Unsupervised Classification of Biomedical Abstracts using Lexical Association*

Jonathon Read, Jonathan Webster, and Alex Chengyu Fang

The Dialogue Systems Group
Department of Chinese, Translation and Linguistics
City University of Hong Kong
Tat Chee Avenue, Kowloon, Hong Kong
{jlread,ctjlw,acfang}@cityu.edu.hk

Abstract. The task of text classification is the assignment of labels that describe texts' characteristics, such as topic, genre or sentiment. Supervised machine learning techniques such as Support Vector Machines or the simple but effective Naïve Bayes have been successfully applied to this task. However, it is not always practical to acquire a sufficient corpus of labelled examples to train these methods. For these cases we describe an unsupervised method for text classification based on two hypotheses. Firstly, we propose that the class of a document may be determined by calculating its constituent features' similarity with prototypical examples of each class. Secondly, we note the importance of class priors in Naïve Bayes classifiers, and hypothesize that class distributions might be estimated using the relative frequency of prototype words. Performing experiments on a corpus of biomedical abstracts with topic information derived from the Medical Subject Headings (MeSH), we investigate the characteristics of the method when used in conjunction with basic, linguistic and knowledge-based features, and find that the performance of the unsupervised method is approximately 80% that of Naïve Bayes. Our research is significant in that it highlights a candidate method with good potential for further improvement when training on unlabelled data.

Keywords: Pointwise mutual information, Text classification, Unsupervised methods

1 Introduction

The RAMCORP project aims to design and construct a telephony-based dialogue system that provides interactive dissemination of knowledge in a variety of domains. In particular, it is intended for use by domain experts who are not native English speakers while engaged in dialogue with other domain experts. While the current focus is with respect to the translation of terminology (Webster *et al.*, 2009), the supporting knowledge base is designed to be easily extensible and interoperable (Fleissner *et al.*, 2010); future iterations of the project will explore dialogue-based information retrieval and question answering in multiple domains. We aim to develop techniques that will automatically construct grammars for automatic speech recognition (ASR) that guide users through topics in various domains. Initially this will employ information automatically generated from articles and terminologies, but eventually we intend for dialogues to be driven by automatically constructed ontologies.

A particular challenge faced by the project is the discrimination of a large number of terms through ASR (Fang *et al.*, 2008); it is desirable to reduce the vocabulary size in order to improve recognition accuracy and promote user satisfaction. One approach to accomplish this is to organise

* The work reported in this paper was supported in part by research grants from the City University of Hong Kong (Project Nos 9610053, 9610126, 7008002, 7002387 and 7002190). We would like to thank the reviewers for their comments.

terms according to topic and first query the user for their desired topic before proceeding with translation. An important step in this process is the classification of documents according to topic, so that the topic of terms may be inferred from the topic of documents in which they typically occur. This paper describes the acquisition of a corpus for use in topic classification experiments, our unsupervised approach to this task, and an evaluation of the efficacy of the approach.

The current domain of experimentation and implementation is that of biomedicine, which is appealing due to the large potential user base and the extensive availability of electronic resources. Of particular interest is the Medical Subject Headings (MeSH), a controlled vocabulary thesaurus describing a hierarchy of concepts and related terminology in biomedicine. Section 2 describes both the MeSH and how we employed the hierarchy it describes to collect a corpus of biomedical article abstracts labelled with topic relevance scores.

While the focus of this paper is with respect to the classification of biomedical articles, the project aims to be readily transferable to other domains. This requirement makes the use of supervised machine learning classifiers problematic and highlights the need for unsupervised learning, as obtaining and labelling training data for each new domain is costly. Instead, we explore an unsupervised method for text classification, wherein the relevance of a text to a particular topic is estimated using the constituent features' degree of association with a small number of manually-selected prototypical features representing each topic. Section 4 presents our classification method in detail.

Section 5 describes experiments that evaluate the performance of the unsupervised method using the biomedical abstract corpus, with respect to both classifying text according to their dominant topic and coring texts according to their relevance to topics. Section 6 discusses these results, presents conclusions and provides indications for future work.

