ELSEVIER

journal homepage: www.intl.elsevierhealth.com/journals/ijmi

Classifying disease outbreak reports using n-grams and semantic features

Mike Conway, Son Doan, Ai Kawazoe, Nigel Collier*

National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, Japan

ARTICLE INFO

Article history:
Received 31 October 2008
Received in revised form
10 February 2009
Accepted 25 March 2009

Keywords:
Text classification
Feature selection
Text mining
Information extraction
Disease tracking

ABSTRACT

Introduction: This paper explores the benefits of using n-grams and semantic features for the classification of disease outbreak reports, in the context of the BioCaster disease outbreak report text mining system. A novel feature of this work is the use of a general purpose semantic tagger – the USAS tagger – to generate features.

Background: We outline the application context for this work (the BioCaster epidemiological text mining system), before going on to describe the experimental data used in our classification experiments (the 1000 document BioCaster corpus).

Feature sets: Three broad groups of features are used in this work: Named Entity based features, n-gram features, and features derived from the USAS semantic tagger.

Methodology: Three standard machine learning algorithms – Naïve Bayes, the Support Vector Machine algorithm, and the C4.5 decision tree algorithm – were used for classifying experimental data (that is, the BioCaster corpus). Feature selection was performed using the χ^2 feature selection algorithm. Standard text classification performance metrics – Accuracy, Precision, Recall, Specificity and *F*-score – are reported.

Results: A feature representation composed of unigrams, bigrams, trigrams and features derived from a semantic tagger, in conjunction with the Naïve Bayes algorithm and feature selection yielded the highest classification accuracy (and *F*-score). This result was statistically significant compared to a baseline unigram representation and to previous work on the same task. However, it was *feature selection* rather than *semantic tagging* that contributed most to the improved performance.

Conclusion: This study has 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 previous work in this domain.

1. Introduction

Reliable document classification is an important pre-processing stage in many Information Extraction and text mining systems [7].

This paper compares the performance of a document representation based on highly discriminating unigrams, bigrams, trigrams and semantic features, against a representation derived from unigram and Named Entity (NE) features reported in Doan et al. [6], for the classification of disease

* *Corresponding author*. Tel.: +81 3 4212 2677.
E-mail address: mike@nii.ac.jp (M. Conway).
1386-5056/$ – see front matter © 2009 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.ijmedinf.2009.03.010

