JNLPBA [5] shared task for named entity recognition. The named entity recognizer outputs marginal distributions of the probability that a substring of the sentence is a protein name. The substrings that have probabilities above some threshold are taken as protein candidates. Then such candidates are mapped to dictionary items whose string edit distance from the candidates are less than some threshold. IDs were taken from Uniprot augmented with the GENA dictionary[6].

When a name is ambiguous, a MaxEnt classifier is used to rank the candidate IDs. The classifier is trained on 296 articles from the training data, using the following features: Similarity between the target article and the MEDLINE articles which is referred to by the Uniprot entry; Similarity between the target article and the MEDLINE articles which include the organism name which is specified by the Uniprot entry; Source dictionary (Uniprot/GENA); Edit distance of the dictionary item and the target string; and Type of the dictionary item (e.g. protein name, gene name, etc.). Similarity was estimated by the cosine measure of the articles represented by tf-idf vectors. Probabilities assigned to each ID by the MaxEnt classifier are output and used by the filtering module (see end of 1.3.1).

1.3 AKANE System

For doing the actual protein pair extraction, the AKANE system [7] was used. It requires AImed Corpus style [2] input for training, so a preprocessor was made to automatically create this kind of co-occurrence sentence collection for the interacting proteins.

The AKANE system parses the input text using the Enju HPSG parser for bio-English. Although the parser has been trained with newswire articles, i.e. Penn Treebank, it can compute accurate analyses of biomedical texts owing to our method for domain adaptation, using the GENIA Treebank [4] to adjust the parsing model. The evaluated bio-performance is 86.9 F-score [3]. The AKANE system combines the output from the parser with the protein pair info from NER, to create the smallest connected parse tree (raw pattern) that covers both proteins. Extra new patterns are also generated by recombining the parts of the raw patterns. Then, counting is done on the training corpus, to evaluate how accurate the patterns are in predicting (only) true interactions. The output from AKANE system lists all possible interactions, so for one mention of an interaction in the text, several interactions are suggested. This is because each protein name is usually ambiguous among several candidate protein IDs, so a postprocessor was made to pick (only) the most likely interaction pair, based on NER probabilities. This is better explained at the end of the following subsection, about pre- and post processing.

1.3.1 Training Data Generation and Pair Filtering based on NER scores

All sentences containing two or more proteins from the PPIs given in the training data files from BioCreative were extracted, and transformed into an AImed style XML marked-up corpus that could be used to train the AKANE system. We assumed that all sentences with a co-occurrence of two interacting (according to the training data from BioCreative) proteins really were describing that

interaction. The accuracy of this assumption, and the effect it had on the prediction phase, was not properly measured (due to lack of time), but some manual inspection of the created corpus indicated reasonable accuracy. Another problem was that only 250 of the total 740 training articles could be used for training. The reason for this is that the AKANE system did not scale well to the large amount of text, compared to the much smaller AImed corpus. So we decided to use only the co-occurrence sentences, and only from the articles where all interacting protein names/IDs could be recognized by NER. Some articles with too many co-occurrence sentences were also dropped, because of the scalability bug in our system.

In order to deal with ambiguity, only the single most likely protein ID were picked from any fragment of ambiguous text, and only the 20 most likely PPI pairs (based on multiplying the NER probabilities) for each article were reported. In run number 2 and 3, a filter was made to remove all pairs that did not have identical species tags in the last part of their protein identifiers. For example, a suggested interaction between P19235 (epor_human) and Q62225 (cish_mouse) is filtered away.

2 Results and Discussion

The three runs were made as follows: Run1 is a version of the system not using the inter-species interaction filter. It achieved an overall F-score of 10.5 (P:8.2% and R:14.6%). Run2 was the best run in terms of F-score based on the training set. On the test data it achieved an overall F-score of 13.7 (P:10.6% and R:19.1%). Run3 was the original AKANE system, trained with the AImed corpus, and optimized for best F-score on the training set. We did not have time to use the machine learning component of AKANE system (F-score 57.3), so instead we used manually tuned parameters and a threshold value reported to achieve 42.0 F-score on AImed (P:70% and R:30%). Still, in the evaluation, Run3 was actually the best one, with an overall F-score of 15.8 (P:15.7% and R:15.9%). This means that training on full text co-occurrence training sentences did not perform any better than training on AImed abstracts alone. The reason for this is that the automatic generation of the training corpus included some noise, in terms of "interacting" co-occurrence like the sentence: *A and B were bought from Santa Cruz inc.*

References

[1] Berger A.L., Pietra S.D., and Pietra V.J.D., A maximum entropy approach to natural language processing, *Computational Linguistics*, 22(1):39–71, 1996.

[2] Bunescu R.C. and Mooney R.J., Subsequence kernels for relation extraction, in *NIPS*, 2005.

[3] Hara T., Miyao Y., and Tsujii J., Adapting a probabilistic disambiguation model of an HPSG parser to a new domain, in *IJCNLP 2005*, vol. 3651 of *LNAI*, 199–210, Springer-Verlag, Jeju Island, Korea, October 2005.

[4] Kim J.D., Ohta T., Tateishi Y., and Tsujii J., GENIA corpus - a semantically annotated corpus for bio-textmining, *Bioinformatics*, 19(suppl. 1):i180–i182, 2003.

[5] Kim J.D., Ohta T., Tsuruoka Y., Tateishi Y., and Collier N., Introduction to the bio-entity recognition task at JNLPBA, in *Proceedings of the JNLPBA-04*, 70–75, Geneva, Switzerland, 2004.

[6] Koike A. and Takagi T., Gene/protein/family name recognition in biomedical literature, in *Proc. Biolink 2004*, 9–16, 2004.

[7] Yakushiji A., *Relation Information Extraction Using Deep Syntactic Analysis*, Ph.D. thesis, University of Tokyo, 2006.