2 Corpus Acquisition

An important resource for experiments in the biomedical domain is the Medical Subject Headings (MeSH) database, a controlled vocabulary thesaurus describing concepts and related terminologies across the breadth of biomedicine[1]. The MeSH represents a gold-standard resource for identifying the topic relevance of both terms (through a hierarchy of terminologies) and biomedical article abstracts (through PubMed, an online information retrieval system). The MeSH maintains four types of terms: descriptors, qualifiers, concepts and supplementary concepts.

Descriptors The MeSH is used to index biomedical articles in PubMed and elsewhere through the application of descriptors. These descriptors form a hierarchy of terms, with 16 classes of terms at the most abstract level. There are 25,588 descriptors, but 13,043 of these occur in more than one position in the hierarchy, and are hence ambiguous to various degrees. For example *protamines* has two parents (namely *non-histone chromosomal proteins* and *nucleoproteins*), both of which are descendants of the root-level class *Chemicals and Drugs*. A more ambiguous example is *toxicogenetics*, which has five parents in three different root-level classes.

Qualifiers The MeSH additional contains 83 subheadings (known as qualifiers) organised into a hierarchy up to three levels deep. Qualifiers may be used in conjunction with descriptors in order to provide specifics about the topic of an article. For instance, an article annotated with the descriptor *liver* and the qualifier *drug effects* is about drug effects on the liver. Qualifiers are subject to certain constraints, such that qualifiers are only applicable to a subset of descriptors.

Concepts Supporting the application of descriptors and qualifiers to articles, the MeSH also provides a term bank to enable the expansion of search queries. The term bank describes 48,953

[1] The MeSH is available for download at `http://www.nlm.nih.gov/mesh/`.

concepts, each concept having an average of 4 synonymous terms and typographical variants. The concepts are also organised according to *is a* relations, from hypernym to hyponym.

Supplementary Concepts The MeSH also provides an additional list of 254,750 supplementary concept records, primarily for detailed descriptions of chemicals and drugs.

From the descriptor hierarchy we collapsed 15 root-level trees into flat sets[2], each representing a distinct topic in biomedicine, with a label derived from the root-level descriptor. Article metadata was obtained from PubMed using the Entrez Programming Utilities[3]. Each PubMed article is annotated with a set of MeSH headings, each composed of one descriptor and, optionally, several qualifiers. Any descriptor or qualifier may be indicated as a major topic of the article. We therefore scored each topic according to the number of descriptors that are members of the topic set, relative to the number of descriptors applied to the article. We weighted descriptors marked as major topics (or those with a major topic qualifier) as three times more indicative of a topic than the other descriptors.[4] Formally, given D, a set of descriptors representing an article and T, a set of descriptors representing a topic, we define the relevance in Equation 1. We say that if the relevance score of a topic is greater than 0.5 then it is the *dominant topic* of that article.

$$relevance_{MeSH}(D,T) = \frac{\sum_{d \in D \bigcap T} weight(d)}{\sum_{d \in D} weight(d)} \quad (1)$$

$$weight(d) = 3 \text{ iff } isMajorTopic(d)\,;\, 1 \text{ otherwise} \quad (2)$$

To collect articles, we first obtained a list of PubMed identifiers by issuing queries for each of the 83 qualifiers. We then selected 1,000,000 (approximately 5%) at random for full article download, ensuring that each article record included an abstract and that it had a dominant topic. This resulted in a corpus containing 108,135,957 word tokens of 1,784,417 lemmatised word types, 44.6% of which occur more than once. Table 1 lists the distribution of articles in terms of relevance score and dominant topic, cross-validated across ten folds. For instance, the mean relevance of all articles to the topic of *anatomy* is 7.73%, whilst *anatomy* is the dominant topic 4.8% of articles (since their relevance score is greater than 0.5).

We performed part-of-speech tagging using the AUTASYS tagger (Fang, 1996) and dependency parsing using the RASP system (Briscoe *et al.*, 2006). Note that these tools are designed for domain-independence; tools specifically designed for the biomedical domain are likely to produce superior results. However, the RAMCORP project is domain-independent, and as we intend to analyse performance across various domains it is important to use a comparable setup.

The AUTASYS tagging system uses a hybrid engine incorporating both probabilistic language modelling techniques and linguistically-inspired heuristics to achieve accuracies of approximately 90%. AUTASYS applies the ICE tag set, which has more than 270 tag-feature combinations, representing perhaps the most detailed automatically-applied grammatical annotation scheme.

