Classifying Disease Outbreak Reports Using N-grams and Semantic Features

Mike Conway, Son Doan, Ai Kawazoe and Nigel Collier
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, Japan
{mike,doan,zoeai,collier}@nii.ac.jp

Abstract

This paper explores the benefits of using n-grams and semantic features for the classification of disease outbreak reports, in the context of a text mining system — BioCaster — that identifies and tracks emerging infectious disease outbreaks from online news. We show that a combination of bag-of-words features, n-grams and semantic features, in conjunction with feature selection, improves classification accuracy at a statistically significant level when compared to previous work. A novel feature of the work reported in this paper is the use of a semantic tagger — the USAS tagger — to generate features.

1 Introduction

Reliable document classification is an important pre-processing stage in many Information Extraction and Text Mining systems (Feldman and Sanger, 2007).[1] This paper compares the performance of a document representation based on highly discriminating unigrams, bigrams, trigrams and semantic features, against a representation based on unigram and Named Entity (NE) features used by Doan et al. (2007), for the classification of disease outbreak reports. While the document representation used by Doan et al. (2007) performed well for this task, a statistically significant improvement in performance was achieved using a representation based around n-grams and semantic features. A novel feature of this work is the use of a general purpose semantic tagger to generate features.

Following a discussion of related work in section 2, we describe in section 3 the feature sets used in this work and how they were derived. Section 4 sets out the methodology used, while section 5 presents results, and some discussion of those results. The final section outlines some broad conclusions and areas for future work.

2 Background

The BioCaster Corpus is a product of a wider project designed to aid in the surveillance and tracking of infectious disease outbreaks using text mining technology. The BioCaster system (Doan et al., 2008) scans online news reports for stories concerning infectious disease outbreaks. An article is of interest if it contains information about newly emerging infectious diseases of potential international significance, such as, the spread of diseases across international borders, the deliberate release of a pathogen, and so on. There are two methods that users can exploit to explore extracted data. First, the pre-interpreted information is publicly available on a web portal (built on Google Maps).[2] Second, registered users can opt to receive information (via email) on diseases, countries or other alerting conditions that interest them. According to Heymann et al. (2001), around 65% of disease outbreaks are first identified from the web.

The BioCaster gold standard corpus is a collection of 1000 news articles selected from the WWW, and subsequently manually categorized and annotated by two PhD students at the National Institute of Informatics (see Figure 1 for

[1] Cohen and Hersh (2005) includes a brief review of important work on text classification in the biomedical domain.

[2] The publicly accessible face of the BioCaster system is a visualization tool called *Global Health Monitor*. It is accessible at the BioCaster Portal (http://www.biocaster.nii.ac.jp).

```
<DOC id="000101" language="en-us"
source="WHO" domain="health"
subdomain="disease" date=2007/3/2
relevancy="publish">
<NAME cl="DISEASE">Avian
Flu</NAME> situation in <NAME
cl="LOCATION">Vietnam</NAME> update
21
 <NAME cl="TIME">16
June 2005</NAME><NAME
cl="ORGANIZATION">WHO</NAME>
is aware of media reports
that <NAME cl="PERSON"
case="true" number="many">six
additional patients</NAME><NAME
cl="CONDITION">infected</NAME>
with <NAME cl="DISEASE">H5N1
avian influenza</NAME> are
undergoing treatment in a <NAME
cl="LOCATION">Hanoi</NAME>
hospital and that <NAME cl="PERSON"
case="true" number="one">a health
care worker</NAME> at the same
hospital may also be <NAME
cl="CONDITION">infected</NAME>.
While these reports have not
yet been officially confirmed by
national authorities, they appear to
be accurate.
 <NAME cl="ORGANIZATION">WHO</NAME>
is seeking confirmation and
further information from the <NAME
cl="ORGANIZATION">Ministry of
Health</NAME>.  </DOC>
```

