Options for the Control of Influenza VI

June 17-23, 2007 • Toronto, Ontario, Canada

Editor
Jacqueline M Katz, PhD

INTERNATIONAL MEDICAL PRESS

Towards Ontology-Driven Influenza Surveillance From Web Rumours

Nigel Collier[1], *Ai Kawazoe*[1], *Mika Shigematsu*[2], *Kiyosu Taniguchi*[2], *Lihua Jin*[1], *John McCrae*[1], *Dinh Dien*[3], *Quoc Hung*[3], *Koichi Takeuchi*[4], *Asanee Kawtrakul*[5]

[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

Surveillance of infectious disease rumours on the Web remains a potentially valuable source of information to public health workers. We argue that computer systems that perform this task require high-quality knowledge sources including a taxonomy of structured concepts, variant terms, including laymen's language, with equivalence relations across languages. The objective of the BioCaster ontology is to help fulfill this role for major languages in the Asia-Pacific region. In this paper we present a summary of the first version of the ontology and briefly describe features which make it useful for influenza surveillance.

Introduction

The recent H5N1 avian influenza epidemic has highlighted the need for improved surveillance to minimize the effect of future pandemics. With timeliness, coverage and the high cost of traditional information infrastructure a concern in the Asia-Pacific region, the ability to exploit unsubstantiated Web news is emerging as a new modality for early detection of disease outbreaks. Several systems have already been deployed including GPHIN [1] and MiTaP [2]. However substantial challenges remain including the massive size of the Web, its multilingual nature and the uncontrolled proliferation of terms. We argue that without knowledge intensive methods search times for Web news will overwhelm scarce expert resources. To support the development of intelligent systems we present a multilingual ontology focused on the needs of infectious disease surveillance with vocabulary in six Asia-Pacific languages (English, Chinese, Japanese, Korean, Thai and Vietnamese) and briefly discuss its application to influenza surveillance. The BioCaster Ontology (BCO) serves the needs of the infectious disease surveillance community by bridging the gap between the uncontrolled use of terminology, including laymen's terms, in online news and the need for a computable semantics in surveillance systems. The ontology therefore seeks to serve the dual purpose of enabling advanced search on newswires by experts using their own vocabulary and also automated understanding and alerting of events reported in online news. Our domain of interest is basically a subset of biomedicine that is focused on mediating the integration of textual content in various languages. Textual content in biomedicine, especially in news reports, exhibits considerable variability which needs to be systematized. A plethora of major nomenclatures and classification systems already exist that we can draw on including SNOMED CT [3], the Unified Medical Language System [4] and ICD10 [5] as well as lexical ontologies such as EuroWordNet [6] each with varying degrees of rigor, coverage and accessibility. Most of these are mono-lingual domain ontologies with a scope far broader and deeper than the application ontology we have in mind for BCO. Few such resources though exist for Asia-Pacific languages, exemplifying the need for high quality cross-language resources to support biomedical applications.