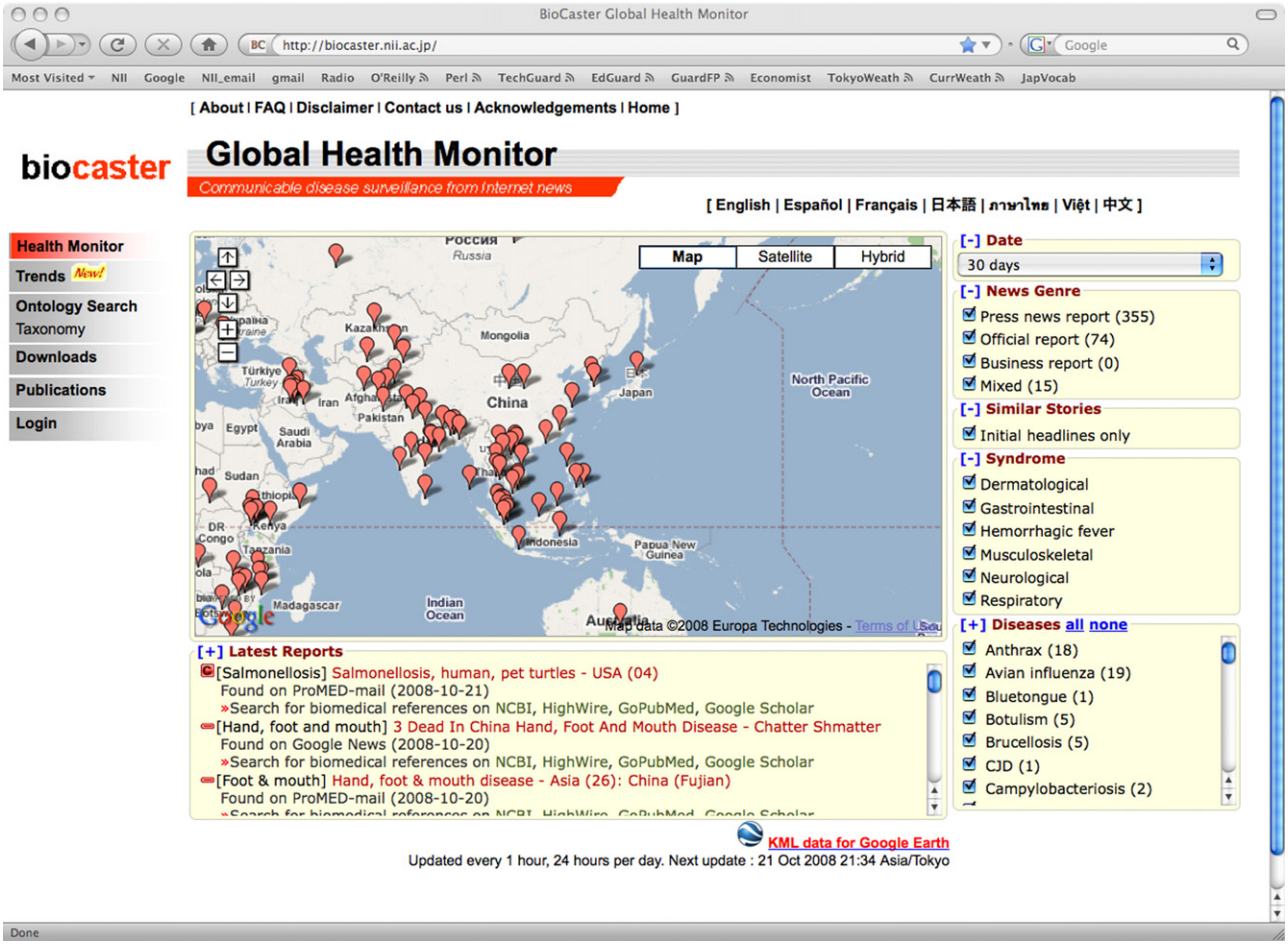

Fig. 1 – BioCaster — *Global Health Monitor* Portal Screenshot

outbreak reports. While the document representation used by Doan et al. [6] performed well for this task, a statistically significant improvement in performance was achieved using a representation built from n-grams and semantic features. A novel feature of the work presented in this paper 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 the current work and how they were derived. Section 4 sets out our methodology, while Section 5 presents results, and some discussion of those results. The final section outlines some broad conclusions on the appropriateness of semantic features and feature selection for the disease outbreak report classification task, and sets out areas for future work.

2. Background

In this section, we will first briefly motivate the current work by describing the application context (the BioCaster epidemiological text mining system), before going on to describe our data (the BioCaster corpus).

2.1. *The* BioCaster *system*

The BioCaster system [5,2] scans online news reports for stories concerning infectious disease outbreaks. An article is of interest if it contains information about newly emerging (or reemerging) infectious diseases of potential international significance, such as, the spread of diseases across international borders, the deliberate release of a pathogen, the discovery of contaminated blood products, and so on. There are two methods that users can exploit to explore extracted data. First, the pre-interpreted information is available from a publicly accessible web portal called *Global Health Monitor* built on Google Maps (Fig. 1 shows the user interface).[1] Second, registered users can opt to receive information (via email) on diseases, countries or other alerting conditions that interest them. According to Heymann and Rodier [10], around 65% of disease outbreaks are first identified from informal sources such as the web, and utilizing text mining technology for epidemiological purposes is becoming increasingly important. There are several systems at various stages of development (includ-

[1] http://www.biocaster.nii.ac.jp.

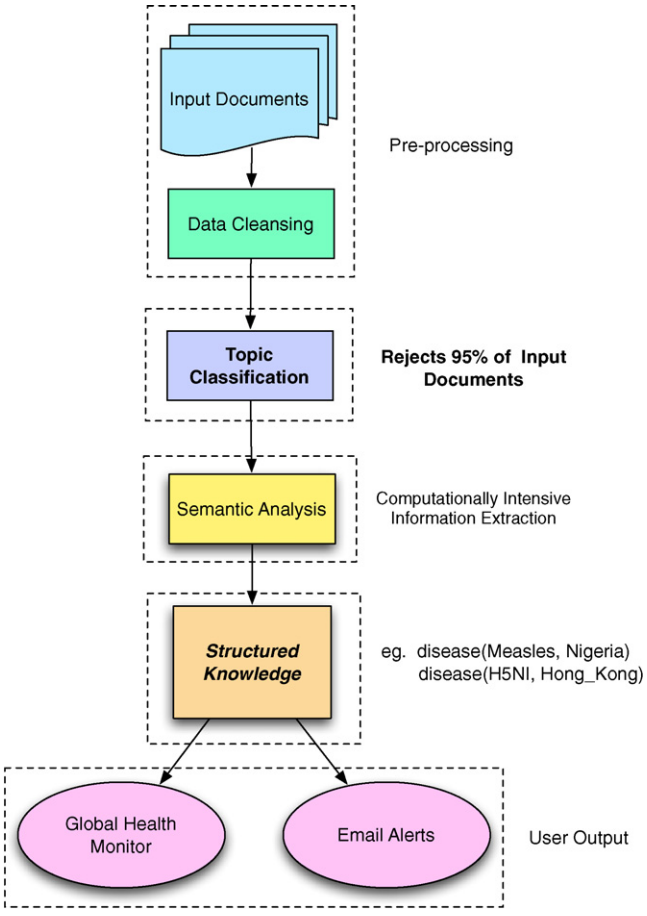

Fig. 2 – BioCaster flowchart.

ing EpiSpider,[2] HealthMap,[3] GPHIN,[4] MedISys,[5] and Project Argus[6]). The BioCaster system differs from existing systems in two significant ways however. First, BioCaster emphasizes Asia-Pacific languages (in addition to English and other languages). Second, one of the primary goals of our research is to reduce the burden on human analysts by trying to automate as much of the information discovery process as possible, although we still regard the human analytic component as a fundamentally necessary part of any effective surveillance system.

The text classification system module (described in this paper) is vital to overall system performance as it filters out irrelevant documents – that is, those documents that are not relevant to disease tracking – before the computationally intensive later stages of deep semantic analysis (see Fig. 2).

2.2. *The* BioCaster *Corpus*

The BioCaster Corpus is a product of the wider BioCaster project. The BioCaster gold standard corpus is a collection

```
<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>
```

Fig. 3 – Example Document from the BioCaster Corpus Showing Named Entity Annotation.