The RASP system uses a probabilistic parse selection model to determine syntactic trees and weighted grammatical relations between lexical heads. It achieved a microaveraged F_1 score of 76% on a gold standard bank of dependency relations (Watson *et al.*, 2007).

We employed an n-gram matching strategy to find all instances of entries in the MeSH vocabulary. This process resulted in the identification of 24,521,588 instances of 45,409 types of single and multi-word expressions of terminology (88% of which are non-unique). There is an average of 19.5 instances of 11.9 types per article.

[2] We disregard the *Publication Characteristics* tree as we did not find instances of its descendent descriptors in PubMed articles.

[3] `http://eutils.ncbi.nih.gov/corehtml/query/static/eutils_help.html`.

[4] This is an arbitrary weighting, but is motivated by the need to represent the MeSH MajorTopic attribute.

Table 1: The cross-validated distribution of articles with respect to the relevance and dominance of topics, in percent.

	Topic	Relevance	Dominance
A	Anatomy	7.73	4.80
B	Organisms	11.40	2.69
C	Diseases	17.82	22.67
D	Chemicals/Drugs	33.84	50.60
E	Analytical/Diagnostic/Therapeutic Techniques/Equipment	8.53	5.91
F	Psychiatry/Psychology	2.97	3.77
G	Phenomena/Processes	7.25	2.98
H	Disciplines/Occupations	0.51	0.22
I	Anthropology/Education/Sociology/Social Phenomena	0.54	0.47
J	Technology/Industry/Agriculture	0.25	0.10
K	Humanities	0.10	0.07
L	Information Science	0.79	0.51
M	Named Groups	3.73	0.67
N	Health Care	3.74	4.52
Z	Geographicals	0.79	0.03

3 Text Classification

The task of text classification is the automatic assignment of a label, or labels, to a group of texts. Supervised machine learning techniques can be very effective when applied to text classification. These techniques usually begin with an indexing procedure that represents a document as a vector of features with associated weights, which are then used to induce classifiers (Sebastiani, 2002). Typically, words form the features of the vector, while weights are derived from the frequency or presence of the words. Sebastiani provides an overview of supervised techniques including: probabilistic Naïve Bayes approaches, inductive rule learners, regression methods, on-line methods, Rocchio classifiers, neural networks, example-based learning, support vector machines and committees of classifiers (composed of several of the above). Nigam *et al.* (2000) described a semi-supervised approach wherein a small collection of labelled documents is used to train a classifier using a combination of Expectation-Maximization and Naïve Bayes, which then generates probability estimates for a larger collection of unlabelled documents. The algorithm trains a new classifier based on the probability estimates, and iterates until achieving convergence.

Some previous studies have investigated the use of linguistically-motivated features for supervised text classification. Sebastiani (2002) notes that phrases (whether syntactically- or statistically-motivated) do not tend to offer any significant improvement in classification effectiveness. Moschitti *et al.* (2004) found that performance is also unaffected by identifying proper nouns or disambiguating word senses. Gonçalves *et al.* (2006) investigated the use of part-of-speech information in selecting features, finding that the feature space could be reduced whilst retaining performance levels by selecting only words with certain parts-of-speech. They also found that the most effective parts-of-speech varied across domains.

Other researchers have investigated classification tasks specifically in the biomedical domain. In the context of the classification of chest radiograph reports, Wilcox and Hripcsak (2003) examined how applying the knowledge of domain experts when constructing document representations can improve classification performance. They found that the application of expert knowledge improved performance beyond that observed when using different learning algorithms. Ruch (2006) developed a system to assign headings from the MeSH and the Gene Ontology. The system employed a combination of term matching and document vector space analysis. Lu *et al.* (2006) describe statistical machine learning techniques to enhance text classification performance when

dealing with both sparse data and large numbers of features. Their approach firstly involved representing texts so that words were collated into semantic concepts using the latent Dirichlet allocation method. Secondly, they experimented with augmenting labelled training data with semi-supervised learning, such that manually labelled training data is supplemented by unlabelled data that are most probably positive cases. The results, obtained from classification experiments on texts regarding Mouse Genome Informatics, indicate that the methods are particularly effective in improving the recall of a classifier.