Figure 1: Example Annotated Entry from the `BioCaster` Corpus

a truncated example, and Kawazoe et al. (2006) for a description of the annotation scheme). The corpus consists of around 290,000 words (excluding annotation). Articles were collected from various online news and non-governmental organization sources, including online news from major newswire publishers.[3] Four *per cent* of the corpus was originally gathered by the International Society for Infectious Diseases, under the ProMED-Mail Programme – a human curated disease outbreak report service.[4] From the perspective of the current work, an important characteristic of the corpus is that each document is classified as belonging to one (and only one) relevancy category with respect to infectious disease outbreaks. There are four categories:

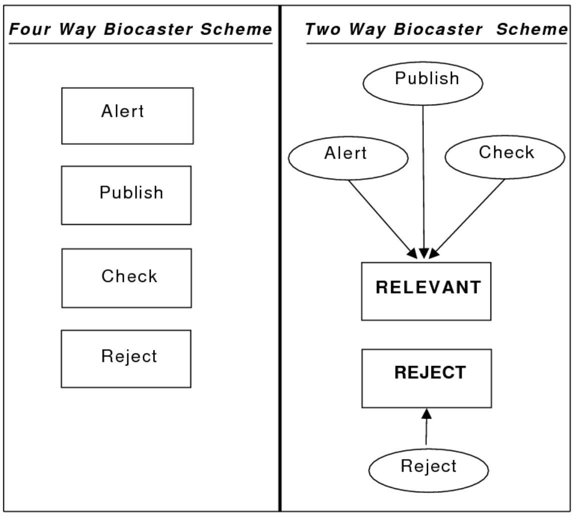


Figure 2: Binary Categories in `BioCaster` Corpus

- **Alert** — News stories tagged “alert” are deemed to be of immediate interest to health professionals.
- **Publish** — News stories tagged “publish” are judged to be of archival importance to health professionals.
- **Check** — News stories tagged “check” are deemed to be of possible interest to health professionals. The category includes suspicious sounding disease outbreak events for which full information is not available.
- **Reject** — News stories tagged “reject” are deemed to be of little or no interest to health professionals.

In situations where annotators disagreed on the class of a document a domain expert was consulted for clarification. All these categories (and guidelines for determining categories) were developed in consultation with the National Institute of Infectious Diseases (Japan) and based on World Health Organization guidelines.[5]

The corpus is composed of news articles from several different domains (see Table 1). Although over half of the documents in the corpus are classified as belonging to the *health* domain, it is important to stress that articles classified as *alert*, *publish* or *check* can also be found in the *business* category (say, the effect of a livestock disease on the agricultural sector) or in the science and technology category. Additionally, an article may be concerned with a specific infectious disease, but not directly concerned with an *out-*

[3] Major sources included the BBC (UK), CBC (Canada), *The Nation* (Thailand), IRIN (United Nations), and the *Sydney Morning Herald*, among others.

[4] `http://www.promedmail.org`

[5] The WHO guidelines can be found at: `www.who.int/gb/ghs/pdf/IHR_IGWG2_ID4-en.pdf`

Domain	Number of Documents
Health	539
Business	173
Society	85
Sport	50
Politics	95
ScienceTech	8
Science	44
Technology	3
Entertainment	3

Table 1: Domains in the BioCaster Corpus

break of that disease. Instead, the article could be about a vaccination campaign or a medical breakthrough. Also, the corpus contains documents which are about serious *non*-infectious diseases, like, for instance, most forms of cancer. These non-infectious disease news stories are marked as *reject*.

In order to create a binary classification scheme, the three categories that can broadly be described as relevant with respect to infectious disease outbreaks (*publish*, *alert* and *check*) were conflated into a single *relevant* category (see Figure 2). The binary corpus consists of 350 *relevant* documents and 650 *non-relevant* documents.

Doan et al. (2007), working on an identical task, points out that a bag-of-words representation struggles to identify biomedically relevant senses of polysemous words like *virus* (computer virus or biological virus) or *control* (control a disease outbreak or control inflation) and proposed the use of NE based semantic features as a possible solution.