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 Fig. 3 for a truncated example, and Kawazoe et al. [11] for a description of the annotation scheme) using guidelines developed in consultation with the National Institute of Infectious Diseases (Japan) and based on the World Health Organization's "Decision instrument for the assessment and notification of events that may constitute a public health emergency of international concern."[7] 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.[8] Four *percent* 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.[9] 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:

- **Alert** —News stories tagged "alert" are deemed to be of immediate interest to health professionals. Examples of stories in this category, could include a new SARS or Japanese Encephalitis outbreak.
- **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. Examples of borderline stories could include an outbreak of Gastroenteritis in a hospital or cruise ship where Norovirus is the suspected cause.

[2] http://www.epispider.org.
[3] http://www.healthmap.org.
[4] http://www.phac-aspc.gc.ca/gphin/index-eng.php.
[5] http://medusa.jrc.it/medisys/homeedition/all/home.html.
[6] http://biodefense.georgetown.edu/projects/argus.aspx.

[7] http://www.who.int/gb/ghs/pdf/IHR_IGWG2_ID4-en.pdf.
[8] Major sources included the BBC (UK), CBC (Canada), *The Nation* (Thailand), IRIN (United Nations), and the *Sydney Morning Herald*, among others.
[9] http://www.promedmail.org.

Table 1 – Domains in the BioCaster Corpus.

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

- **Publish** —News stories tagged "publish" are judged to be of archival importance to health professionals. Examples of stories in this category might include an update on an ongoing outbreak of Dengue Fever in India, or a small scale Botulism outbreak in the US.
- **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.

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 *outbreak* 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 Fig. 4). The two-class corpus consists of 350 *relevant* documents and 650 *non-relevant* documents.

Doan et al. [6], 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 Named Entity based semantic features as a possible solution.

The approach outlined in this paper extends the work reported in Doan et al. [6] for binary classification of the BioCaster corpus. We take Doan et al.'s work further by employing n-grams, a semantic tagger and feature selection to achieve enhanced classification accuracy.[10]

[10] A preliminary version of this paper appeared as [3].

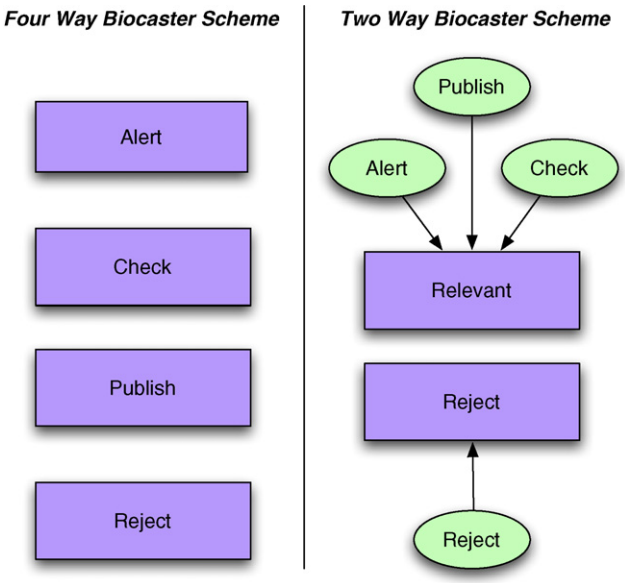

Fig. 4 – Binary categories in BioCaster Corpus.

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 [12] 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 [15], who found "overwhelming evidence" that syntactic features fail to improve topic based classification. Scott and Matwin [20] 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." Previous work using our data has shown however that domain sensitive *semantic* representations can be useful [6,4].

3.1. *Named Entity based features*

Doan et al. [4], in previous work on this task, used the 18 Named Entity tag types (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 NEs and all NE attributes (BOW+NE+roles). Fig. 5 shows how features were generated from a sentence snippet of the BioCaster corpus.

Doan et al. [6] extended this work using a larger data-set — the same data-set used in the current work — and aug-

Table 2 – Named Entities and roles in the BioCaster Named Entity Annotation Scheme.

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

mented the features with a bespoke "semantic dictionary." This approach depended on the creation of domain specific WordNet style synonym sets of verbs and nouns designed to capture the distinctive semantic characteristics of disease outbreak reports. For example:

spread_verbs(spread, circulate, progress, carry)
report_verbs(report, confirm)
examine_verbs(examine, check, screen)
detect_verbs(detect, find, discover, confirm, diagnose)

victim(death, fatality, case, victim, patient)
medical_occupation(doctor, nurse, physician, surgeon)
medical_facility(hospital, clinic, ward, center, center)
spokesman(official, doctor, authority, officer, chief, spokesman)

If one of the words in brackets is matched in the text, then its associated semantic category (spread_verbs, report_verbs, etc.) is added as a feature. We combined these semantic features with the BOW+NE+Roles feature set to create the BOW+NE+Roles+VN feature set. [6] achieved their best result using this representation (93.4% accuracy and 0.91 F-score).

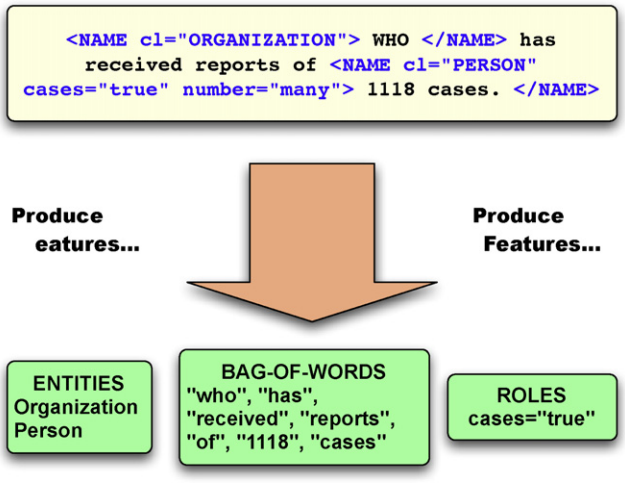

Fig. 5 – Generating BOW+NE+roles features (based on [4]).