4 Lexical Association for Unsupervised Topic Classification

The underlying technique for text classification employed in our research is based on two hypotheses. The first is derived from Turney's (2003) notion that the sentiment of text (i.e. whether it conveys a generally positive or negative opinion) can be determined by estimating the similarity of its constituents with prototypical examples of positivity (e.g. *excellent*) and negativity (e.g. *poor*). This hypothesis is easily transferable to the classification of documents according to topic providing good-quality prototypes are available. However, whereas Turney's method determined the sentiment of phrases using pointwise mutual information, the derived method we employ is applicable to tasks with several non-specific classes. Furthermore, the similarity method is generic, so as to allow future investigations of other similarity methods.

We estimate the similarity between features and prototypes by measuring their lexical association. Measures of *lexical association* examine first-order similarity between features (Grefenstette, 1994). That is, they determine the similarity of a pair of features by considering how likely they are to co-occur. There are many measures of lexical association, including various likelihood measures and hypothesis tests (Evert, 2004). We employ pointwise mutual information (PMI) to measure lexical association. Church and Hanks (1990) introduced the use of PMI in order to measure word association, in order to discover collocations such as those indicating semantic relations and lexico-syntactic constraints. Yang and Pedersen (1997) selected features for supervised machine learning text classifiers using PMI. Turney (2001) employed PMI measured using hit counts from a Web search engine to answer multiple-choice synonymy questions, achieving an accuracy of 73.8%. Turney (2003) evaluated PMI (also using search engine hit counts) in classifying product reviews by sentiment. PMI can also be effective for sentiment analysis when using cooccurrence frequencies observed in large corpora, may also be employed to automatically acquire entries for a sentiment lexicon, and moderately correlates with strength of sentiment (Read and Carroll, 2009).

The second intuition of our method stems from the importance of the class prior in Naïve Bayes classifiers[5]. However, it is not possible to directly estimate the probability of classes without supervised learning. Instead, we hypothesise that the frequency of prototype words correlates with the frequency of the classes they represent. For instance, in the PubMed corpus described above, the frequency of the prototype 'organism' relative to frequency of all topic prototypes is 3%, hence we apply a prior probability estimate of 0.03 when evaluating texts' relevance to the Organisms topic. We accordingly apply a prior to all topic relevance calculations, based on the relative frequency of a topic's prototypes in the training data.

Formally, given a set of features representing a document, F, its relevance to a topic, t (from a set of topics, T), may be estimated as:

$$relevance_I(F,t) = \frac{\Pr(t_P)}{\sum_{u \in T} \Pr(u_P)} \sum_{f \in F} \max_{p \in t_P} \log \frac{\Pr(f,p)}{\Pr(f)\Pr(p)} \quad (3)$$

where t_P is a set of prototypical features of that topic. Our approach is to generate a vector of estimated relevance scores, $\overrightarrow{e} = (relevance_I(F,t_1), relevance_I(F,t_2), ..., relevance_I(F,t_n))$

[5] We found that the cross-validated accuracy of a Naïve Bayes classifier trained on bags of words from the above data set fell by 18% points when deprived of the class prior information.

for each document. To classify according to dominant topic, we choose the most relevant topic: $t^* = \arg\max_t relevance_I(F,t)$.

5 Experiments

5.1 Set-up

In the following experiments we assess a variety of features for use in the classification method described above. These include basic features supplemented by a stop list of function words (*unigrams*, *bigrams*, *trigrams*), grammatical features (*nouns*, *verbs*, *adjectives*, *adverbs*, singleton and compound nouns (*nouns+CN*)), grammatical *relations*, instances of *terminology* and unions of these feature sets. As the sets vary in size we evaluate a logarithmic range of sizes, $s = (2^4, 2^5, 2^6, ..., 2^{23})$, where features are selected according to descending frequency in the training corpus. We estimate the probability of features using relative-frequencies, with Laplacian smoothing. When estimating joint probabilities we consider all features that occur in the same training example as cooccurring. Prototypical example words are straightforward to derive for topic-based classification; in these experiments we simply use the unigrams and bigrams indicated by the topic labels (see Table 1).

To consider the efficacy of the prototype-frequency estimate of the class prior, we evaluate two versions of our method. The first does not include the prototype frequency-informed class prior estimate and is hereafter referred to as the *unweighted* method. The second is the full function as specified by Equation 3, which we refer to as the *weighted* method.