The approach outlined in this paper develops the work reported in Doan et al. (2007) for binary classification of the BioCaster corpus. We take Doan et al. (2007)'s work one stage further by employing n-grams, a semantic tagger and feature selection to achieve enhanced classification accuracy.

3 Feature Sets

The text classification community has expended a huge amount of research effort on identifying the most effective features for representing text documents. Yet the simplest and most commonly used text representation — the so-called "bag-of-words" representation where each distinct word in a document collection acts as a feature — has proven stubbornly effective. Lewis (1992) compared simple phrase based features with a bag-of-words representation and found that classification performance deteriorated when more complex features were used. The use of syntactic features was again assessed by Moschitti and Basili (2004), who found "overwhelming evidence" that syntactic features fail to improve topic based classification. Scott and Matwin (1999) in a series of experiments using Reuters news wire data reported that phrase based representations (in this case, noun phrases) failed to improve topic classification compared to bag-of-words, and concluded that, "it is probably no worth pursuing simple phrase based representations any further." Domain sensitive *semantic* representations have however been shown to enhance text representations in some situations (Doan et al., 2007).

Named Entity	Attributes
PERSON	case,number
ORGANIZATION	none
LOCATION	none
TIME	none
DISEASE	none
CONDITION	none
NON-HUMAN	transmission
VIRUS	none
OUTBREAK	none
ANATOMY	transmission
SYMPTOM	non
CONTROL	none
CHEMICAL	therapeutic,transmission
DNA	none
RNA	none
PROTEIN	none

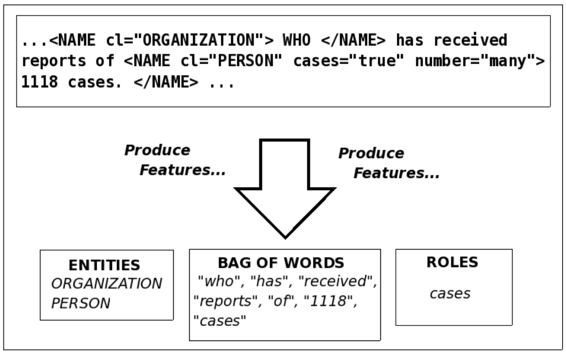
Table 2: Named Entities and Roles in the BioCaster Named Entity Annotation Scheme

3.1 Named Entity Based Features

Doan et al. (2007) used the 18 NE tags (some of which have associated attributes or "roles") in the BioCaster annotation scheme to augment bag-of-words features (see Table 2 for a list of NEs and their associated roles), increasing classification accuracy from 74% accuracy with a bag-of-words representation (BOW) to 84.4 % accuracy with a feature set consisting of BOW plus all NS and all NE attributes (BOW+NE+roles). Figure 3 shows how features were generated from a sentence snippet of the BioCaster corpus.

3.2 N-gram Features

N-grams were used (where $n > 1$) as they may help reduce the problems presented by polysemous words and identify concepts highly char-



...<NAME cl="ORGANIZATION"> WHO </NAME> has received reports of <NAME cl="PERSON" cases="true" number="many"> 1118 cases. </NAME> ...
Produce Features...
Produce Features...
ENTITIES
ORGANIZATION
PERSON
BAG OF WORDS
"who", "has", "received", "reports", "of", "1118", "cases"
ROLES
cases

Figure 3: Generating BOW+NE+roles Features (Based on Doan et al. (2007))

acteristic of disease outbreak reports. The trigram `ministry_of_health` may help identify disease outbreak reports more effectively than its constituent unigrams `ministry`, `of` and `health`. Unigrams were derived from the `BioCaster` corpus itself, whereas bigrams and trigrams were acquired from a larger in-domain corpus of 874,000 words from ProMED-Mail disease outbreak report service. This was used in preference to the `BioCaster` corpus because of its size. Only bigrams and trigrams that occurred at least twice in the ProMED-Mail corpus were retained and used in our document representation.

3.3 USAS Semantic Tagger Features

The semantic tags used in this work were generated using the USAS semantic tagger (Rayson et al., 2004).[6] The USAS tag scheme consists of 21 major discourse categories and 232 fine grained semantic tags and relies heavily on a lexicon to assign semantic classes.[7] Figure 4 shows the twenty-one top level categories.