3.2. *N-gram features*

N-grams were used (where $n > 1$) as they may help reduce the problems presented by polysemous words (for example "H5N1 **virus**" vs. "computer **virus**") and identify concepts highly characteristic 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. To give a concrete example:

- ministry could plausibly refer to *religious* ministry or some other non-health related arm of government ("In Oklahoma's first execution in 24 years, a man who started a religious **ministry** in prison was put to death by lethal injection early today for . . ."[11]).
- The isolated function word of has no semantic content.
- health can be used in a non medical context "(. . . the **health** of the San Diego economy had been based on the health of the housing market . . .)"[12].

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 [19,18].[13] 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.[14] Fig. 6 shows the 21 top level categories.

According to [19] 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.

- 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

[11] Leading article in *The New York Times* 10th September 1990.
[12] Article in *The Guardian* 4th October 2008.
[13] 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/. A web based interface to the system – Wmatrix – is available at http://ucrel.lancs.ac.uk/wmatrix.
[14] The tagset used in the USAS semantic tagger was loosely based on that developed by [13].

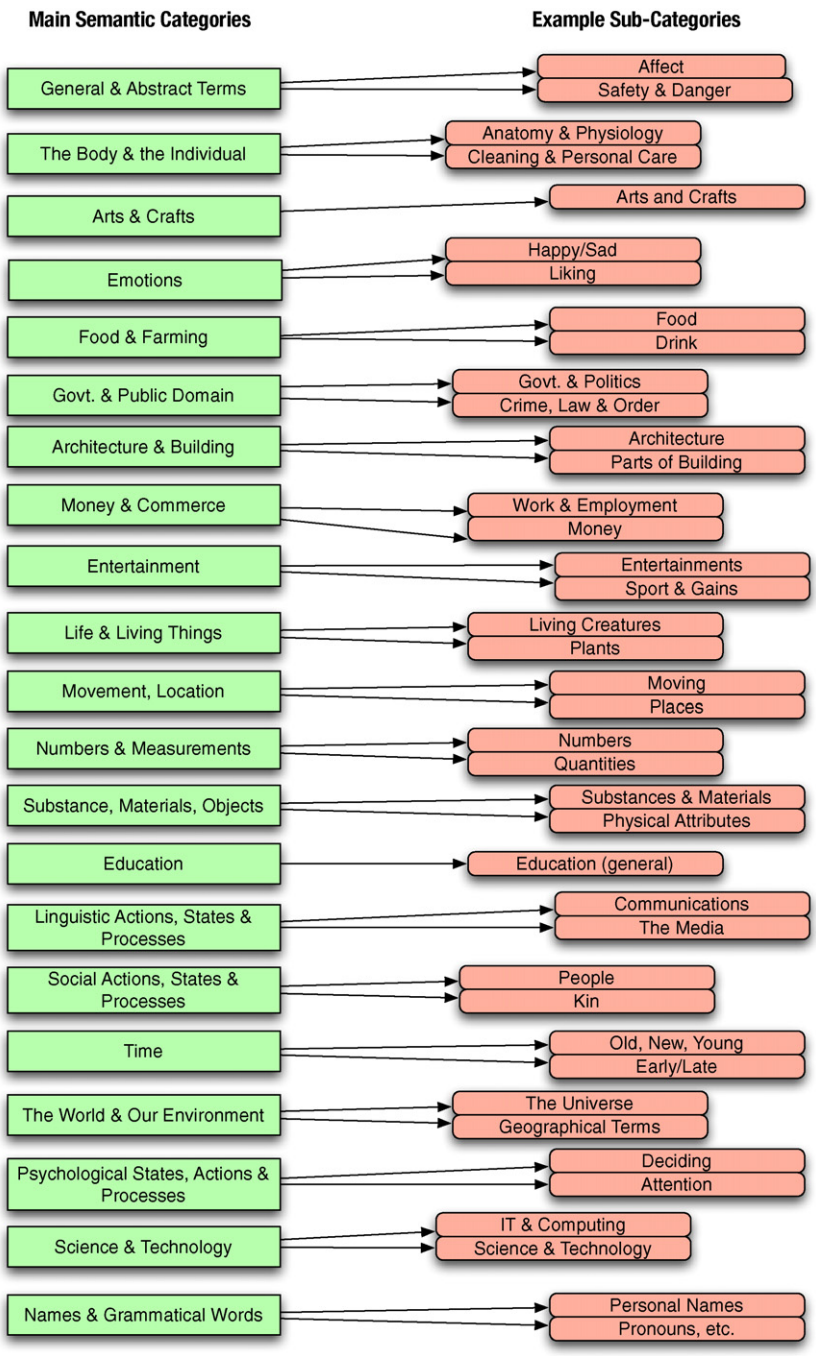

Fig. 6 – UCREL Semantic Tag Scheme.

semantic tags. For example, in the food domain, "battered" is more likely to refer to the cooking technique ("As a kid, seafood meant **battered** cod, boil-in-the-bag haddock or crab paste ..."[15]), rather than suggest conflict or violence ("A **battered** Gordon Brown Faces More Blows."[16])

- 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's account 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

[15] Cookery article in *The Guardian* 12th June 2007.
[16] Headline in *International Herald Tribune* 13th May 2008.

the purposes of this work, multiword units were not used in the representation due to some difficulties reliably extracting them from the USAS output format. 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 [21], and has been used in the study of the compositionality of multiword expressions [17]. An important difference between the USASsemantic tagger and other more well known lexical semantic resources, like WordNet[8] 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, and is not specifically designed for the biomedical domain.[17] 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.