5.2 Baseline and upper-bound

For a baseline approach we created an estimate vector using the frequency of each topic's *prototype* words, as observed in the training data. To provide an indication of the expected performance upper-bound we compiled a vector using the probability estimates from a supervised Naïve Bayes classifier (see Sebastiani, 2002 for details): $relevance_{NB}(F,t) = \Pr(t) \prod_{f \in F} \Pr(f|t)$ where the values of $\Pr(t)$ and $\Pr(f|t)$ are estimated using relative frequencies of tokens (subject to a stop list of functional words) observed in the training data, with Laplacian smoothing.

5.3 Performance Evaluation

We calculated the correlation of the estimated relevance scores $\overrightarrow{e}$ and true scores $\overrightarrow{t}$ (derived from the PubMed metadata using Equation 1) using Pearson's correlation coefficient (Kenney and Keeping, 1962):

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{e_i - \bar{e}}{s_e} \right) \left(\frac{t_i - \bar{t}}{s_t} \right) \quad (4)$$

where n is the number of topics, $\bar{p}$ and $\bar{t}$ are mean values, and s_p and s_t are the standard deviations. r ranges from 1 (the scores are associated) to -1 (the scores are inversely associated), with 0 indicating there is no relation between the scores. We report the mean correlation over all examples in a test set. We selected the topic with the greatest predicted relevance as the dominant topic and calculated the resulting harmonic mean of precision and recall (F_1). To assess the variability of results we carried out ten-fold cross validation and calculate the standard error (σ). To judge the significance of differences in feature types we used a two-tailed Paired t-test over the results on each fold, taking a significance level of $p < 0.05$.

5.4 Results

Figure 1 depicts the performance of the unweighted and weighted $relevance_I$ methods with respect to predicting the true topic relevance scores as defined by the $relevance_{MeSH}$ method, while Figure 2 shows the performance when choosing dominant classes. In order to promote clarity we