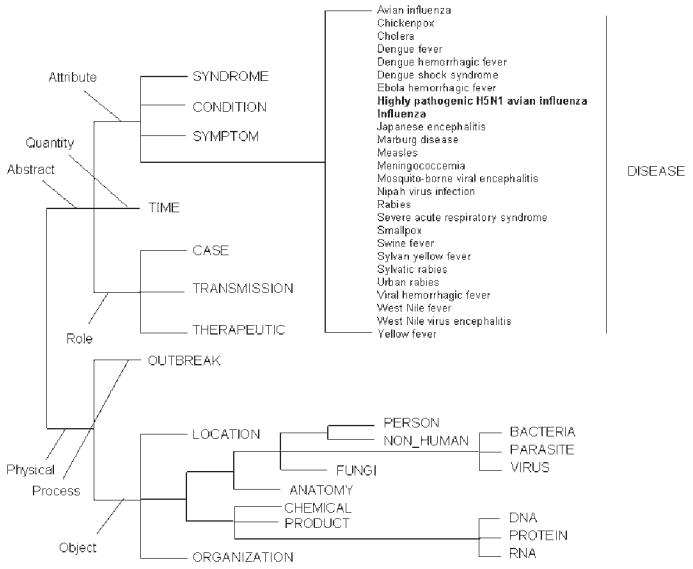
Materials and Methods

Discussions with epidemiologists and analysis of WHO consultation reports (e.g. [7]) revealed several scenarios for information surveillance including: the moment of transition from animal-to-human transmission and sustained human-to-human transmission, the spread of a virulent pathogen across international borders and the deliberate release of a virulent pathogen into the population. For the top level of our ontology various extant ontologies were considered and SUMO (the IEEE Suggested Upper Merged Ontology) [7] was chosen. This essentially covers non-lexicalized domain independent classes such as attribute, role or process and should enable interoperability across other ontologies. We then surveyed a collection of 1000 news articles and identified eighteen target entity classes of terms including virus, bacteria, disease, syndrome, symptom and host. The classes were chosen to reflect the granularity of entities that could be found in news articles. In practice the survey of the inventory for entity types is often not trivial because ad-hoc concept classes incorporate a mixture of substance and role viewpoints for example the class of people and cases. We were helped in our analysis by adapting the formal method suggested by Guarino and Welty [8] and detailed in [9]. Key relations were also identified between target classes such as host-pathogen, disease-symptom and pathogen-transmission. A list of high priority pathogens was then formed from a survey of notifiable diseases in the region by a geneticist supported by an epidemiologist and a computational linguist. Pathogens included the H5N1 subtype of influenza A virus. Following this groundwork, workflow to fill in the domain terms then generally followed the procedure set out in the EuroWordNet project [6]. An ontology fragment was first identified that was based on the first 27 pathogens in our high priority list. We then harvested terms from news articles using named entity recognition [10] before validating and defining them. Among 28140 English documents processed between April 2006 and early July 2006, we harvested 1695 unique entities for diseases, 399 for symptoms, 37825 for locations, and 262 for anatomy. In order to bootstrap the ontology development we looked for term pairs in known associative relations using mutual information [2] as the measure. Once the English term set was established for the fragment we proceeded to describe term equivalence across the six languages. This work was undertaken by linguists fluent in each of the languages.

Results and Discussion

In the first release version of the BCO we have described terminology in six languages for 27 pathogens and the diseases they cause, 108 symptoms, 10 routes of transmission, and 6 syndromes of those diseases. In total the synonyms for each language consist of 479 English terms, 274 Japanese terms, 296 Korean terms, 283 Chinese terms, 361 Thaiterms and 265 Vietnamese terms. Links were made to external ontologies and nomenclature such as ICD10, LOINC, MedDRA, MeSH and SNOMED CD. The resulting ontology has given us a compact structuring focused on one domain and application. As can be seen from the structural framework given in Fig. 1 the class of diseases includes influenza and its daughter Highly pathogenic H5N1 avian influenza.Within the ontology the two are related by a narrower term relation and each has a set of synonymous terms in the six languages. A single term was also selected as the preferred term for each concept. This is useful for systems that need to unify output so that users do not need to be concerned with variations. In our work so far we have presented a taxonomy of objects which aims to meet the need of a computable semantics for disease outbreak surveillance from news. In future work we expect to keep expanding the ontology and term banks year by year. We are now focusing on the design of a hierarchy of events related to disease outbreaks. The major challenges here though include considerations of event granularity, event inclusion (e.g. a single case and a group case) and temporality. (BCO is freely available to browse and download at http://biocaster.nii.ac.jp).

Figure 1. Structural overview of the BioCaster Ontology. Entity classes are shown capitalized.



Acknowledgements

We gratefully acknowledge funding from the Research Organi-zation of Information System's Fusion Fund.

References

1. The Global Public Health Intelligence Network (GPHIN): http://www.phacaspc.gc.ca/media/nr-rp/2004/2004_gphin-rmispbk_e.html Accessed July 10, 2007
2. Damianos L, Ponte J, Wohlever S, et al. 2002, 'MiTAP, Text and Audio Processing for Bio-security: A Case Study'. In: *Proc. Fourteenth Innovative Applications of Artificial Intelligence (IAAI-200), Alberta, Canada*.
3. Stearns MQ, Price C, Spackman KA, Wang AY. 2001, Clinical Terms: overview of the development process and project status. In: *Proc. American Medical Informatics Association (AMIA) Symposium*. pp. 662-666.
4. Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. *Methods of Information in Medicine*. 1993;32:281-291.
5. WHO International Classification of Diseases (ICD-10): http://www.who.int/classifications/icd/en Accessed July 10, 2007
6. Vossen P. Introduction to EuroWordNet. *Computers and the Humanities*. 1998;32:73-89.
8. World Health Organization WHO consultation on priority public health interventions before and during an influenza pandemic. Technical report.
9. Niles I, Pease A. Origins of the Standard Upper Merged Ontology'. In: *Working Notes of the IJCAI-2001 Workshop on the IEEE Standard Upper Ontology, Seattle, Washington*.
10. Guarino N, Welty, C. A formal ontology of properties. In: R. Dieng and O. Corby (eds.): *EKAW-2000: Proc. 12th Int. Conf. on Knowledge Engineering and Knowledge Management*. pp. 97-112.
11. Kawazoe A, Jin L, Shigematsu M, et al. The development of a schema for the annotation of terms in the BioCaster disease detection/tracking system. In: *KR-MED 2006: Proc. Int. Workshop on Biomedical Ontology in Action, Baltimore, USA*.
12. Takeuchi K, Collier N. Bio-medical entity extraction using support vector machines. *Artificial Intelligence in Medicine*. 2005;33(2):125-137. DOI information: 10.1016/j.artmed.2004.07.019.
13. Church K, Hanks P. Word association norms, mutual information, and lexicography. *Computational Linguistics*. 1990;16(1):22-29.