Lang Resources & Evaluation (2006) 40:405–413
DOI 10.1007/s10579-007-9019-7

A multilingual ontology for infectious disease surveillance: rationale, design and challenges

Nigel Collier · Ai Kawazoe · Lihua Jin · Mika Shigematsu · Dinh Dien · Roberto A. Barrero · Koichi Takeuchi · Asanee Kawtrakul

Published online: 26 June 2007

Abstract A lack of surveillance system infrastructure in the Asia-Pacific region is seen as hindering the global control of rapidly spreading infectious diseases such as the recent avian H5N1 epidemic. As part of improving surveillance in the region, the BioCaster project aims to develop a system based on text mining for automatically monitoring Internet news and other online sources in several regional languages. At the heart of the system is an application ontology which serves the

N. Collier (✉) · A. Kawazoe · L. Jin
National Institute of Informatics, Tokyo, Japan
e-mail: collier@nii.ac.jp

A. Kawazoe
e-mail: zoeai@nii.ac.jp

L. Jin
e-mail: lihua-jin@mp.nii.ac.jp

M. Shigematsu
National Institute of Infectious Diseases, Tokyo, Japan
e-mail: mikas@nih.go.jp

D. Dien
Vietnam National University (HCM), Ho Chi Minh City, Vietnam
e-mail: ddien@fit.hcmuns.edu.vn

R. A. Barrero
Murdoch University, Perth, Australia
e-mail: rbarrero@ccg.murdoch.edu.au

K. Takeuchi
Okayama University, Okayama, Japan
e-mail: koichi@cl.it.okayama-u.ac.jp

A. Kawtrakul
Kasetsart University, Bangkok, Thailand
e-mail: ak@ku.ac.th

dual purpose of enabling advanced searches on the mined facts and of allowing the system to make intelligent inferences for assessing the priority of events. However, it became clear early on in the project that existing classification schemes did not have the necessary language coverage or semantic specificity for our needs. In this article we present an overview of our needs and explore in detail the rationale and methods for developing a new conceptual structure and multilingual terminological resource that focusses on priority pathogens and the diseases they cause. The ontology is made freely available as an online database and downloadable OWL file.

Keywords Infectious disease surveillance · Multilingual ontology · Text mining

1 Introduction

Recent epidemics among both humans (SARS) and animals (avian influenza) have shown clear gaps in the disease surveillance systems of Asia-Pacific region countries. Although surveillance should be the cornerstone of the defense against such rapidly spreading diseases, a lack of timely information has been seen to hinder the control efforts of public agencies. In the BioCaster project we are developing a text mining system for outbreak surveillance from Internet news and academic literature which can aid public health experts in recognizing clusters of potentially rapidly spreading infectious disease outbreaks. The overall benefit should be to raise awareness of threats and to reduce uncertainty in order to make informed interventions.

Among a handful of currently active surveillance systems for monitoring early developments of internationally spreading diseases is the Public Health Agency of Canada's GPHIN system (Public Health Agency of Canada 2004). This system represents the state-of-the-art and is credited by the World Health Organization (WHO) (Grein et al. 2000) with the earliest detection of the SARS (severe acute respiratory syndrome) epidemic. However it does have some limitations such as not currently having specialized terminological coverage in some regional languages such as Japanese, Korean, Thai, or Vietnamese. Additionally the knowledge sources behind the system are not publicly available, limiting the ability of users to review or expand them.

The need for local language processing capability becomes clear when we consider that timeliness is one of the key factors in the value of information for minimizing morbidity and mortality. An outbreak or incidence is likely to be first mentioned publicly in the local media but further time will pass before the story is translated and published in the international media, if it is published at all.

At the heart of BioCaster is a multilingual application ontology which serves as the computable semantics for the text mining system. The ontology should serve the purpose both of enabling advanced searches on the mined facts and of allowing the system to make intelligent inferences for assessing the priority of events so that alerts can be automatically sent. However, it became clear early on in the project

though that existing classification schemes did not have the necessary language coverage or semantic specificity for our needs.