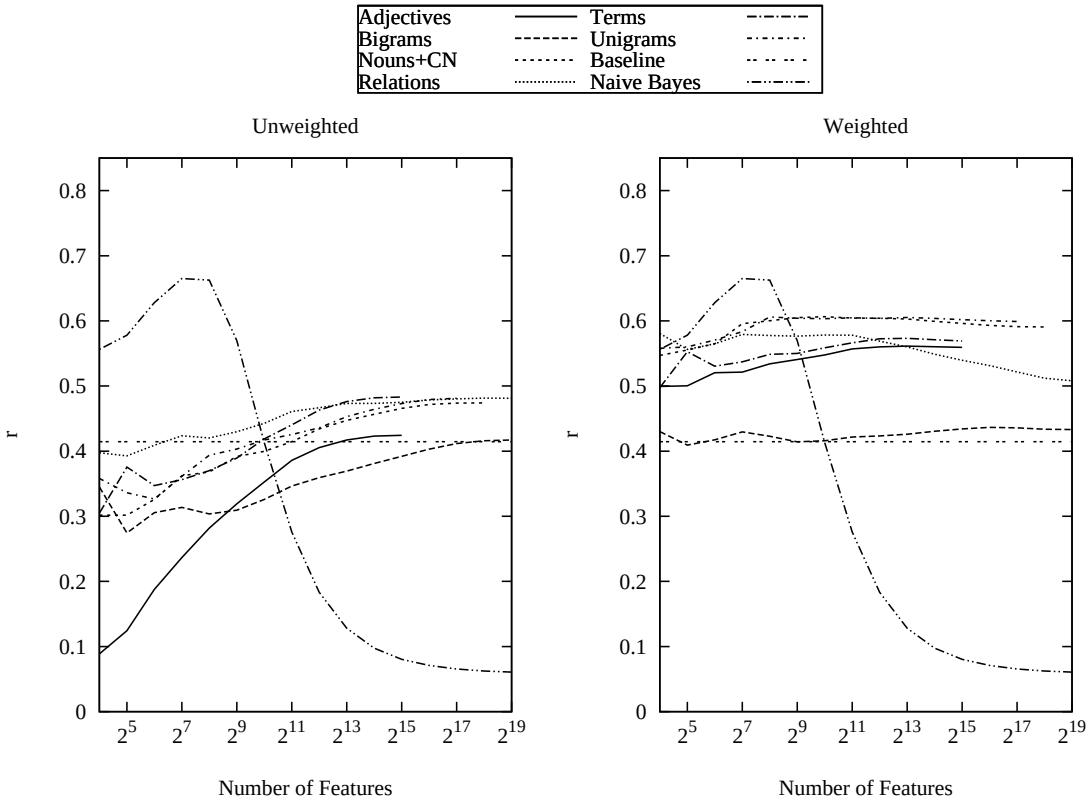

Figure 1: The cross-validated correlation (r) achieved when using different feature types and set sizes with the unweighted and weighted $relevance_I$ measure to predict true topic relevance scores, with supervised performance indicated by Naïve Bayes probability estimates and baselines calculated by counting the frequency of prototypical topic words.

do not depict the performance of verbs, adverbs, trigrams or any of the unified feature sets as they are consistently weaker than other feature sets. We also do not show lemmas or nouns because their performance is consistently similar to that of unigrams (and is very slightly inferior). Standard error bars are also omitted as they are too small to depict clearly (all standard errors are $<= 0.0012$).

In both tasks, all feature types outperform the baseline method given enough features, but are below the supervised upper-bound. When examining the difference in performance of pairs of feature types (using the optimal feature set sizes for a given type), we found that all differences were significant, except for: [relations, unigrams] in the unweighted relevance task and [singleton and compound nouns, unigrams] in the unweighted classification task.

Term features are best for unsupervised prediction of relevance scores in the unweighted method with $r = 0.483$, though nouns+compound nouns, relations and unigrams also achieve similar results. In the unweighted methods all feature types perform best with the greatest feature set size possible (e.g. $32,768$ terms). This situation changes when the class prior estimate is applied as in the weighted version Nouns+CN perform best ($r = 0.606$), and the most effective number of features is considerably lower ($1,024$ nouns/compound nouns).

In the dominant topic classification task, grammatical relations are the best performing feature type in the unweighted method with $F_1 = 0.480$, using the most features of any type ($524,288$). When the weight is applied, however, the best feature type is unigrams ($F_1 = 0.638$) with a set size of $131,072$ features.