According to Rayson et al. (2004) assigning a semantic tag is a two stage process. First, assigning a list of *possible* semantic tags to a word. Second, identifying the contextually appropriate sense from the list of *possible* tags. A combination of several different methods are used to disambiguate word senses.

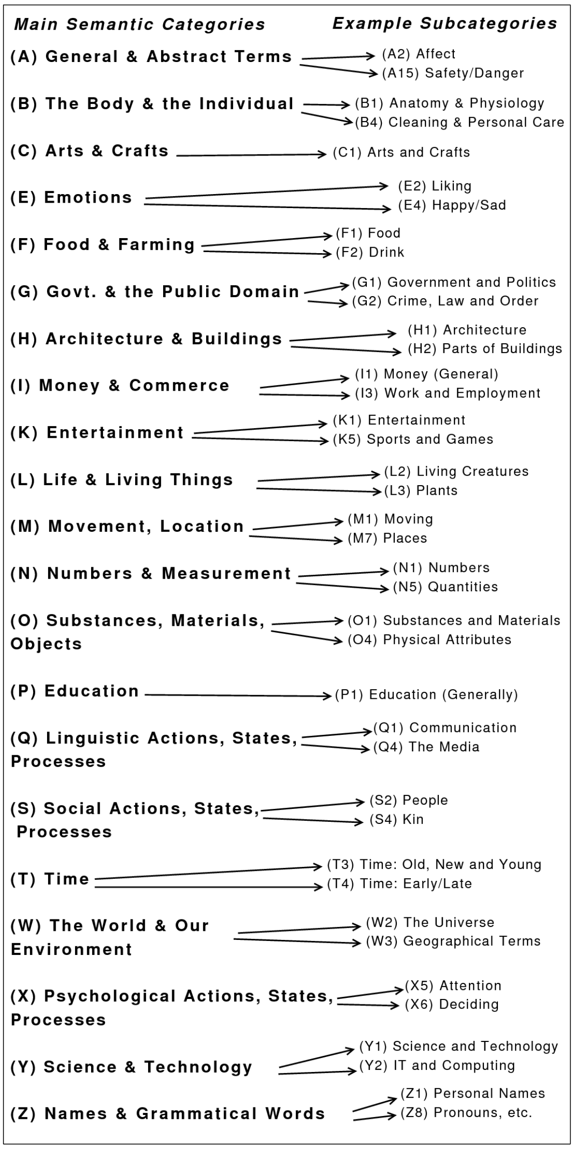


Figure 4: UCREL Semantic Tag Scheme

- FILTER BY POS TAG. For example, "spring" (season) and "spring" (jump) can be disambiguated using their POS tag. One is a temporal noun and the other is a verb.
- GENERAL LIKELIHOOD RANKING. For example, "green" is used more frequently as a colour term rather than meaning "naïve."
- DOMAIN OF DISCOURSE. The domain of discourse can be specified, and this extra information used in assigning semantic tags. For example, in the food domain, "battered" is more likely to refer to the cooking technique, rather than, say violence.
- TEXT-BASED DISAMBIGUATION. Leverages the fact that a word is likely to retain the same sense throughout a given text.
- CONTEXTUAL RULES. Templates are used to identify some senses. For example, if the noun "account" occurs in the pattern "NP ac-

[6] The USAS (UCREL Semantic Analysis System) was developed at the University Centre for Computer Corpus Research on Language (UCREL) at the University of Lancaster. More details of the tagger can be found at: `http://ucrel.lancs.ac.uk/usas/`

[7] The tagset used in the USAS semantic tagger was loosely based on that developed by McArthur (1981).

count of NP" it is likely to be concerned with narrative explanation.

- LOCAL PROBABILISTIC DISAMBIGUATION. Uses local context and collocational information to determine the correct tag. This method is only partially implemented.

