Improving Biomedical Text Categorisation with NLP

Michael Matthews
Informatics Department, University of Edinburgh, Buccleuch Place,Edinburgh, UK
Email: m.matthews@ed.ac.uk;

Abstract

Background: Text categorisation has been used in bioinformatics to help identify documents containing protein-protein interactions. Standard text categorisation methods have used the bag-of-words approach with little input from NLP. While this has proved effective in the past, there is some evidence that the techniques are not adequate in some biological domains. Here we examine how chunking, named-entity recognition and relationship extraction can be combined with traditional text categorisation techniques to improve the classification of documents containing protein-protein interactions.

Conclusions: A system that combines the output of an NLP system with the standard techniques of text categorisation can produce results that exceed the performance of either system on its own. The F_1 of a system that combined features of an NLP system with standard text categorisation features was 68.1 compared with 62.0 using text categorisation alone and 61.9 using relationship extraction alone.

1 Background

1.1 Introduction

Automatic text categorisation has been used in the biomedical domain to help find documents that contain protein-protein interactions. Typically, text categorisation largely ignores the structure and semantics of a document and rather treats a document as a bag-of-words. There is some evidence, however, that standard techniques are not always sufficient. Therefore it is worth considering how NLP can possibly improve performance. In particular, we examine how chunking, named-entity recognition (NER), and relationship extraction (RE) can be used to improve text categorisation.

1.2 Previous Work

In text categorisation, a document is typically represented as a vector of the words in the document with associated weights. The words are frequently stemmed using the porter stemmer and words in a stop word

list are often removed. The weight is usually some function of the number of times the word occurs in the document [1]. A subset of all of the features is often selected using some metric, most commonly information gain [2]. Finally, given a training corpus of documents which have each been marked with their true class, a model can be created which predicts the most likely class of a document given the document representation. There are many different classification techniques, some of the most common being naive Bayes [3], SVMs [4] and KNN [1]. Many of these techniques have been successfully applied in the bioinformatics domain. [5] use a Bayesian approach while both [6] and [7] use SVMs to classify documents containing protein-protein interactions. However, in the KDD Challenge 2002 [8], Regev et al [9] found evidence that in at least that specific classification task, the identification of papers suitable for FlyBase gene-expression database curation, information extraction techniques were more suitable than classic text categorisation techniques. Information extraction is a subset of NLP which covers a broad range of techniques for processing human language with computers with the general goal of extracting information from text with minimal human interaction. NLP has been used extensively in bioinformatics [10], particularly in NER to identify proteins and other biological entities [11] and in RE to extract protein-protein interactions [12, 13].

2 Data and Methods

The experiments were run on a set of 2025 PubMed abstracts that were all analysed to determine if they contained protein-protein interactions as part of the Text Mining programme (TXM). Abstracts containing protein-protein interactions were considered curatable while documents not containing protein-protein interactions were considered not-curatable. In all, 467 documents were found to be curatable and 1558 not-curatable. The collection was split 64% for training, 16% for heldout testing and 20% for testing with each set having the same proportion of curatable and not-curatable documents. Each document was processed with an NLP pipeline consisting of a tokeniser and chunker based on the LTG toolkit [14], a part-of-speech tagger based on the Curran and Clark tagger (C&C) [15] trained on the MedPost data [16], a NER tagger also based on the C&C tagger trained on documents with proteins annotated, and a maximum entropy RE model [17] trained on documents with protein-protein interactions annotated. Results are given for naive Bayes [1] using all features occurring at least 3 times and selecting the top 1500 features based on information gain. Numbers were all converted to the # symbol, punctuation was removed and Greek symbols were converted to their English equivalents. The term frequency is used as the weight for each feature. The prior probability used for the naive Bayes classifier was modified to optimise the F_1 score.

[1] Experiments were also run with SVMs and Maximum Entropy Models, but naive Bayes performed the best. The reasons for this are being studied and may be the subject of a future paper.

Results are given for 10 fold cross validation using the training and heldout sets, for the heldout set when training on the training set and for the test set when training on the train and heldout sets. The test set was not used until the final evaluation.

3 Results and Discussion

The following experiments were run with results reported in Table 1.

Chunks We compare the results of using bigrams as features with those of using the chunks provided by the chunker as features under the hypothesis that the chunks will provide more meaningful groupings of words and thus higher performance.

NER We compare the performance of using the proteins identified using NER as features with the results of using words matching a list of 500,000 proteins names derived from RefSeq as features under the hypothesis that NER will provide a more reliable indication of proteins than a word list.

RE The RE module predicts protein-protein interactions and assigns a probability indicating the confidence of the interaction. This output can be used on its own to classify documents as containing protein-protein interactions or can be used as an additional feature for the text classification system. We compare the results of using standard text categorisation on its own, RE on its own, and results of combining both sets of features. For the combined results, we also experiment with weighting the features by their F_1 calculated as described in [2].