Consider for example the following scenario which illustrates the semantic-driven search capability of such an ontology: A public health expert is interested in finding out about a possible incidence of viral reassortment occurring in a H5N1 avian influenza case in Vietnam. The expert logs in to the BioCaster portal and enters *H5N1 avian influenza* as the search term along with *Vietnam*, the date range of interest and requests only English language news articles. Internally BioCaster recognizes that the first term is an English variant of a root-term in its disease concept hierarchy *highly pathogenic H5N1 avian influenza* and that there are a number of English synonyms such as *H5N1 disease*, *HPAI (H5N1)* and *A(H5N1) flu* which it can use to expand the query. The search is performed but the results are not relevant to the user's information need. The system then offers the user the choice of searching using related symptoms which are based on the <hasSymptom> relation as well as the pathogenic agent found by the <causedBy> relation. The user selects this option and the search is performed again using symptoms such as *cough*, *pneumonia* and *acute respiratory distress* and the agent name *influenza A virus subtype H5N1*. This time an article is found but the report is already 2 weeks out of date and missing some vital pieces of information about the name of the location. The user then chooses to search the Vietnamese news and the search is repeated using Vietnamese term equivalents. After the system retrieves the Vietnamese news, a structured translation is generated for each event summarizing mined information in English by following the <isSynonymOf> relation to the root term and from there the <preferredTerm> is found for English. Each term is given in its <preferred-Term> form, making events easier to compare. The expert then finds the event that she is searching for where the location name is clearly identified. In this scenario the system has helped the user to quickly find relevant information by expanding the query with semantically related terms and also to cross the language barrier.

In this introductory paper we present a brief discussion of the rationale, design and challenges for our multilingual ontology. In the following section we provide a brief survey of some major related resources and comment on their influence on the BioCaster Ontology (BCO); Sect. 3 outlines the general methodology, the design process, details about scope and the priority for populating BCO with terms; in Sect. 4 we conclude by discussing the ongoing work.

2 Related work

Our domain of interest is basically a subset of biomedicine that is focussed 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 GALEN Core (Rector et al. 1995), SNOMED CT (Stearns et al. 2001), and the Unified Medical Language System (Lindberg et al. 1993), each with varying degrees of rigor, coverage and accessibility. Most of these are monolingual 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. Below we survey a few mostly multilingual resources, examining each for overlap with our objective.

EuroWordNet (Vossen 1998) is a widely used multilingual lexical ontology for general language processing. The basic unit of class is the *synset* which aims to group words and expressions with the same meaning in a given context. Synsets are related through hyponym, hyperonym, meronym and various other relations. As the structuring of WordNets is essentially language dependent, EuroWordNet provides a bridge between language specific WordNets by adopting *Inter-Lingual-Index* (ILIs) for relating synsets in different languages. However EuroWordNet was not intended to be domain specific and as such it lacks depth of terminological coverage and more importantly domain-specific relations. Our purpose on the other hand is to make explicit the relations between the disease, the pathogenic agent, the typical location of incidence, symptoms and the mode of transmission. We take inspiration from EuroWordNet in several areas such as the use of a top level structure that is broadly similar and the use of a mediating node which we call a *root term*. Where we differ is by setting the role of the root term to be both the container linking synonyms across languages and also in being the object to which various domain sensitive relations point. By using the root term essentially as an interlingual pivot, we simplify the construction and maintenance of domain-specific relations.

Wikipedia is a large-scale multilingual source of encyclopedic knowledge created by collaborative effort on the Web with over 1.3 million articles in English, over 100,000 in Japanese, and over 10,000 in Chinese, Korean, Thai and Vietnamese. Articles in different languages can be linked with interlanguage links. It has many valuable resources for our purpose such as lists of infectious diseases and their relations to symptoms, transmission agents and pathogens. Two potential disadvantages though are that its links encode a variety of association relations and the entries themselves may vary in quality, timeliness and coverage due to an open editorial system (Giles 2005). Categorization of articles may also not necessarily strictly reflect ontological principals. Naturally also the entries are written for human readers and require structuring. None of these though are serious barriers to knowledge reuse and in practice we have found Wikipedia to be a valuable resource.

