Using Hedges to Enhance a Disease Outbreak Report Text Mining System

Mike Conway, Nigel Collier
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku
Tokyo 101-8430, Japan
{mike|collier}@nii.ac.jp

Son Doan
Vanderbilt University Medical Center
2525 West End Ave., Suite 800
Nashville, TN 37235, USA
son.doan@vanderbilt.edu

1 Introduction

Identifying serious infectious disease outbreaks in their early stages is an important task, both for national governments and international organizations like the World Health Organization. Text mining and information extraction systems can provide an important, low cost and timely early warning system in these circumstances by identifying the first signs of an outbreak automatically from online textual news. One interesting characteristic of disease outbreak reports — which to the best of our knowledge has not been studied before — is their use of speculative language (*hedging*) to describe uncertain situations. This paper describes two uses of hedging to enhance the BioCaster disease outbreak report text mining system.

Following a brief description of the BioCaster system and corpus (section 2), we discuss in section 3 previous uses of hedging in NLP and the methods used to identify hedges in the current work. In section 4 we describe some initial classification experiments using hedge features. Section 5 describes a "speculative" method of tagging disease outbreak reports with a metric designed to aid users of the BioCaster system in identifying articles of interest.

2 BioCaster System & Corpus

The BioCaster system scans online news reports for stories concerning infectious disease outbreaks (e.g. H5N1, Ebola) and makes its results available to registered users as email alerts (Collier et al., 2008). In addition to this email service, data that has been filtered through a topic classifier but which is still uninterpreted is used to populate a Google Map application called the *Global Health Monitor*.[1]

The BioCaster corpus consists of 1000 news articles downloaded from the WWW and then manually categorized and annotated with Named Entities by two PhD students. Articles were collected from various news sources (e.g. *BBC*, *New York Times* and ProMED-Mail[2]). Each document is classified as either *relevant* (350) or *reject* (650).[3]

The corpus is designed to include difficult borderline cases where more advanced understanding of the context is required. For example, an article may be about, say, polio, but not centrally concerned with specific outbreaks of that disease. Instead, the article could report a vaccination campaign or research breakthrough.

3 Hedges

According to Hyland (1998), in an extensive study of speculative language in science writing, hedges "are the means by which writers can present a proposition as an opinion rather than a fact." More recently, Kilicoglu and Bergler (2008) have presented a method for automatically identifying hedges in the biomedical domain. In the current work, we used a science orientated hedge lexicon derived from Mercer et al. (2004). The lexicon consisted of 72 verbs (including *appear, appears, appeared, appearing, indicate, indicates, indicated, indicating*, and so on) and 32 non-verbs (including, *about, quite, poten-*

[1] www.biocaster.org

[2] ProMED-Mail is a human curated service for monitoring disease outbreak reports (www.promedmail.org.)

[3] For copyright reasons, the BioCaster corpus is not publicly available.

Proceedings of the Workshop on BioNLP, pages 142–143,
Boulder, Colorado, June 2009. ©2009 Association for Computational Linguistics

Rank	Hedge	Rank	Hedge
1	reported	9	suggests
2	suspected	10	estimated
3	probable	11	appeared
4	suspect	12	appearing
5	usually	13	mostly
6	see	14	assumes
7	reports	15	predicted
8	sought	16	suggested

Table 1: Statistically Significant Hedges

Features	Naive Bayes		SVM	
	Acc	F	Acc	F
9000 χ^2	94.8	0.93	92.2	0.89
Unigram	88.4	0.85	90.9	0.87
Unigram+hedge	88.0	0.85	91.7	0.89

Table 2: Classification Results

tially, likely and so on). Preliminary work showed that the frequency of hedge words differs in the two categories of the BioCaster corpus (*relevant* and *reject*) at a highly significant level using the χ^2 test ($P < 0.01$). Table 1 shows the 16 most discriminating hedge words in the BioCaster corpus (identified using the χ^2 feature selection method.)

4 Classification Experiment

The current BioCaster system uses n-gram based text classification to identify disease outbreak reports, and reject other online news. We used hedging features to augment this classifier, and evaluated the results using a subset of the BioCaster corpus. One binary hedging feature was used. The feature was "true" if and only if one of the 105 hedge lexemes identified by Mercer et al. (2004) occurred in the input document within 5 words of a disease named entity. Results are shown in Table 2, where it can be seen that the addition of a single binary hedge feature to the unigram feature set increases accuracy by 0.8%. The performance does not however reach the level achieved by the χ^2 9000 n-gram feature set described in Conway et al. (2008).

5 Towards a "Speculative" Metric

	Accept (%)	Reject (%)
High	64.2	48.3
Medium	29.5	36.7
Low	6.3	15.0

Table 3: Proportion of Articles in Each Category

Users of the BioCaster system would benefit from an indicator of how "speculative" each news article is, as breaking news regarding disease outbreaks is characterized by uncertainty, which is encoded using hedging. We use the Mercer list of 105 hedging words as described above, in conjunction with statistics derived from a 10,000 document section of the Reuters corpus to provide a "speculative" metric.[4] We calculated total frequencies for all 105 hedge words in each of the 10,000 Reuters documents — that is, the *total* number of hedge words per document — then ranked these frequencies (after normalizing the frequencies to take account of document length). The bottom third of documents had hedge percentages in the range 0% - 0.2544% (Low). The middle third had hedge percentages in the range 0.2545% - 1.0574 (Medium). The range for the top third was 1.0575% - 100% (High). Documents inputted to the BioCaster system automatically have their proportion of hedge words calculated and are assigned a value according to their position on the scale (Low, Medium or High). Table 3 shows that a majority of the documents in the *accept* segment of the BioCaster corpus can be tagged as highly speculative using this method.

References

N. Collier, S. Doan, A. Kawazoe, R. Matsuda-Goodwin, M. Conway, Y. Tateno, Q-H. Ngo, D. Dien, A. Kawtrakul, K. Takeuchi, M. Shigematsu, and K. Taniguichi. 2008. BioCaster: Detecting Public Health Rumors with a Web-based Text Mining System. *Bioinformatics*, 24(24):2940–2941.

M. Conway, S. Doan, A. Kawazoe, and N. Collier. 2008. Classifying Disease Outbreak Reports Using N-grams and Semantic Features. *Proceedings of the Third International Symposium on Semantic Mining in Biomedicine (SMBM 2008), Turku, Finland*, pages 29–36.

K. Hyland. 1998. *Hedging in Scientific Research Articles*. John Benjamins, Amsterdam.

H. Kilicoglu and S. Bergler. 2008. Recognizing Speculative Language in Biomedical Research Articles: a Linguistically Motivated Perspective. *BMC Bioinformatics*, 9(Suppl 11):S10.

R. Mercer, C. DiMarco, and F. Kroon. 2004. The Frequency of Hedging Cues in Citation Contexts in Scientific Writing. In *Proceedings of the Canadian Conference on AI*, pages 75–88.

[4] Reuters Corpus, Volume 1, English language, 1996-08-20 to 1997-08-19 (Release date 2000-11-03, Format version 1, correction level 0).