Testing Extensive Use of NER tools in Article Classification and a Statistical Approach for Method Interaction Extraction in the Protein-Protein Interaction Literature

Anália Lourenço[1] , Michael Conover[2,3] , Andrew Wong[4] , Fengxia Pan[5] , Alaa Abi-Haidar[2,3] , Azadeh Nematzadeh[2,3] , Hagit Shatkay*[6] , Luis M. Rocha*[2,3]

[1]IBB/CEB, University of Minho, Campus Gualtar, Braga, Portugal
[2]School of Informatics and Computing, Indiana University, USA
[3]FLAD Computational Biology Collaboratorium, Instituto Gulbenkian de Ciência, Portugal
[4] School of Computing, Queen's University, Kingston, ON, Canada
[5] Microsoft Corp., Redmond, WA, USA
[6] Dept. of Computer and Information Sciences, University of Delaware, USA

Email: Anália Lourenço - analia@deb.uminho.pt; Michael Conover - midconov@indiana.edu; Andrew Wong - 3aw14@queensu.ca; Fengxia Pan - fepan@microsoft.com; Alaa Abi-Haidar - aabihaid@indiana.edu; Azadeh Nematzadeh - azadnema@indiana.edu; Hagit Shatkay*- shatkay@cis.udel.edu; Luis M. Rocha*- rocha@indiana.edu;

*Corresponding author

Abstract

We participated (as Team 81) in the *Article Classification* (ACT) and *Interaction Method* (IMT) subtasks of the *Protein-Protein Interaction* task of the Biocreative III Challenge. For the ACT we pursued an extensive testing of available Named Entity Recognition (NER) tools, and used the most promising ones to extend our the *Variable Trigonometric Threshold* (VTT) linear classifier we successfully used in BioCreative II and II.5. Our main goal was to exploit the power of available NER tools to aid in the document classification of documents relevant for Protein-Protein Interaction. We also used a Support Vector Machine Classifier on NER features for comparison purposes. For the IMT, we experimented with a primarily statistical approach, as opposed to a deeper natural language processing strategy; in a nutshell, we exploited classifiers, simple pattern matching, and ranking of candidate matches using statistical considerations. We will also report on our efforts to integrate our IMT method sentence classifier into our ACT pipeline.

Article Classification Task

We participated in both the online submission with our own annotation server implementing the VTT algorithm via the BioCreative MetaServer platform, as well as the offline component of the Challenge. We used three distinct classifiers: (1) the lightweight *Variable Trigonometric Threshold* (VTT) linear classifier that employs word-pair textual features and protein counts extracted using the ABNER tool [1], and which we successfully introduced in the abstract classification task of BioCreative II [2] as well as on the full-text scenario of Biocreative II.5 [3], (2) a novel version of VTT that includes various NER features as well as various

sources of textual features, and (3) a Suport Vector Machine (using SVM^{light}) that takes as features various entity count features from the NER tools we tested.

In the novel version of VTT that included various NER features, a document d is considered to be relevant if:

$$M.\sum_{f=1}^{F}\frac{P_f(d)}{N_f(d)} \geq \lambda_0 + \sum_{\pi=1}^{EP}\frac{\beta_\pi - n_\pi(d)}{\beta_\pi} - \sum_{\nu=1}^{EN}\frac{\beta_\nu - n_\nu(d)}{\beta_\nu} \quad (1)$$

where λ_0 is a constant threshold for deciding whether a document is positive/relevant or negative/irrelevant. $P_f(d)$ and $N_f(d)$ are occurrence counts of discriminative features (see [3] for details) for feature set f. These features can be textual features (such as bigrams) or features from entity recognition tools. EP is the number of entity count features, π, correlated with relevant documents, and EN is the number of entity count features, ν, correlated with irrelevant documents; $M = EN + EP$.

In addition to testing the power of available NER tools to aid in the document classification of documents relevant for Protein-Protein Interaction, we were interested in answering a few other questions: (1) is there a benefit to using word bigrams as textual features, in comparison to the simpler word-pairs we previously employed [2, 3]? (2) Is it advantageous to use additional PPI classification data from previous Biocreative challenges, or is it best to use only Biocreative III data? (3) how much, if at all, does full-text data help on the classification? Given the time limitations of the challenge, the submitted runs will only allow us to respond to our main question (the utility of existing NER tools) and additional question (1) above. We intend to test questions (2) and (3) post-challenge.