All the diseases we are interested in fall within the International Statistical Classification of Diseases and Related Health Problems, ICD10, (WHO 2004a). This is a detailed and widely used coding system for diseases published by the World Health Organization with various national extensions. Diseases are structured within a classification and given with diagnosis and a unique code. One point of concern is that we need to take care to consider the level of granularity that is practical for terms that will appear in the news sources. For example ICD10 makes fine-grained distinctions between four classes of tuberculosis which in turn have 38 subclasses. Issues are also raised by how ICD-10 partitions the domain where terms are often composite entities incorporating a disease base plus a condition and can also include underspecification (Bodenreider et al. 2004).

Finally, PHSkb (Doyle et al. 2005) from the Centers for Disease Control and prevention is a coding system to support the exchange of electronic data about observations of notifiable diseases between public health professionals in the United States (US). It provides extensive coverage of notifiable diseases and their causal agents in the US. There are some points of divergence though with our approach: (a) the coding system supports only English and is focussed on terminology applicable to the US situation, (b) the relations sometimes lack rigor, e.g. there are separate subtrees for organisms whose role is *Vector* and *transmission_mode_values* which includes *vector borne* but nothing to relate the two, as well as underspecified relations such as *associated substances*, (c) the terminology coverage does not directly include synonyms although this may be recoverable from the links to controlled vocabularies.

3 Method

In the BCO we have initially started with six languages: Chinese (simplified), English, Korean, Japanese, Thai and Vietnamese. After completion of the top level structure by a computational linguist consisting of essentially non-lexicalized domain independent classes, leaf classes are constructed that correspond to domain-dependent entity classes (target entity classes) which have been (Guarino and Welty 2000) and detailed in (Kawazoe et al. 2006).

Terms were then gathered for English, Korean, Japanese and Chinese by a biologist with support from an epidemiologist, a geneticist and the computational linguist. Sources include those surveyed in Sect. 2 in addition to terminology harvesting which is done on an automatically annotated corpus using named entity recognition (NER). Quality control starts at the design stage with best practice coming from the Open Biomedical Ontologies initiative (OBO) (Smith et al. 2005) guidelines. Support for Vietnamese and Thai terminology is from linguists fluent in those languages.

The target for the first release is to construct 200 *root terms* (synonymous term clusters) with their definitions and relations. While this is modest in comparison to established classifications, it should allow us to have up to 1000 verified surface level terms spread across the six languages, giving us a compact structuring focussed on one domain and application. Following from this we expect to keep expanding the ontology and term banks year by year.

For the domain dependent classes and relations we follow a broadly similar work flow to the EuroWordNet project. We first identify an ontology fragment based on a list of priority pathogens which are gathered from lists of notifiable diseases on ministry of health Web sites. This is designed to concentrate expert resources where they will be of most value. The priority pathogen list leads naturally to a specification of vocabulary scope and the collection of terminology such as the diseases they cause in hosts, the symptoms they exhibit, etc. Terms are then encoded and their equivalence and associative relations identified. Following from this we perform quality checking and release the new version for public feedback and

evaluation. The second stage comes where we compare, mediate and restructure ontology fragments.

For tool support BCO is being developed using the Protege ontology editor (Noy et al. 2001) with the Web Ontology Language (OWL) plug-in allowing for export to a description logic formalism and integration with a reasoner for validation. Versioning is controlled by Subversion.

3.1 Scope

The backbone of the ontology is the familiar subsumption hierarchy (hypernym, hyperonym relations). The scope of the vocabulary and relations were determined through joint discussions between computational linguists and domain experts. Several scenarios were revealed for disease surveillance and those which received high priority include: (a) the moment of transition from animal-to-human transmission to limited or sustained human-to-human transmission of a pathogen; (b) the spread of an infective and virulent pathogen across international borders; (c) the deliberate release of a virulent pathogen into the human population. Supported by WHO consultation reports (WHO 2004b) our discussions revealed the need to focus on detection and tracking of unusual clusters rather than individual cases.

