Assessing the Coverage of BioCaster Terms in Web News

Nigel Collier, PhD, Reiko Matsuda Goodwin, PhD, Ai Kawazoe, PhD, Son Doan, PhD

National Institute of Informatics, Tokyo, Japan

OBJECTIVE

We describe here a multilingual ontology to support disease surveillance by intelligent text mining systems from Web-based rumours. We informally assessed the coverage of its English terms on large sample of news collected from the Web.

BACKGROUND

The BioCaster ontology (BCO) [1-2] contains a structured hierarchy of high level categories related to infectious diseases and has been in use at the National Institute of Informatics in an intelligent text mining system called BioCaster since 2006. BioCaster aims to surveillance the Web for disease outbreaks and other public health hazards mentioned in news and other online texts.

Text mining (TM) – the extraction of structured content from unstructured free text - is increasingly being used in conjunction with knowledge-based resources to provide systems with a limited level of expert-like background understanding. An ontology for text mining needs to fulfill two major criteria: (1) to provide a subset of normalized terms that appear within the collection of texts being analysed, and (2) to sufficiently describe the relations between terms so that the TM system can infer what the typical user needs to know from the limited information that it sees at the surface level in the text. Examples include knowing that a particular pathogen caused a disease, knowing that a province is part of a specific country or knowing that a pathogen is a subtype of another. (2) is also of key importance in helping to limit the number of rules that need to be written by hand.

METHODS

Constructing the term set in the BCO referenced a variety of extant ontologies but critical to our aim was term harvesting using standard text mining techniques on large collections of newswire and official reports. All terms were hand curated by a domain expert, a linguist and a computational linguist before being entered into the ontology. Multilingual terms were separated from the main hierarchy in synonym sets much like EuroWordNet [3] allowing the text mining system to establish cross-language equivalence.

RESULTS

The second release of the BCO saw the total number of root terms for infectious diseases rise from 34 to 95 with corresponding increases in the number of pathogens from 30 to 102, symptoms from 116 to 144 and animals from 44 to 81. All root terms are cross-referenced against external resources. Term coverage which includes significant numbers of near synonyms was greatly expanded. The rise in terms for each language was as follows: Chinese from 503 to 729, English from 494 to 1091, Japanese from 297 to 812, Korean from 510, Thai from 380 to 746 and Vietnamese from 672 to 973. Spanish and French were newly added.

The present study aims to shed light on term representativeness in real texts and to highlight areas where future extensions must focus. We analysed approximately 29,000 news articles and measured English disease term coverage from the ontology. It was found for example that 78% of disease terms directly matched those found in the corpus. Other term classes had a similarly high level of representativeness.

CONCLUSIONS

The BCO is a multilingual application ontology focused on the needs of natural language processing applications in public health. Since the release of its second version in April 2008 the ontology, open source and freely available in OWL (Web Ontology Language) format, has been downloaded and used by a growing number of medical and academic groups worldwide. We freely welcome feedback and contributions. Future work will focus on expanding the number of public health threats and to extend the languages.

REFERENCES

[1] Collier, N., Kawazoe, A., Jin, L., Shigematsu, M., Dien, D. Barrero, R., Takeuchi , K.and Kawtrakul, A. (2007), "A multilingual ontology for infectious disease surveillance: rationale, design and challenges", Language Resources and Evaluation, Elsevier.

[2] Kawazoe, A., Chanlekha, H., Shigematsu, M. and Collier, N. (2008), "Structuring an event ontology for disease outbreak detection", in BMC Bioinformatics, 9 (Suppl 3): S8.

[3] Vossen, P. (ed.), (1998), "EurWordNet: a multilingual database with lexical semantic networks", Dordrecht, Netherlands: Kluwer.

Further Information:
Nigel Collier, collier@nii.ac.jp
www.biocaster.nii.ac.jp