Towards responding to our main question, we utilized the following NER tools and dictionaries: ABNER [1], NLProt, Oscar 3, CHEBI (Chemical names), PSI-MI, MeSH terms, and BRENDA enzyme names. With each one of these tools, we extracted various types of features in abstracts and in figure and table captions. We then computed occurrence counts of the various feature types, for instance: Number of protein mentions in an abstract identified by ABNER, or PSI-MI method mentions in figure captions. Finally, we selected those *entity feature counts* that best discriminated relevant and irrelevant documents in the training and development data. This was done via the analysis of charts such as those described in Figure 1, which depicts a comparison of the counts of ABNER protein mentions in abstracts and BRENDA enzyme names in figure captions on Biocreative III training data (excluding development data). As can be seen, the counts of BRENDA enzyme name mentions in figure captions of documents in the training data does not discriminate well between relevant and irrelevant documents. In contrast, counts of ABNER protein mentions in abstracts are distinct for relevant and irrelevant documents. We used this type of plot to identify which features from NER tools and which document portions behaved differently for relevant and irrelevant documents. For our extended VTT classifier, we used the following five entity feature counts: ABNER protein mentions in abstracts, NLProt protein mentions in abstracts, PSI-MI methods in abstracts, ABNER protein mentions in figure captions, and Oscar compound names in figure captions—which were all positively correlated with relevant documents (therefore $EN = 0$ and $M = EN = 5$ in equation 1) We rejected many other entity feature counts, but provide the community with our feasibility study of the various NER Tools as aids for PPI-relevance article classification. Moreover, we used all entity count features to a SVM classifier to understand the performance of those features alone in classifying PPI-relevant documents.

The Interaction Method Task

We note that the BioCreative training set consisted of full-text articles along with the identifiers of the PPI detection methods that were judged to be discussed in them, *without* any tagging of the sentences that formed the actual evidence for the method. Hence the training corpus could not be used to directly train a classifier to identify PPI method sentences. To make up for this shortcoming, we used a corpus that was developed independently and used in a previous work by Shatkay et al [4]. In that work, Support Vector Machine (SVM) and Maximum Entropy classifiers were trained using a corpus of 10,000 sentences from full-text biomedical articles, which were tagged at the sentence-fragment level, along five dimensions: *focus* (methodological, scientific or generic), *type of evidence* (experimental, reference, and a few other types), *level of confidence* (from 0 - no confidence, to 3 - absolute certainty), *polar-*

ity (affirmative or negative statement), and *direction* (e.g. up-regulation vs. down-regulation). Notably, that corpus had little or nothing to do with protein-protein interaction, but a classifier trained on the *Focus* dimension showed high sensitivity and specificity in identifying *Methods sentences*, and as such we have used it without any retraining. We also used classifiers trained to tag text along the other dimensions, but as almost all sentences were of affirmative polarity and high confidence, we decided to use only the Focus classifier (particularly, whether or not a sentence was classified as a *Methodology* sentence). Using the converted text files provided by BioCreative, we applied a simple strategy for breaking the corpus into sentences based on a modified version of the Lingua-EN-Sentence Perl module [5]), and eliminated any text segment that looked like a bibliographic reference using a simple rule-based strategy. The remaining sentences were converted into a simple binary term-vector representation, for the purpose of classifying each sentence by Focus, utilizing a SVM classifier [4]. This classification step did not identify which method is discussed; rather, it only identifies candidate sentences that may discuss methods.

The specific Method Identifiers (MIs) were then associated with sentences by simple pattern-matching to PSI-MI ontology terms (the primary name and synonyms characterizing each concept)), loaded using the *OBO::Parser::OBOParser* Perl module (part of *ONTO-Per*l package [6]). To allow, to some extent, partial matches, and shuffling of word-order in matches we used two Perl modules: *Text::Ngramize* [7], and *Text::RewriteRules* [8]. The module *Lingua::StopWords* was used to avoid the matching of common English words [9]. As such simple pattern matching can lead to many spurious matches, we scored matches such that exact matches are scored higher than partial ones and longer matches score higher than shorter ones.