Genetic epidemiology adds another dimension to the information needs as viral DNA/RNA and their interaction with the host's genes play a key role in determining susceptibility or resistance to pathogens. We therefore plan on adding in a further level of detail about the pathogenic agent and host which includes genes and their products. The strategy behind including such information in the BCO is to obtain a total picture of each pathogen in terms of its life cycle with enhanced reference capability for human experts and potential to understand articles in the lifescience literature database given in MEDLINE.

3.2 Design

The BCO at the top level consists of a foundation ontology taken from the OWL formatted version of the Suggested Upper Merged Ontology, SUMO, (Niles and Pease 2001)—for a discussion on foundation ontologies see e.g. (Farrar 2003). The SUMO ontology provides very general classes such as an *Entity* with subclasses *Attribute*, *Quantity*, *Object*, and *Process* giving a potential source of integration with other ontologies. SUMO also includes a much more extensive taxonomy and a rich axiomization in both SUO-KIF and OWL full. For our purpose SUMO's hierarchy was simplified in order to remove superfluous details by requiring all non-leaf classes to have at least two children. One potential difficulty we noted with SUMO is that it was not so clear how to relate some of our event classes (e.g. outbreak) under *Process* (perdurant) as the subclassification here is not exhaustive and does not cover some event types. Since most of the taxonomy we have developed so far pertains to objects (endurants) a discussion of this will be left for future work.

The mid-level of the BCO consists of a disjoint set of target entity classes. These were chosen by considering the level of granularity that could be achieved using

automatic term recognition and grounding techniques, e.g. (Takeuchi and Collier 2005). This has resulted in a relatively shallow ontology which nevertheless has expressive relations. The current set of target entity classes is shown in Fig. 1.

When developing a multilingual ontology the question arises about how to establish cross language relationships between terms and whether it is possible to adopt a strict notion of synonymy when we are considering non-cognate language pairs. In general language this notion of synonymy would be fraught with difficulty but we expect that for specialized domains such as the infectious disease outbreak domain term correspondence might be simpler to maintain. For example the multilingual correspondence between translations of *weakness* and *fatigue* are shown in Fig. 2. Our experience so far for Disease, Virus, Bacteria and Symptom classes has shown this to be a manageable issue within the framework we have established although we expect that other classes such as Location or Anatomy may be more difficult to unify.

As outlined earlier we have adopted the notion of a *root term* which serves as an interlingual pivot between terms within and across languages connecting to language specific terms with the <synonymTerm> relation and to preferred terms in each of these languages with the <preferredTerm> relation. Each root term takes extra properties including a unique identifier, a definition, an editor note, a scope note and various links to external vocabularies and resources such as ICD10, MeSH, SNOMED CT and Wikipedia. Language specific terms take properties that include a unique identifier, an ISO 639 language identifier, and whether it is an abbreviation or a colloquialism.