4. Methodology

In our experiments we used two feature representations; term frequency and binary. Term frequency was used in order to facilitate a meaningful comparison between Doan et al. [6] and the current approach. A binary representation was used as early experimental work indicated that binary features performed better than weighted features for these typically short documents. This position is supported by [24], who found that for non-topical text classification — in [24]'s case the classification of literary text — binary feature representations produce higher accuracy. Stopword removal was not used in any experiment (and was not used by [6]). Three machine learning algorithms were employed: Naïve Bayes, Support Vector machines and the C4.5 decision tree algorithm [22,14]. The Wekadata mining toolkit[18] 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 [1]. Accuracy — the main metric used in this work — is the percentage of correctly defined documents (defined as the number of correctly assigned instances divided by the total number of instances). However, we also report other common text classification metrics (Recall, Precision, Specificity and *F*-score). A contingency table is used to perform calculations (see Table 3). Accuracy, Precision, Recall, Specificity and *F*-Score are defined (respectively) in Eqs. (1)–(5).

Table 3 – Contingency table for calculating classification accuracy (REL is "Relevant" and non-REL is "Non-Relevant").

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

$$\text{Accuracy} = \frac{a+d}{a+b+c+d} \tag{1}$$

$$\text{Precision} = \frac{a}{a+b} \tag{2}$$

$$\text{Recall} = \frac{a}{a+c} \tag{3}$$

$$\text{Specificity} = \frac{d}{b+d} \tag{4}$$

$$\text{F-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

Specificity and recall are especially important to note here, as they are not dependent on the proportion of *relevant* and *reject* documents in the corpus, and hence are possibly more indicative of real world performance. That is, in the BioCaster corpus, 35% of documents belong to the relevant category, whereas in the working BioCaster system, less than 5% of the input documents are relevant.

Feature selection techniques are central to this work. Yang and Pedersen [23] 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 [9] provides a survey of feature selection methods for text classification.

The χ^2 method — implemented in Weka — was used for feature selection as it has shown to be effective in the context of text classification [23]. For more on the χ^2 method see Oakes et al. [16].

5. Results and discussion

In order to compare our results with [6], we performed two sets of experiments. The first set of experiments used the same pre-processing steps as [6](that is, no stopword removal and a term frequency document representation). The second set of experiments used a binary document representation and no stopword removal. This second set of experiments is the focus of our discussion, as binary features yielded better results than term frequency representations (although not for every feature/classifier combination). Our chosen baseline is the BOW+NE+Roles+VN feature set identified by [6].

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,

[17] 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.

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

Table 4 – Initial results (Note that "BOW" is "Bag-of-Words.") Results for binary and term frequency document representations are presented.

Features	No. features	Accuracy (binary/term freq.)		
		NB	SVM	C4.5
SEMTAG	580	78.8/67.4	82.8/85.3	76.9/79.9
SEMTAG (COMP)	263	78.4/68.7	82.9/85.3	74.1/80.1
UNIGRAMS	21322	88.4/85.5	90.9/90.0	80.8/82.1
BIGRAMS	1567	87.6/82.7	87.1/85.6	83.5/81.2
TRIGRAMS	2345	82.5/80.9	81.1/78.9	82.2/77.0
BOW+NE+ROLES	21334	88.3/83.9	90.4/89.3	84.1/80.5
BOW+NE+ROLES+VN	21408	88.3/85.7	89.9/89.9	84.0/82.8
χ^2 (CHI-SQUARED)	9000	**94.8**/89.9	92.2/93.9	81.6/88.1

worked best of all. Table 4 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 dimension itself as a feature. For example, if we take the USAS tag E2 (Liking/Disliking dimension), 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.