The tagger is also designed to identify multi word units (For example, "United States" is tagged as a multiword unit with a geographical tag) using various techniques, but for the purposes of this work, multiword units were ignored. Also, in some instances the tagger presents two tags as joint equal in likelihood. For example, in the sentence, "County health officials said the baby also exposed about 58 children at the Murray Callan Swim **School**, also in Pacific Beach," the highlighted word "**School**" is classified as both *Education in general* and *Architecture: Kinds of Houses and Buildings*. In this kind of situation – where two tags are presented as equally likely, both tags are retained and used in the document representation.

The tagger has previously been embedded in a translation support system for English and Russian (Sharoff et al., 2006), and has been used in the study of the compositionality of multiword expressions (Piao et al., 2006). An important difference between the USAS semantic tagger and other more well known lexical semantic resources, like WORDNET (Fellbaum, 1998) is that the USAS tagger *disambiguates* between word-senses (albeit without 100% accuracy), rather than providing sets of synonyms for each word sense. Like WORDNET, the USAS semantic tagger is designed for general purpose use, rather than specifically built for use in the biomedical domain.[8] However, 7.7% of words in the taggers lexical database (3,511 words from a total of 45,870) do have *the body* or *life and living things* as their primary semantic category. Table 3 shows a breakdown of the number of words for which a biological sense is dominant.

Tag	Tag Gloss	Lexemes
B1	Anatomy & Physiology	756
B2	Health & Disease	25
B3	Medicines & Medical Treatment	348
L1	Life & Living Things	14
L2	Living Creatures Generally	300
L3	Plants	371

Table 3: Biology Related USAS Semantic Tagger Tags

	REL correct	non-REL correct
Assigned REL	a	b
Assigned non-REL	c	d

Table 4: Contingency Table for Calculating Classification Accuracy (REL is "Relevant" and non-REL is "Non-Relevant")

4 Methodology

In all our experiments, we used a binary feature representation. That is, if a feature X occurs at least once in a document, the feature value for X in that document is 1, otherwise the value is 0. This binary representation was used as early experimental work indicated that binary features performed better than weighted features. Three machine learning algorithms were used: Naïve Bayes, Support Vector machines and the C4.5 decision tree algorithm (Witten and Frank, 2005; Mitchell, 1997). The Weka data mining toolkit[9] was used for all the reported machine learning work, and the classification accuracy levels reported (that is, per cent of correctly assigned instances) are the results of 10-fold cross validation. Where statistical significance levels are reported, 10 × 10-fold cross validation is used in conjunction with the corrected resampled t-test as presented in Bouckaert and Frank (2004). Accuracy is the percentage of correctly defined documents (defined as the number of correctly assigned instances divided by the total number of instances). This can easily be calculated from a contingency table (see Table 4) as $accuracy = (a + d)/(a + b + c + d)$.

Feature selection techniques are central to this work. Yang and Pedersen (1997) showed that aggressive feature selection can increase classification accuracy for certain kinds of texts (in their case, newswire articles). Of the various different algorithms tested, they found that χ^2 and information gain proved most effective. Forman (2003) provides a survey of feature selection methods for text classification.

The χ^2 method was used for feature selection[10]

[8] Note that the general purpose biological categories used by the USAS tagger, while appropriate for disease related newspaper texts in the BioCaster corpus, may well be insufficiently fine grained for effectively representing academic papers in the biology domain.

[9] http://www.cs.waikato.ac.nz/ml/weka/

[10] The Weka implementation of the χ^2 feature selection algorithm was used.