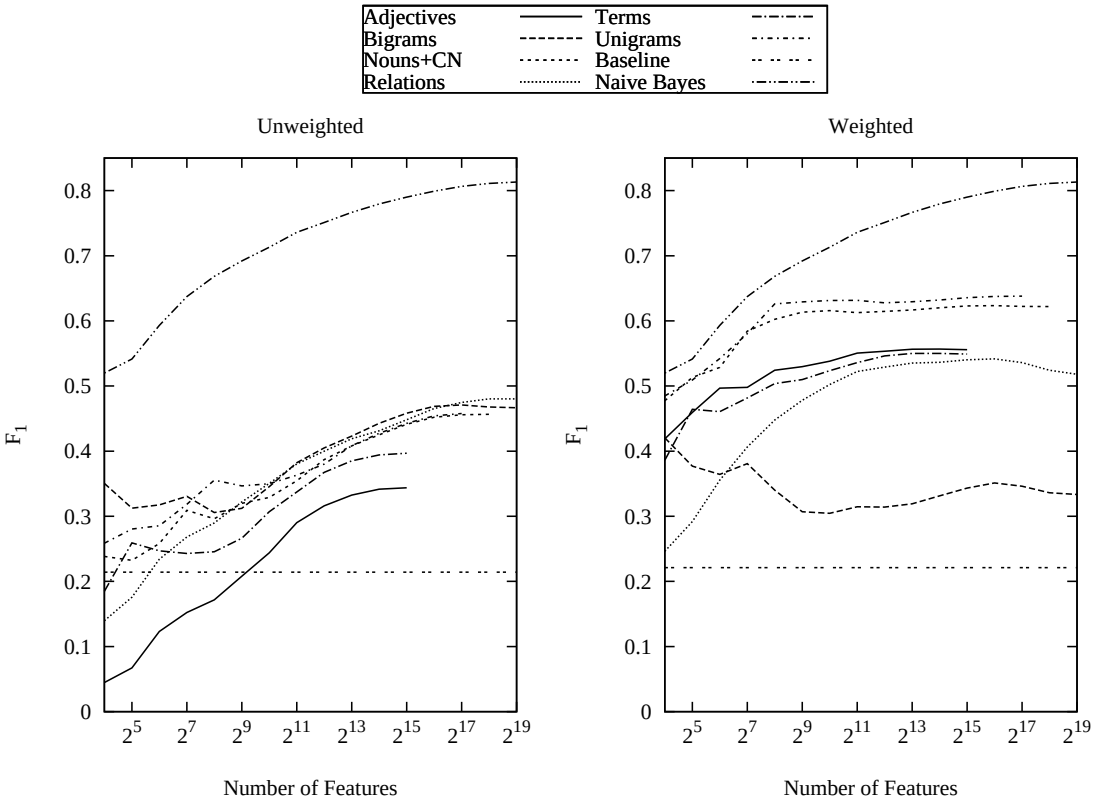

Figure 2: The cross-validated performance (F_1) achieved when using different feature types and set sizes with the unweighted and weighted $relevance_I$ measure to predict dominant topics, with supervised performance indicated by Naïve Bayes probability estimates and baselines calculated by counting the frequency of prototypical topic words.

6 Discussion and Conclusion

The research reported in this paper was motivated by the need to enable rapid adaptation of a telephony terminology translation system to different domains. The need prompted the use of unsupervised classification of topics so that topic-related terms might subsequently be extracted. A novel technique was proposed that uses prototypical words and a measure of lexical association determine documents' relevance to topics. A corpus of PubMed abstracts was created and the experimental results reported here indicate that the prototype similarity method can effectively estimate the relevance of biomedical abstracts according to several topics and classify the documents according to the dominant topic.

A class-prior estimate informed by prototype-frequency is beneficial for most feature types (though experimentation may be necessary in order to determine the most viable feature set size). In the relevance task the weighting resulted in an improvement in correlation of 0.133 for the nouns+compound nouns feature set, and in the classification task the F_1 improved by 0.180 when using unigram features with the prior estimate weighting. In subsequent measurements we found a strong degree of correlation between the prototype frequencies and topic distribution ($r = 0.707$), indicating that prototype frequency can be a useful estimate of topic prior probabilities for unsupervised classifiers. However, some feature types (such as grammatical relations) were not as strongly affected by the weighting, and others (such as bigrams) were negatively affected.

The performance of feature types and set sizes set-ups are considerably different in the relevance and classification tasks. In particular, it seems that smaller feature sets are more effective when determining topic relevance. One possible explanation for this is that the correlation score

is more sensitive to extreme false-positive scoring, whereas the basis of the F_1 measure is binary (i.e. the predicted dominant topic either is correct, or incorrect). Further work will consider other approaches to measuring the accuracy of the prediction vectors.

While in this paper we considered pointwise mutual information as a word similarity estimate, there are a variety of different measures of lexical association. Although it is effective in our task, it may not always be the most effective measure (Yang and Pedersen, 1997). Furthermore, there are alternative strategies for similarity calculation such as considering second-order similarity. Future work will therefore investigate these alternative techniques.

Also, we have considered only one corpus in the biomedical domain. One next step will include an evaluation of inter-domain operability. This will be achieved using two standard text classification data sets: RCV1, a set of newswire articles collected and annotated according to topic, industry and location by Reuters (Lewis *et al.*, 2004), and 20news[6], a collection of articles posted to twenty different Usenet newsgroups. We also plan to compare other approaches for unsupervised text classification, including the use of keywords to retrieve training data for supervised classification (Ko and Seo, 2004) and Intensional Learning (Gliozzo *et al.*, 2005).

To conclude, this paper has shown that measuring documents' constituents similarity with prototypical words can achieve reasonably accurate topic classification. When coupled with an estimate of class prior probabilities (derived from prototype relative-frequency) the performance increases significantly to around 80% of what can be expected by a supervised Naïve Bayes classifier.

References

Briscoe, Ted, John Carroll, and Rebecca Watson. 2006. The second release of the RASP system. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pp. 77–80, Sydney, Australia.

Church, Kenneth W. and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16, 22–29.

Evert, Stefan. 2004. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis, University of Stuttgart.