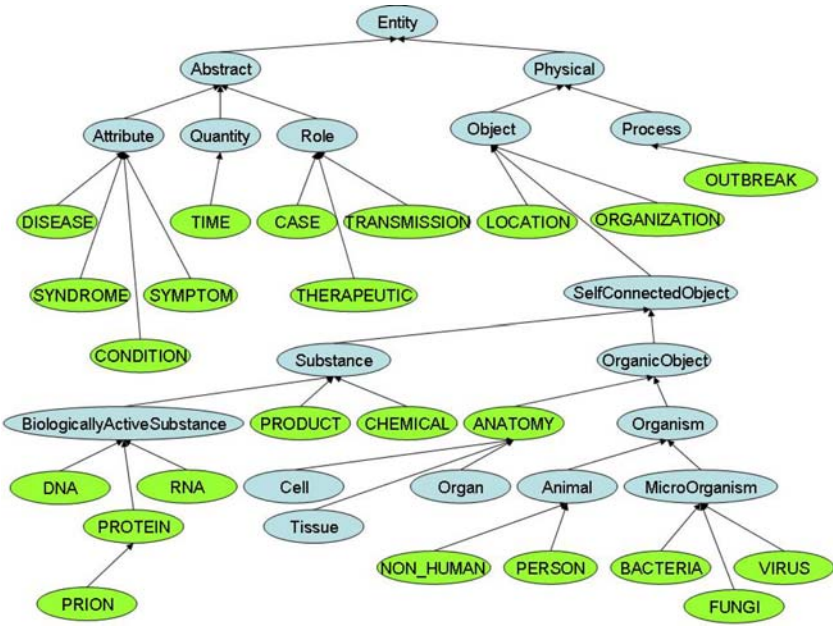

Fig. 1 Partial BCO class hierarchy showing the top level and target entity classes (capitalized)

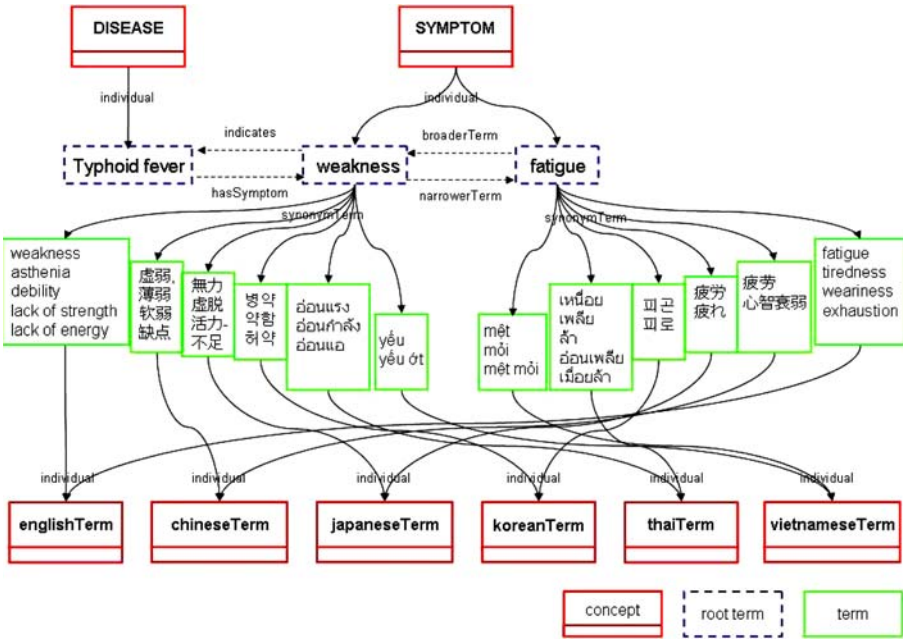

Fig. 2 Example of multilingual term correspondence for *weakness* and *fatigue*

The set of relations we have currently identified as being the core of our application includes associations between pathogen and host (<hasPrimaryHost>), disease and symptom (<hasSymptom>), disease and country (<notifiableIn>), pathogen and organ (<infectsOrgan>), pathogen and mode of transmission to humans (<hasTransmissionMode>) and disease and symptom (<hasSymptom>).

4 Conclusion

The BCO outlined in this article arose from the need for a multilingual ontology to underpin the development of an infectious disease surveillance system. It is an application ontology which is being transparently and collaboratively developed to support infectious disease surveillance. At the same time we expect that it can also be used in the future to bootstrap the development of monolingual biomedical text mining systems for Asia-Pacific languages where specialized nomenclatures are much in need. The first version of the BCO was released in January 2007 at http://www.biocaster.nii.ac.jp. We actively solicit feedback for improvement and extension of the ontology.

In future publications we will discuss further about several key issues such as quality control, event semantics and automatic term harvesting from corpora.