Features	No. Features	NB	SVM	C4.5
semtag	580	78.8	82.8	76.9
semtag (comp)	263	78.4	82.87	74.14
unigrams	21322	88.4	90.9	80.8
bigrams	1567	87.6	87.1	83.5
trigrams	2345	82.5	81.1	82.2
BOW+NE+roles	20889	88.4	90.6	82.2
χ^2 (chi-squared)	9000	**94.8**	92.2	81.6

Table 5: Initial Results (Note that "BOW" is "Bag-of-Words")

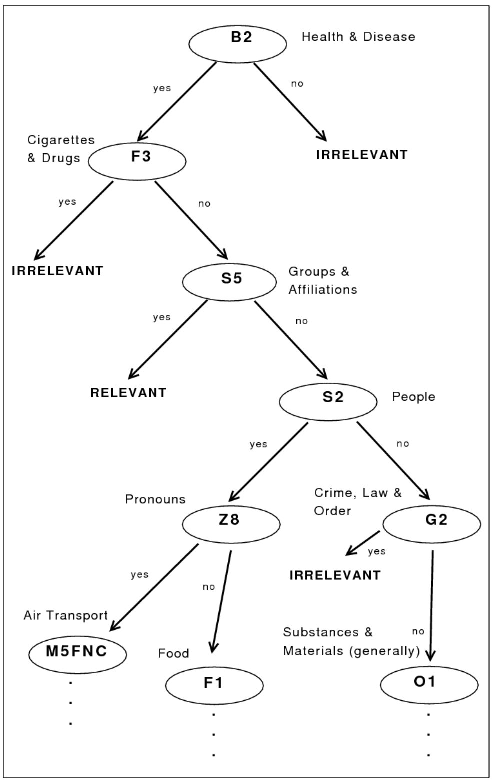


Figure 5: Partial C4.5 Decision Tree for Semantically Tagged `BioCaster` Corpus

as it has shown to be effective in the context of text classification (Yang and Pedersen, 1997). For more on the χ^2 method see Oakes et al. (2001).

5 Results and Discussion

Our chosen baseline in this work is the BOW+NE+roles feature set identified by Doan et al. (2007) using similar data (that is, an earlier, smaller version of the `BioCaster` corpus consisting of 500 documents). This baseline feature set achieved a classification accuracy of 88.4%, the same as the unigram feature set. This was surprising as Doan et al. (2007) found that the BOW+NE+roles achieved higher accuracy than the unigram feature set. These differing results could be accounted for by Doan et al. (2007)'s use of term weighting rather than a binary representation, and the use of a smaller corpus.

Initial comparisons of the several feature representations show that n-gram representations achieved better results than a semantic tag based feature representation. However, a *mixture* of unigrams, bigrams, trigrams and semantic tag features, worked best of all. Table 5 summarizes these initial results. Note that two different document representations based on the USAS semantic tagger were used. The *compressed* representation discarded directionality indicators along a given dimension, and instead used the presence or absence of the dimension itself as a feature. For example, if we take the USAS tag `E2` (Liking), those words tagged `E2+` (like *adore* and *beloved*) and those words tagged `E2-` (like *detest* and *abhor*) will be reduced to one feature (`E2`) reflecting the liking/disliking dimension, although this change had little impact on the results, which are very similar for both of the semantic tagger based representations.

The C4.5 decision tree algorithm seems to perform consistently worse than both the Naïve Bayes and SVM[11] algorithms. One of the advantages of the decision tree algorithm however, is its potential for data exploration purposes. Figure 5 shows the root of a partial decision tree derived from the (full) USAS semantic tag representation of the `BioCaster` corpus. Working from the root of the tree, it can be seen that if the document does not contain any words that are tagged *Health & Disease* then the document is immediately classified as irrelevant (that is, not a disease outbreak report). At the next level, if the document contains a *Cigarettes & Drugs* tag, then the document is classed as irrelevant as diseases *directly* related to cigarettes and non-medicinal drug use are normally chronic rather than highly infectious. The next level down refers to *Groups and Affiliations*, which in the USAS semantic tagger guidelines is described as "Terms relation to groups/the level of association/affiliation between groups,"[12] with prototypical examples like *alliance, caste, community* and so on. The importance of this category for classification accuracy is explained by the inclusion of the word "epidemic" (a strong in-

[11] Default `Weka` parameters were used for the SVM algorithm.