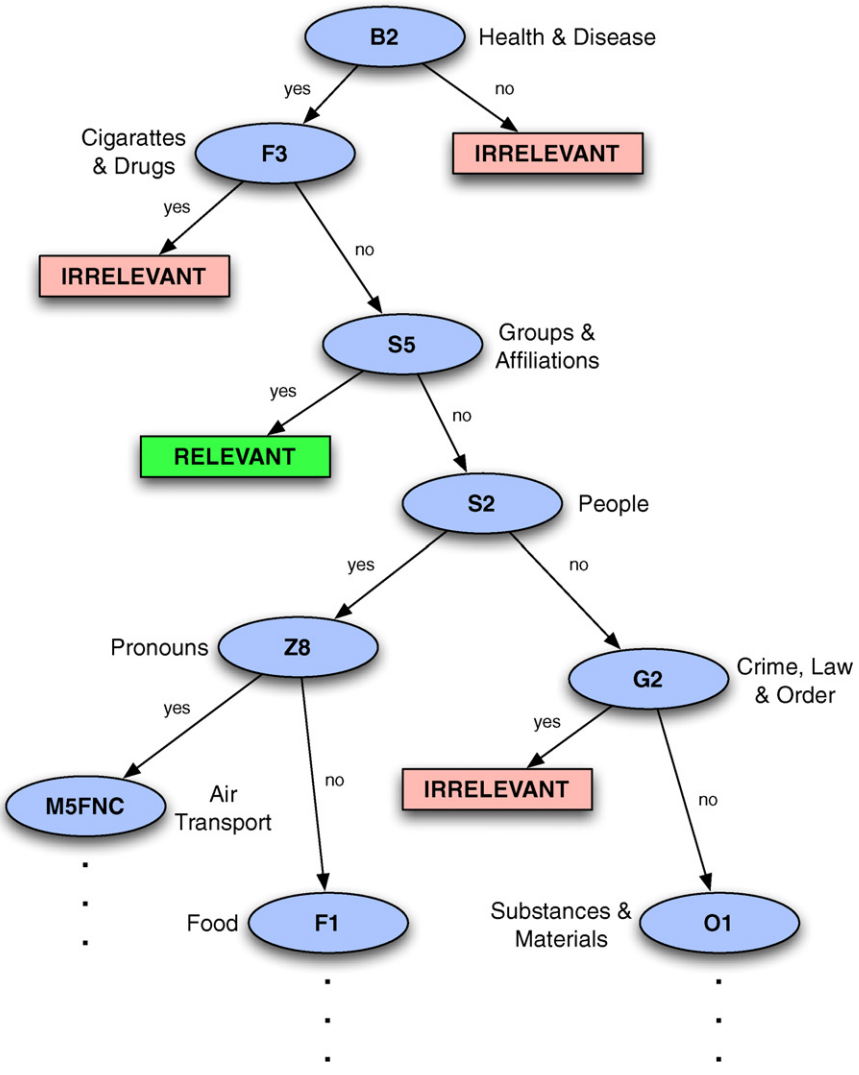

Fig. 7 – Partial C4.5 decision tree for semantically tagged BioCaster corpus.

The C4.5 decision tree algorithm seems to perform consistently worse than both the Naïve Bayes and SVM[19] algorithms. One of the advantages of the decision tree algorithm however, is its potential for data exploration purposes. Fig. 7 shows the root of a partial decision tree derived from the (full) USAS semantic tag representation of the BioCaster corpus (using binary features). 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," 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 indicator that a document is concerned with disease outbreaks) in the *groups and affiliations* tag.[20]

The best performing feature set (94.8% accuracy using the Naïve Bayes algorithm – see Table 4) 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+VN feature set. This was true for both binary and term frequency based document representations, although the χ^2 9000 result for the term frequency representation was a little lower at 93.9% (using the SVM algorithm). (Note that the "term frequency" result is directly comparable to [6].) Rather than choosing an arbitrary cut off point for feature selection, the optimal number of features was derived experimentally, using stratified 10-fold cross validation in conjunction with the χ^2 feature selection method. For each feature, the mean χ^2 value is calculated based on the 10 stratified cross validations. It is this mean that is then used to rank features. We used cross validation in order to help eliminate positive bias, while at the same time using all our limited data. Fig. 8 shows that accuracy peaks at around 9000 features for Naïve Bayes, and gradually decreases when more features are added. The performance of the other two classifiers used is also shown. Note that for C4.5, classification accuracy peaks with a small number of features, then declines as features are added. It can be seen that the SVM algorithm performs very consistently as more features are added.

The 9000 most powerfully discriminating 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 5, the two most discriminating *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. The ten most discriminating features are unigrams. All these unigrams are not however *specific* to the disease outbreak domain. For instance, *outbreak, confirmed, reported, death* and so on can be used in the context of war, disasters and social emergencies generally. This suggests that the classifier may perform less well when processing newspaper reports concerning war, social collapse or civil unrest. The role of government is also clear in Table 5; *ministry* occurs as a unigram, in a bigram and in two trigrams.

For data exploration purposes, we also performed feature selection using only semantic tags. The 20 most discriminating semantic features (including example concordances)

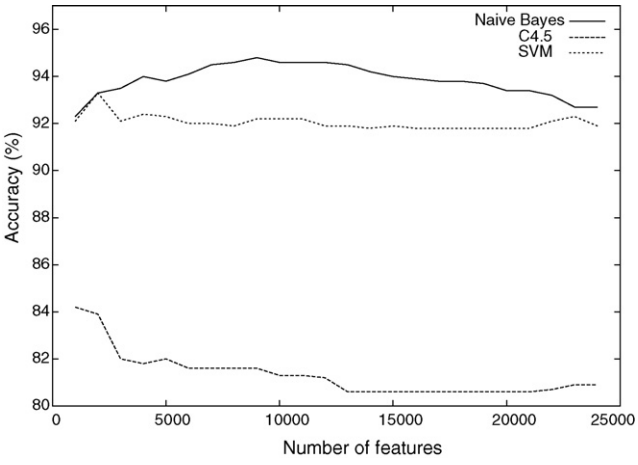

Fig. 8 – Comparison of feature selection thresholds.

Table 5 – Most discriminating features in the BioCaster Corpus.

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

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

[20] 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.

Table 6 – χ^2 Semantic features with examples—note that features 5 and 11 are incorrectly tagged.

RANK	SEMANTIC TAG	EXAMPLE
1	Health & Diseases [B2]	... have been *infected* and chickens killed ...
2	Life & Living Things [L1]	... in Sukabumi infected by *wild* polio virus ...
3	Medicine [B3]	... polio *immunization* for 4000 children ...
4	Living (General) [L2]	... polio *virus* spread, the government ...
5	Money (General) [I1]	... the death *toll* in the diarrhea outbreak ...
6	Wanting; Planning, etc. [X7]	... DOES NOT OCCUR IN RELEVANT DATA ...
7	Objects (General) [O2]	... of food *items* like unpacked bread ...
8	Crime, Law [G2]	... eating *improperly* cooked fish ...
9	Living (Gender) [L2mfn]	... a *monkey* whose remains were found ...
10	Deciding [X6]	... school has *decided* to shut its doors ...
11	Sports [K5]	... extend the *exercise* by two days ...
12	Entertainment [K1]	... they attended a wedding *party* ...
13	Places [M7]	... confirmed in Ibbi *local* government *area* ...
14	Business [I2]	... banned the *sale* of food items ...
15	People [S2mfc]	... on the part of the *people* of Kenema ...
16	Warfare [G3]	... are a medical *officer* of health ...
17	Movement [M3]	... on the *roadside* to ensure ...
18	Arts [C1]	... an *artist*, died on 8 July ...
19	Pronoun [Z8]	... government area and *it* has spread ...
20	Moving [M1]	... the spread of *waterborne* diseases ...