Each sentence was thus tentatively associated with all the MIs whose terms hit the sentence. Statistical considerations were used to post-process this many-to-many mapping, selecting one MI among multiple MIS that hit the same sentence, while selecting a single sentence as evidence for each matched MI. Employing several scoring schemes similar in spirit to TF*IDF, we scored each sentence for each candidate MI, based on the length of the match (the higher the better), how rare or frequent the matched terms were in the corpus, in the sentence, and in the methods ontology (rare terms score higher, frequent - lower), and increasing the score for sentences that were classified as *Methodology* in the classification step described earlier. The MIs that scored the highest were reported, and the sentence that gave rise to the score was provided as evidence. The different runs we have submitted varied in the scoring methods used, and in the thresholds placed over the scores to select the MIs that were actually reported.

Integrating the ACT and IMT pipelines

While we were unable to integrate both pipelines for an ACT submission, we are working post-challenge to utilize the output of our IMT pipeline as additional entity features in our ACT pipeline. We will report on this development at the Biocreative III workshop.

Authors contributions

Anália Lourenço was responsible for all NER tool extraction for the ACT and participated in the development and testing of the IMT pipeline. Michael Conover was responsible for the processing of all NER entity features in the ACT, design and implementation of data model for ACT pipeline, analysis of entity features, and design and implementation of SVM classifier. Andrew Wong participated in the development and testing of the IMT pipeline. Fengxia Pan developed, implemented, trained and tested the classifiers which were used in the IMT pipeline, as part of her MSc work at the School of Computing, Queen's University, Kingston, Ontario. Alaa Abi-Haidar produced the code necessary for implementing the VTT method. Azadeh Nematzadeh produced code to extract textual features from documents. Hagit Shatkay developed the methodology and experimental set up for the IMT. Luis M. Rocha developed the methodology and experimental set up for the ACT.

Acknowledgements

Andrew Wong's work was supported by NSERC Discovery Grant, NSERC Discovery Acceleration Supplement, and Ontario's Early Researcher Award, awarded to Hagit Shatkay. Michael Conover, Alaa Abi-Haidar and Azadeh Nematzadeh were supported with a grant from the FLAD Computational Biology Collaboratorium

at the Instituto Gulbenkian de Ciencia in Oeiras, Portugal. We also thank support from this grant for travel, hosting and providing facilities used to conduct part of this research. We thank Artemy Kolchinsky for assistance in setting up the online server for the ACT.

References

1. Settles B: **ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text**. *Bioinformatics* 2005, **21**(14):3191–3192.
2. Abi-Haidar A, Kaur1 J, Maguitman A, Radivojac P, Retchsteiner A, Verspoor K, Wang Z, Rocha LM: **Uncovering protein interaction in abstracts and text using a novel linear model and word proximity networks**. *Genome Biology* 2008, :9(Suppl 2):S11, [http://genomebiology.com/2008/9/s2/S11/abstract/].
3. Kolchinsky A, Abi-Haidar A, Kaur J, Hamed AA, Rocha LM: **Classification of Protein-Protein Interaction Full-Text Documents Using Text and Citation Network Features**. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2010, **7**:400–411.
4. Shatkay H, Pan F, Rzhetsky A, Wilbur WJ: **Multi-Dimensional Classification of Biomedical Text: Toward Automated, Practical Provision of High-Utility Text to Diverse Users**. *Bioinformatics* 2008, **24**(18):2086–2093.
5. **CPAN module, Lingua-EN-Sentence**[http://search.cpan.org/~shlomoy/Lingua-EN-Sentence-0.25/lib/Lingua/EN/Sentence.pm].
6. **CPAN module, ONTO-PERL**[http://search.cpan.org/~easr/ONTO-PERL-1.23/].
7. **CPAN module, Text-Ngramize**[http://search.cpan.org/~kubina/Text-Ngramize-1.03/lib/Text/Ngramize.pm].
8. **CPAN module, Text-RewriteRules**[http://search.cpan.org/~ambs/Text-RewriteRules-0.23/lib/Text/RewriteRules.pm].
9. **CPAN module, Lingua::StopWords**[http://search.cpan.org/dist/Lingua-StopWords/].