[12] Technical material on the USAS semantic tag scheme can be found at: `http://ucrel.lancs.ac.uk/usas/`

dicator that a document is concerned with disease outbreaks) in the *groups and affiliations* tag.[13]

The best performing representation (94.8% using the Naïve Bayes algorithm – see Table 5) was derived by performing feature selection on *all* the features used (that is, all unigrams, bigrams, trigrams and semantic features). This result was statistically significant when compared to the BOW+NE+roles feature set. Rather than choosing an arbitrary cut off point for feature selection, the optimal number of features was derived experimentally. Figure 6 shows that accuracy peaks at around 9,000 features, and gradually decreases when more features are added.

The 9,000 most powerfully discriminatory features, as determined by the χ^2 method, consist of a mixture of unigrams, bigrams and semantic features, suggesting that a mixed approach to document representation is optimal, rather than relying on a single *type* of feature. Of the one hundred most discriminating features, 50% were unigrams, 37% were bigrams, 8% were trigrams and 5% were semantic tags. As can be seen from Table 6, the two most discriminatory *semantic* features are B2 (health and diseases) and L2 (living creatures), results that are in line with intuitions regarding the subject matter of disease outbreak reports.

Of the 9,000 most discriminating features derived using the χ^2 method, only 130 are semantic tags (<2%), and as semantic tagging is a relatively complex procedure, we investigated the performance of the 9,000 feature set with all 130 semantic features removed, in order to test how much the inclusion of semantic tag features improves accuracy. Running the classifier with the 130 semantic tags removed led to a 0.5% reduction in classification accuracy; not a statistically significant difference.

1	health	16	the outbreak
2	cases	17	case
3	outbreak	18	the ministry
4	confirmed	19	hospital
5	died	20	cases of
6	disease	21	poultry
7	symptom	22	outbreak in
8	reported	23	suspected
9	ministry	24	the ministry of
10	death	25	fever
11	virus	26	h5n1
12	the disease	27	have died
13	of health	28	provinces
14	**B2**	29	**L2**
15	ministry of health	30	the virus

Table 6: Most Discriminating Features in the `BioCaster` Corpus

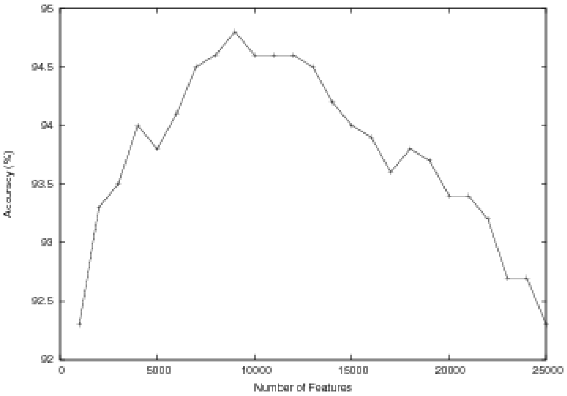


Figure 6: Comparison of Feature Selection Thresholds

6 Conclusion

In conclusion, we have shown that for the classification of disease outbreak reports, a combination of bag-of-words, n-grams and semantic features, in conjunction with feature selection, increases classification accuracy at a statistically significant level compared to a "BOW+NE+roles" representation. A novel feature of this work is the use of a semantic tagger — the USAS semantic tagger — to generate semantically rich features. However, most of the increase in classification accuracy arose from the inclusion of n-grams in the feature set, rather than the USAS tagger derived semantic features. It is possible that the thesaurus derived scheme used by the tagger is insufficiently fine grained to capture some important biological concepts, but that the tagger's ability to disambiguate between potentially polysemous biological words (like "virus") was enough to increase accuracy slightly.

Further work will fall into two broad areas:

- Developing and testing further domain specific semantic features (including adding Doan et al. (2007)'s BOW+NE+roles to the feature selection operation).
- Semantic features derived from the USAS tagger will be considered to enhance other

[13] As stated above, if the semantic tagger's disambiguation mechanisms cannot decide between two tags, both are included in the document representation. For example, "epidemic" counts as both a *Health and Disease* word, and also as a *Groups and Affiliations* word.

modules of the BioCaster text mining system.