Table 7 – Binary document representation. Note that "F" is "F-score" and "Spe" is "Specificity".

Features	Naïve Bayes				SVM				C4.5			
	F	Pre	Rec	Spe	F	Pre	Rec	Spe	F	Pre	Rec	Spe
SEMTAG	0.72	0.68	0.78	0.79	0.76	0.75	0.77	0.86	0.68	0.67	0.68	0.82
SEMTAG(COMP)	0.71	0.67	0.77	0.79	0.76	0.76	0.77	0.86	0.65	0.63	0.66	0.79
UNIGRAMS	0.85	0.76	0.98	0.83	0.87	0.86	0.89	0.92	0.74	0.72	0.75	0.84
BIGRAMS	0.82	0.86	0.78	0.93	0.81	0.84	0.79	0.91	0.76	0.78	0.74	0.87
TRIGRAMS	0.72	0.85	0.62	0.94	0.72	0.75	0.70	0.87	0.73	0.79	0.68	0.90
BOW+NE+RO.	0.85	0.76	0.97	0.83	0.87	0.85	0.88	0.92	0.78	0.76	0.80	0.86
BOW+NE+RO+VN	0.80	0.76	0.98	0.83	0.86	0.83	0.89	0.93	0.75	0.74	0.75	0.87
χ^2 9000	**0.93**	0.89	0.97	0.94	0.89	0.88	0.90	0.93	0.75	0.73	0.76	0.84

are shown in Table 6, where we can see an example of a typical tagging mistake. The world "toll" is tagged as belonging to the class **Money (general)**, whereas the context of the sentence "the death *toll*" clearly refers to human mortality.

Of the 9000 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 9000 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.

In order to gain a better understanding of the results presented, we calculated f-score, precision, recall and speci-

Table 8 – Term frequency representation. Note that "F" is "F-score" and "Spe" is "Specificity".

Features	Naïve Bayes				SVM				C4.5			
	F	Pre	Rec	Spe	F	Pre	Rec	Spe	F	Pre	Rec	Spe
SEMTAG	0.66	0.52	0.90	0.55	0.80	0.80	0.79	0.88	0.71	0.73	0.69	0.86
SEMTAG(COMP)	0.66	0.53	0.90	0.54	0.79	0.79	0.79	0.88	0.73	0.71	0.75	0.83
UNIGRAMS	0.81	0.76	0.86	0.85	0.86	0.83	0.89	0.90	0.74	0.75	0.74	0.86
BIGRAMS	0.78	0.70	0.86	0.81	0.79	0.81	0.77	0.90	0.72	0.75	0.70	0.84
TRIGRAMS	0.72	0.74	0.70	0.86	0.67	0.73	0.62	0.88	0.62	0.74	0.53	0.90
BOW+NE+RO	0.79	0.72	0.87	0.82	0.85	0.82	0.88	0.90	0.72	0.72	0.73	0.85
BOW+NE+RO+VN	0.80	0.76	0.86	0.85	0.86	0.83	0.89	0.90	0.75	0.74	0.75	0.86
χ^2 9000	0.87	0.81	0.93	0.88	**0.91**	0.91	0.91	0.95	0.83	0.82	0.84	0.90

ficity for each classifier (see Table 7 for binary results. In the interests of completeness, term frequency results are presented in Table 8). It it is clear that the Naïve Bayes/ χ^2 classifier delivers *very* high recall (0.97), although this statement must be qualified with the observation that the baseline feature set (in conjunction with Naïve Bayes) provides slightly higher recall (0.98). It is also notable that specificity is very high for the χ^2 9000 feature set in conjunction with the Naïve Bayes algorithm (equal first with the trigram feature set in conjunction with Naïve Bayes algorithm at 0.94). The difference in *precision* between the two classifiers is much more stark, with a thirteen point difference between the baseline BOW+NE+Roles+VN feature set and the best performing feature set (0.76 and 0.89), though the BOW+NE+Roles+VN feature set in conjunction with the SVM algorithm performs a little better (0.83). In the context of a system that identifies disease outbreaks from newspaper texts, the cost of failing to identify a relevant text is very high, therefore our priority is to maximize recall, but maintain precision at acceptable levels. The Naïve Bayes/ χ^2 classifier meets this goal, as it provides very high recall (0.97), while providing the best precision of all the classifiers we have studied.

One further point to bear in mind is that the BioCaster corpus is *manually* tagged for Named Entities. In the context of a working system, where Named Entity recognition is performed on input documents *automatically* (and with mistakes) it is likely that performance will reduce. The approach suggested in this paper does not rely on human intervention – in our evaluation we use text automatically tagged using the USAS tagger – and is thus more likely to reflect "real world" performance.

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+VN" 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.'s [6] BOW+NE+roles to the feature selection operation).
- Semantic features derived from the USAS tagger will be considered to enhance other modules of the BioCaster text mining system.

Summary points

What was known before the study?