Acknowledgements This study was supported by a grant from the Research Organization of Information Systems (ROIS). We also gratefully acknowledge useful discussions with Abla Mawudeku and Michael Blench at GPHIN about their system.

References

Bodenreider, O., Smith, B., & Burgun, A. (2004). The ontology-epistemology divide: a case study in medical terminology. In *Proceedings of the Third International Conference on Formal Ontology in Information Systems (FOIS 2004)* (pp. 185–195). IOS Press.

Doyle, T. J., Ma, H., Groseclose, S. L.,& Hopkins, R. S. (2005). PHSkb: A knowledgebase to support notifiable disease surveillance. *BMC Medical Informatics and Decision Making, 5*(27). PMID:16105177.

Farrar, S. (2003). An ontology for linguistics on the semantic web. Ph.D. thesis, Department of Linguistics, The University of Arizona.

Giles, J. (2005). Internet encylopaedias go head to head. *Nature, 438*, 900–901. doi:10.1038/438900a.

Grein, T. W., Kamara, K. B., Rodier, G., Plant, A. J., Bovier, P., Ryan, M. J., Ohyama, T., & Heymann, D. L. (2000). Rumours of disease in the global village: Outbreak verification. *Emerging Infectious Diseases, 6*, 97–102.

Guarino, N., & Welty, C. (2000). Ontological analysis of taxonomic relations. In A. Laender, S. Liddle, V. E. Storey (Eds.), *Proceedings of ER-2000: The International Conference on Conceptual Modeling* (pp. 210–224). Berlin, Germany: Springer Verlag LNCS.

Kawazoe, A., Jin, L., Shigematsu, M., Barerro, R., Taniguchi, K.,& Collier, N. (2006). 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* (pp. 77–85). Baltimore, USA.

Lindberg, D. A., Humphreys, B. L., & McCray A. T. (1993). The Unified Medical Language System. *Methods of Information in Medicine, 32*, 281–291.

Niles, I., & Pease, A. (2001). Origins of the standard upper merged ontology. In *Working Notes of the IJCAI-2001 Workshop on the IEEE Standard Upper Ontology*. Seattle, Washington.

Noy, N. F., Sintek, M., Decker, S., Crubezy, M., Fergerson, R. W., & Musen, M. A. (2001). Creating semantic web contents with protégé-2000. *IEEE Intelligent Systems, 16*(2), 60–71.

Public Health Agency of Canada. (2004). Global Public Health Intelligence Network (GPHIN). http://www.phac-aspc.gc.ca/media/nr-rp/2004/2004_gphinrmispbke.html.

Rector, A., Solomon, W., Nowlan, T., & Rush, A. (1995). A terminology server for medical language and medical information systems. *Methods of Information in Medicine, 34*, 147–157.

Smith, B., Ceusters, W., Klagges, B., Kohler, J., Kumar, A., Lomax, J., Mungall, C. J., Neuhaus, F., Rector, A., & Rosse, C. (2005). Relations in biomedical ontologies. *Genome Biology, 6*, R46.

Stearns, M. Q., Price, C., Spackman, K. A., & Wang, A. Y. (2001). Clinical terms: Overview of the development process and project status. In *Proc. American Medican Informatics Association (AMIA) Symposium* (pp. 662–666).

Takeuchi, K., & Collier, N. (2005). Bio-medical entity extraction using support vector machines. *Artificial Intelligence in Medicine, 33*(2), 125–137. DOI information: 10.1016/j.artmed.2004.07.019.

Vossen, P. (1998). Introduction to EuroWordNet. *Computers and the Humanities, 32*, 73–89.

WHO. (2004a). *ICD-10, International Statistical Classification of Diseases and Related Health Problems, Tenth Revision*. World Health Organization.

WHO. (2004b). WHO consultation on priority public health inverventions before and during an influenza pandemic. Technical report, World Health Organization. http://www.who.int/csr/disease/avian_influenza/final.pdf.