Figures

Figure 1 - Entity Counts Analysis

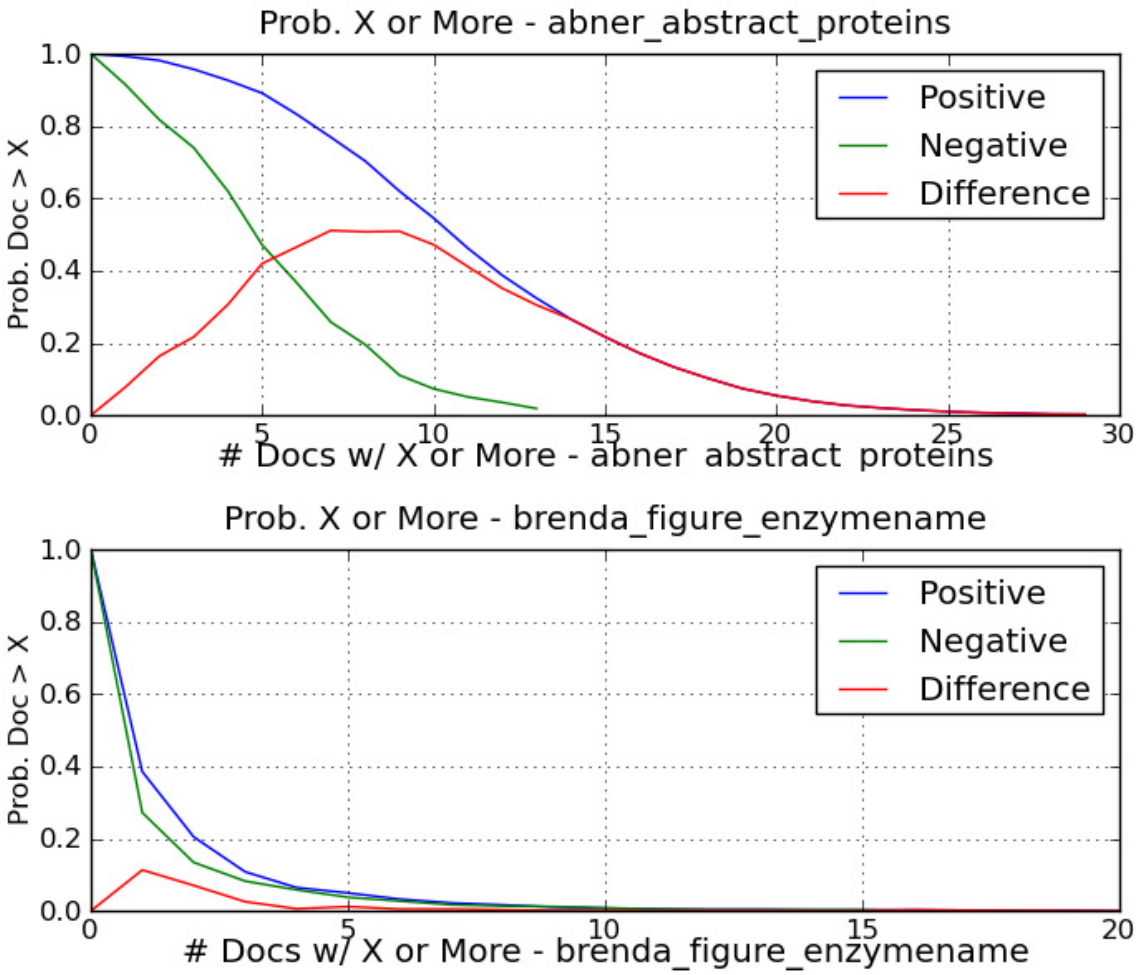

Figure 1: Comparison of the counts of protein mentions as identified by ABNER in abstracts of the articles (top), and BRENDA enzyme names in figure captions (bottom). Results shown for iocreative III training data (excluding development data). The horizontal axis represents the number of mentions x, and the vertical axis the probability $p(x)$ of documents with at least x mentions. The blue lines denote documents labeled relevant, while the green lines denote documents labeled irrelevant; the red lines denote the difference between blue and red lines.