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-Mail2). 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, potential

---

1www.biocaster.org
2ProMED-Mail is a human curated service for monitoring disease outbreak reports (www.promedmail.org.)
3For copyright reasons, the BioCaster corpus is not publicly available.
Table 1: Statistically Significant Hedges

| 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 2: Classification Results

| Features       | Naive Bayes | SVM |
|----------------|-------------|-----|
|                | Acc         | F   | Acc | F |
| 9000 $\chi^2$  | 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 3: Proportion of Articles in Each Category

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

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).