Features	Cross-Validation			Held Out			Test		
	Prec	Rec	F_1	Prec	Rec	F_1	Prec	Rec	F_1
Chunk	54.4	64.0	58.6	57.0	60.8	58.8	55.0	71.0	62.0
Bigram	53.6	62.2	57.4	55.8	58.1	57.0	55.8	67.7	61.2
NER	54.0	69.4	60.5	54.1	71.6	61.6	57.7	76.3	65.7
Protein List	53.4	65.1	58.4	54.8	62.2	58.2	53.5	73.1	61.8
Text Categorisation Alone	54.4	64.0	58.6	57.0	60.8	58.8	55.0	71.0	62.0
RE Alone	54.8	71.9	62.1	59.0	66.2	62.4	55.6	69.9	61.9
RE Combined Simple	56.4	73.4	63.6	54.0	63.5	58.4	57.5	74.2	64.8
RE Combined F_1 Weighting	59.6	79.7	68.0	62.1	79.7	69.8	59.1	80.6	68.2

Table 1: Experimental Results

4 Conclusions

The chunker appears to provide a slight advantage over simple bigrams and NER provides an improvement over using a gazetteer. Not surprisingly, RE provides the greatest improvement. Adding features derived

from the output of the RE module to a text classification system and weighting the features in proportion to their individual F_1 scores resulted in an improvement of 10% over using either RE or text categorisation alone. Thus a system that combines the output of an NLP system with standard techniques of text categorisation can produce results that exceed the performance of either system on its own.

5 Acknowledgements

The work reported here was supported by the ITI Life Sciences Text Mining programme (www.itilifesciences.com).

References

1. Sebastiani F: **Machine learning in automated text categorization**. *ACM Computing Surveys* 2002, **34**:1–47.
2. Forman G: **An extensive empirical study of feature selection metrics for text classification**. *J. Mach. Learn. Res.* 2003, **3**:1289–1305.
3. McCallum A, Nigam K: **A comparison of event models for Naive Bayes text classification**. In *AAAI-98 Workshop on Learning for Text Categorization* 1998.
4. Joachims T: **Text categorization with support vector machines: learning with many relevant features**. In *Proceedings of ECML-98, 10th European Conference on Machine Learning*. Edited by Nédellec C, Rouveirol C, Chemnitz, DE: Springer Verlag, Heidelberg, DE 1998:137–142.
5. Marcotte E, Xenarios I, Eisenberg D: **Mining literature for protein-protein interactions.** *Bioinformatics* 2001, **17**:259–363.
6. Polavarapu N, Navathe SB, Ramnarayanan R, ul Haque A, Sahay S, Liu Y: **Investigation into Biomedical Literature Classification Using Support Vector Machines.** In *CSB* 2005:366–374.
7. Donaldson I, Martin J, de Bruijn B, Wolting C, Lay V, Tuekam B, Zhang S, Baskin B, Bader GD, Michalickova K, Pawson T, Hogue CWV: **PreBIND and Textomy - mining the biomedical literature for protein-protein interactions using a support vector machine.** *BMC Bioinformatics* 2003, **4**:11.
8. Yeh A, Hirschman L, Morgan A: **Background and overview for KDD Cup 2002 task 1: information extraction from biomedical articles**. *SIGKDD Explor. Newsl.* 2002, **4**(2):87–89.
9. Regev Y, Finkelstein-Landau M, Feldman R, Gorodetsky M, Zheng X, Levy S, Charlab R, Lawrence C, Lippert RA, Zhang Q, Shatkay H: **Rule-based extraction of experimental evidence in the biomedical domain: the KDD Cup 2002 (task 1)**. *SIGKDD Explor. Newsl.* 2002, **4**(2):90–92.
10. Blaschke C, Hirschman L, Valencia A: **Information extraction in molecular biology.** *Briefings in Bioinformatics* 2002, **3**(2):154–165.
11. Finkel J, Dingare S, Manning CD, Nissim M, Alex B, Grover C: **Exploring the boundaries: gene and protein identification in biomedical text**. *BMC Bioinformatics* 2005, **6**:S5.
12. Blaschke C, Valencia A: **The Frame-Based Module of the SUISEKI Information Extraction System**. *IEEE Intelligent Systems* 2002, **17**(2):14–20.
13. Hao Y, Zhu X, Huang M, Li M: **Discovering patterns to extract protein-protein interactions from the literature: part II.** *Bioinformatics* 2005, **21**(15):3294–3300.
14. Grover C, Tobin R: **Rule-Based Chunking and Reusability**. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy 2006.
15. Curran JR, Clark S: **Language Independent NER using a Maximum Entropy Tagger**. In *Proceedings of CoNLL-2003*. Edited by Daelemans W, Osborne M, Edmonton, Canada 2003:164–167.
16. Smith L, Rindflesch T, Wilbur WJ: **MedPost: a part-of-speech tagger for bioMedical text**. *Bioinformatics* 2004, **20**(14).
17. Nielsen LA: **Extracting protein-protein interactions using simple contextual features**. In *BioNLP'06 Linking natural language processing and biology: towards deeper biological literature analysis*, Brooklyn,USA 2006.