Acknowledgments

We would like to express thanks to Dr Paul Rayson, Directer of UCREL (University Centre for Computer Corpus Research on Language) at Lancaster University for providing access to the USAS semantic tagger. This work was funded in part by grants from the Japanese Society for the Promotion of Science (grant no: P07722) and the Research Organization of Information Systems.

References

R. Bouckaert and E. Frank. 2004. Evaluating the Replicability of Significance Tests for Comparing Learning Algorithms. In *Advances in Knowledge Discovery and Data Mining*, pages 3–12. Springer, Berlin.

A. Cohen and W. Hersh. 2005. A Survey of Current Work in Biomedical Text Mining. *Briefings in Bioinformatics*, 6(1):57–71.

S. Doan, Q. Hung-Ngo, A. Kawazoe, and N. Collier. 2008. Global Health Monitor - A Web Based System for Detecting and Mapping Infectious Diseases. *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP08) - Companion Volume*, pages 951–956.

S. Doan, A. Kawazoe, and N. Collier. 2007. The Role of Roles in Classifying Annotated Biomedical Text. *BioNLP 2007: A Workshop of ACL 2007*, pages 17–24.

R. Feldman and J. Sanger. 2007. *The Text Mining Handbook: Advanced Approaches to Analyzing Unstructured Data*. CUP.

C. Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, Mass.

George Forman. 2003. An Extensive Empirical Study of Feature Selection Metrics for Text Classification. *Journal of Machine Learning Research*, 3:1289–1305.

D. Heymann, G. Rodier, and WHO. 2001. Hot spots in a wired world: WHO surveillance of emerging and re-emerging infectious diseases. *The Lancet*, 1(5):345-353.

A. Kawazoe, L. Jin, M. Shigematsu, R. Barrero, K. Taniguchi, and N. Collier. 2006. The Development of a Schema for the Annotation of Terms in the BioCaster Disease Detection/Tracking System. In *Proceedings of the Second International Workshop on Formal Biomedical Knowledge Representation*, pages 77–85.

David D. Lewis. 1992. *Representation and Learning in Information Retrieval*. Ph.D. thesis, Department of Computer Science, University of Massachusetts, Amherst, US.

T. McArthur, editor. 1981. *Longman Lexicon of Contemporary English*. Longman, London.

Tom Mitchell. 1997. *Machine Learning*. McGraw-Hill International, Singapore.

A. Moschitti and R. Basili. 2004. Complex Linguistic Features for Text Classification: A Comprehensive Study. In *Proceedings of the 26th European Conference on Information Retrieval Research*, pages 181–196.

M. Oakes, R. Gaizauskas, H. Fowkes, A. Jonsson, V. Wan, and M. Beaulieu. 2001. Comparison Between a Method Based on the Chi-Square Test and a Support Vector Machine for Document Classification. In *Proceedings of the 24th ACM Special Interest Group on Information Retrieval (SIGIR01)*, pages 440–441.

S. Piao, P. Rayson, O. Mudraya, A. Wilson, and R. Garside. 2006. Measuring MWE Compositionality Using Semantic Annotation. *Proceedings of COLING/ACL 2006 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 2–11.

P. Rayson, D. Archer, S. Piao, and T. McEnery. 2004. The UCREL Semantic Analysis System. *Proceedings of the Workshop on Beyond Named Entity Recognition: Semantic Labeling for NLP Tasks in association with the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, pages 7–12.

Sam Scott and Stan Matwin. 1999. Feature Engineering for Text Classification. In *Proceedings of the 16th International Conference on Machine Learning*, pages 379–388.

S. Sharoff, B. Babych, P. Rayson, P. Mudraya, and S. Piao. 2006. ASSIST: Automatic Semantic Assistance for Translators. *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006)*, pages 139–132.

I.H. Witten and E. Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan-Kaufmann, San Francisco, second edition edition.

Y. Yang and J. Pedersen. 1997. A Comparative Study on Feature Selection in Text Categorization. In *Proceedings of ICML-97, 14th International Conference on Machine Learning*, pages 412–420.