- High quality document classification is essential for an epidemiological text mining system.
- Unigram based features have proven stubbornly effective for general document classification.

What this study has added to the body of knowledge?

- A combination of n-gram and semantic features (generated by the USAS tagger), combined with feature selection improves classification accuracy at a statistically significant level compared to previous work.
- The use of a general purpose semantic tagger — the USAS tagger — is useful for exploring our corpus of disease outbreak reports.

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

[1] R. Bouckaert, E. Frank, Advances in Knowledge Discovery and Data Mining, in: chapter Evaluating the Replicability of Significance Tests for Comparing Learning Algorithms, Springer, Berlin, 2004, pp. 3–12.
[2] N. Collier, S. Doan, A. Kawazoe, R. Matsuda-Goodwin, M. Conway, Y. Tateno, Q.-H. Ngo, D. Dien, A. Kawtrakul, K. Takeuchi, M. Shigematsu, K. Taniguichi, BioCaster: Detecting Public Health rumors with a Web-based Text Mining System, Bioinformatics 24 (24) (2008) 2940–2941.
[3] M. Conway, S. Doan, A. Kawazoe, N. Collier, Classifying Disease Outbreak Reports Using N-grams and Semantic Features, in: Proceedings of the Third International Symposium on Semantic Mining in Biomedicine (SMBM 2008), Turku, Finland, 2008, pp. 29–36.
[4] S. Doan, A. Kawazoe, N. Collier, The Role of Roles in Classifying Annotated Biomedical Text, in: Proceedings of BioNLP 2007: A Workshop of ACL 2007, 2007, pp. 17–24.
[5] S. Doan, Q. Hung-Ngo, A. Kawazoe, N. Collier, Global Health Monitor - A Web Based System for Detecting and Mapping Infectious Diseases, in: Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP), 2008, pp. 951–956.
[6] S. Doan, A. Kawazoe, M. Conway, N. Collier, Towards Role-based Filtering of Disease Outbreak Reports. Journal of Biomedical Informatics, 2009. doi:10.1016/j.jbi.2008.12.009.
[7] R. Feldman, J. Sanger, The Text Mining Handbook: Advanced Approaches to Analyzing Unstructured Data, CUP (2007).

[8] C. Fellbaum (Ed.), WordNet: An Electronic Lexical Database, MIT Press, Cambridge, Mass, 1998.
[9] G. Forman, An Extensive Empirical Study of Feature Selection Metrics for Text Classification, Journal of Machine Learning Research 3 (2003) 1289–1305.
[10] D. Heymann, G. Rodier, WHO Hot spots in a wired world: WHO surveillance of emerging and re-emerging infectious diseases, The Lancet 1 (5) (2001) 345–353.
[11] A. Kawazoe, J. lihua, M. Shigematsu, R. Barrero, K. Taniguchi, N. Collier, The Development of a Schema for the Annotation of Terms in the BioCaster Disease Detecting/Tracking System, in: Proceedings of the Second International Workshop on Formal Biomedical Knowledge Representation, 2006, pp. 77–85.
[12] D. Lewis, Lewis Representation and Learning in Information Retrieval, Ph.D. Thesis, Department of Computer Science, University of Massachusetts, Amherst, US, 1992.
[13] T. McArthur (Ed.), Longman Lexicon of Contemporary English, Longman, London, 1981.
[14] T. Mitchell, Machine Learning, McGraw-Hill International, Singapore, 1997.
[15] A. Moschitti, R. Basili, Complex Linguistic Features for Text Classification: A Comprehensive Study, in: In Proceedings of the 26th European Conference on Information Retrieval Research, 2004, pp. 181–196.
[16] M. Oakes, R. Gaizauskas, H. Fowkes, A. Jonsson, V. Wan, M. Beaulieu, A Method Base on the Chi-Square Test for Document Classification, in: Proceedings of the 24th annual international ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 21), 2001, pp. 440–441.
[17] S. Piao, P. Rayson, O. Mudraya, A. Wilson, R. Garside, Measuring MWE Compositionality USING Semantic Analysis, in: Proceedings of COLING/ACL 2006 Workshop on Multiword Expresions: Identifying and Exploiting Underlying Properties, 2006, pp. 2–12.
[18] P. Rayson, From Key Words to Key Semantic Domains, International Journal of Corpus Linguistics 13 (4) (2008) 519–549.
[19] P. Rayson, D. Archer, S. Piao, T. McEnery, The UCREL Semantic Analysis System, in: Proceedings of the Workshop on Beyond Named Entity Recognition: Semantic Labelling for NLP Tasks in association with the Fourth International Conference on Language Resources and Evaluation (LREC 2004), 2004, pp. 7–12.
[20] S. Scott, S. Matwin, Feature Engineering for Text Classification, in: In Proceedings of the 16th International Conference on Machine Learning, 1999, pp. 379–388.
[21] S. Sharoff, B. Babych, P. Rayson, P. Mudraya, S. Piao, ASSIST: Automatic Semantic Assistance for Translators, in: Proceedings of the 11th Conference of the European Chapter of the Association for Computational LInguistics (EACL 2006), 2006, pp. 132–139.
[22] I.H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, 2nd edition, Morgan-Kaufmann, San Francisco, 2005.
[23] Y. Yang, J. Pedersen, A Comparative Study on Feature Selection in Text Categorization, in: Proceedings of ICML-97, 14th International Conference on Machine Learning, 1997, pp. 412–420.
[24] B. Yu, An Evaluation of Text Classification Methods for Literary Studies, Literary and Linguistic Computing 23 (3) (2008) 327–343.