Fang, Alex Chengyu. 1996. AUTASYS: Automatic tagging and cross-tagset mapping. In Sidney Greenbaum, ed., *Comparing English World Wide: The International Corpus of English*, pp. 110–124. Oxford University Press, Oxford.

Fang, Alex Chengyu, Weigang Li, and Jonathan Webster. 2008. A word-probabilistic interface to dialogue modules. In *Proceedings of the Twelfth Workshop on the Semantics and Pragmatics of Dialogue*, pp. 183–190, London, UK.

Fleissner, Sebastian, Xiaoyue Liu, Jonathan Webster, and Alex Chengyu Fang. 2010. RAMCORP-KB: An interoperable knowledge base for linguistic applications. In *Proceedings of the Second International Conference on Global Interoperability for Language Resources*, pp. 75–84, Hong Kong.

Gliozzo, Alfio, Carlo Strapparava, and Ido Dagan. 2005. Investigating unsupervised learning for text categorization bootstrapping. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pp. 129–135, Vancouver, Canada.

Gonçalves, Teresa, Cassiana Silva, Paulo Quaresma, and Renata Vieira. 2006. Analysing part-of-speech for Portuguese text classification. In *Computational Linguistics and Intelligent*

[6] `http://people.csail.mit.edu/jrennie/20Newsgroups/`

Text Processing, number 3878 in Lecture Notes in Computer Science, pp. 551–562. Springer, Berlin/Heidelberg.

Grefenstette, Gregory. 1994. Corpus-derived first-, second- and third-order word affinities. In *Proceedings of Euralex*, pp. 279–290, Amsterdam.

Kenney, J. F. and E. S. Keeping, 1962. *Mathematics of Statistics, Pt. 1, 3rd ed.*, ch. Linear Regression and Correlation. Van Nostrand, Princeton, NJ.

Ko, Y. and J. Seo. 2004. Learning with unlabeled data for text categorization using bootstrapping and feature projection techniques. In *Proceedings of the Forty-Second Annual Meeting on Association for Computational Linguistics*, Barcelona, Spain.

Lewis, David D., Yiming Yang, Tongy G. Rose, and Fan Li. 2004. RCV1:a new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5, 361–397.

Lu, Xinghua, Bin Zheng, Atulya Velivelii, and ChengXiang Zhai. 2006. Enhancing text categorization with semantic- enriched representation and training data augmentation. *The Journal of the American Medical Informatics Association*, 13(5), 526–535.

Moschitti, Allesandro and Roberto Basili. 2004. Complex linguistic features for text classification: A comprehensive study. In Sharon McDonald and John Tait, eds., *Advances in Information Retrieval*, number 2997 in Lecture Notes in Computer Science. Springer, Berlin/Heidelberg.

Nigam, Kamal, Andrew Kachites McCallum, Sebastian Thurn, and Tom Mitchell. 2000. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39, 103–134.

Read, Jonathon and John Carroll. 2009. Weakly supervised techniques for domain-independent sentiment classification. In *Proceedings of the 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion Measurement*.

Ruch, Patrick. 2006. Automatic assignment of biomedical categories: Toward a generic approach. *Bioinformatics*, 22(6), 658–664.

Sebastiani, Fabrizio. 2002. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1), 1–47.

Turney, P. D. 2001. Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the Twelfth European Conference on Machine Learning*, pp. 419–502.

Turney, Peter D. and Michael L. Littman. 2003. Measuring praise and criticims: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21(4), 315–346.

Watson, Rebecca, Ted Briscoe, and John Carroll. 2007. Semi-supervised training of a statistical parser from unlabeled partially-bracketed data. In *Proceedings of the Tenth International Conference on Parsing Technologies*, pp. 23–32, Prague, Czech Republic.

Webster, Jonathan, Alex Chengyu Fang, Xiaoyue Liu, and Weigang Li. 2009. Multi-factor evaluation of speech recognition for better dialogue system design. In *Proceedings of the Conference of the Pacific Association for Computational Linguistics*, pp. 274–282, Sapporo, Japan.

Wilcox, Adam B. and George Hripcsak. 2003. The role of domain knowledge in automating medical text report classification. *The Journal of the American Medical Informatics Association*, 10, 330–338.

Yang, Y. and J. O. Pedersen. 1997. A comparative study on feature selection in text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning (ICML'97)*, pp. 412–420.