Detecting Web Rumours with a Multilingual Ontology-Supported Text Classification System

Nigel Collier[1], PhD, Ai Kawazoe[1], PhD, Son Doan[1], PhD, Mika Shigematsu[2], MD, Kiyosu Taniguchi[2], MD, Lihua Jin[1], PhD, John McCrae[1], MSc, Hutchatai Chanlekha[1], MSc, Dinh Dien[3], PhD, Quoc Hung[3], MSc, Van Chi Nam[3], MSc, Koichi Takeuchi[4], DEng, Asanee Kawtrakul[5], PhD

[1]National Institute of Informatics, Tokyo, Japan; [2]National Institute of Infectious Diseases, Tokyo, Japan; [3]Vietnam National University (HCM), Vietnam; [4]Okayama University, Okayama, Japan; [5]Kasetsart University, Bangkok, Thailand

OBJECTIVE

In this paper we present a summary of the BioCaster system architecture for Web rumour surveillance, the rationale for the choices made in the system design and an empirical evaluation of topic classification accuracy for a gold-standard of English and Vietnamese news.

BACKGROUND

Timely surveillance of disease outbreak events of public health concern currently requires detailed and time consuming manual analysis by experts. Recently in addition to traditional information sources, the World Wide Web (Web) has offered a new modality in surveillance, but the massive collection of multilingual texts which must be processed in real time presents an enormous challenge.

Among currently active Web surveillance systems is the Public Health Agency of Canada's GPHIN system [1] and the MiTAP system [2]. Several key issues remain including the need for increased automation of relevance detection, extending surveillance to cover languages in the Asia-Pacific region and the need for a quantitative evaluation of system accuracy.

We present a new system called BioCaster, based on a multilingual ontology of terms in six Asia-Pacific languages [3] whose purpose is (a) to provide the computable semantics for 18 named entity (NE) classes, 3 role types and 7 domain relationships in this domain [4], (b) to bridge the gap between laymen's terms that are commonly used in newswire and expert conceptualization, and (c) to mediate translation of equivalent terms in different languages.

METHODS

The overall target of our system is to classify articles according to a simple four class standard: r*eject, publish, check* (borderline) and *alert*. After data is downloaded from the Internet using an RSS aggregator and cleansed we perform NE and role analysis, and then topic classification. At this early stage we aim simply to separate *reject* articles from everything else. Further down the pipeline event analysis will be used to make fine-grained distinctions with a knowledge of modality, negation, temporality etc. We leave this for future work and focus here on the early stage tasks. To test the ability of the system to classify topicality correctly we collected 1000 news texts in English and annotated them by hand for terms in the 18 NEclasses, their roles as well as topical relevance. 350 were judged positive. This was repeated for 334 Vietnamese news texts with 167 judged positive.

RESULTS

We compared Naïve Bayes (NB) against Support Vector Machines (SVM)[6]. On the English corpus we attained an accuracy with NB of 88.1%. For Vietnamese accuracy was 91.3% using SVM. While the size of the gold standard did not allow us to achieve performance closure we believe that the results show promising levels of performance and furthermore highlighted interesting trends in the task such as the contribution made by specific entity types in combination with roles such as *case*.

CONCLUSIONS

The BioCaster system is currently operational on a cluster computer and downloads in excess of 5000 news reports each day from over 1000 feeds. From these approximately 40.6 are found to be relevant each day and made available for online search by registered users.

REFERENCES

[1] Public Health Agency of Canada. 2004. Global Public Heath Intelligence Network (GPHIN). http://www.gphin.org

[2] Damianos, L., Ponte, J., Wohlever, S., *et al.* 2002, 'MiTAP, Text and Audio Processing for Bio-security: A Case Study'. In: *Proc. IAAI-2002, Alberta, Canada.*

[3] Collier, N., 2007, Kawazoe, A., Jin, L. *et al.* "A multilingual ontology for infectious disease surveillance: rationale, design and challenges", J. Language Resources and Evaluation (in press).

[4] Kawazoe,A., Jin, L., Shigematsu, M., *et al.*. 2006. "The development of a schema for the annotation of terms in the BioCaster disease detection/tracking system" *Proc KR-MED*, pp. 77–85.

[6] V. N. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, New York, 1995.