Proceedings of the Twenty-Third International Florida Artificial Intelligence Research Society Conference (FLAIRS 2010)

Ontology-Based Text Mining for Predicting Disease Outbreaks

Nicolae Dragu
Trinity College
Hartford, Connecticut
nicolae.dragu@trincoll.edu

Fouad Elkhoury*
University of Hartford
West Hartford, Connecticut
elkhoury@hartford.edu

Takunari Miyazaki*
Trinity College
Hartford, Connecticut
takunari.miyazaki@trincoll.edu

Ralph A. Morelli*
Trinity College
Hartford, Connecticut
ralph.morelli@trincoll.edu

Nicolás di Tada
InSTEDD
Buenos Aires, Argentina
nditada@instedd.org

Abstract

We have implemented an ontology-based text-mining tool for predicting disease outbreaks. This tool is designed to be used as a free and open-source plug-in for InSTEDD's interactive biosurveillance system *Riff*.

Availability. This tool, in its source code, is freely available from `http://code.google.com/p/e-dop/`.

Introduction

Biosurveillance systems monitor, integrate and analyze various forms of information of potential value to detect and predict infectious disease outbreaks and bioterrorism events. In our efforts to prepare for and respond to public health emergencies, which have become increasingly prominent issues, a great deal of interest has emerged in such systems in recent years (see, e.g., (Center for Disease Control and Prevention 2009)).

Many early biosurveillance systems have been designed to deal with specific data sources and thus applied detection algorithms specifically designed for them. However, much less effort has been focused on how to deal with diverse data types and how human users should effectively interact with multiple systems. The *Riff* project, by InSTEDD, is a solution to deal with such complex diversity issues of biosurveillance. Riff is an interactive decision support environment that combines the power of human experts and biosurveillance services to allow its users to collaborate around streams of information to detect, characterize and respond to emerging threats, providing more comprehensive analysis and deeper insight than what conventional systems provide (Kass-Hout and di Tada 2009).

In this note, we propose a basic architecture of an interactive ontology-based text-mining tool to analyze online news reports concerning potential infectious disease outbreaks. Ultimately, this tool is designed to be incorporated into Riff as a plug-in module. At the moment, we have implemented a functional prototype, and we report herein how it works. To provide the most possible flexibility and interactivity with human users, a novel feature of this tool is an ability to accommodate multiple disease ontologies described in the *Web Ontology Language* (OWL). In our prototype, we have chosen the highly-regarded *BioCaster Ontology* (BCO) (Collier et al. 2006) as default. To navigate and manage an ontology, we have employed *Protégé-OWL* (Musen and others 2009) as an underlying OWL platform. All of this project (including BioCaster and Protégé) is free and open source.

This work was completed while the first two authors (undergraduates) participated in the 2009 Humanitarian FOSS Summer Institute at Trinity College.

Underlying tools

In text mining, ontologies have been widely used as means to define explicit computable semantics; in effect, ontologies enable semantic search and furthermore making intelligent inferences.

BioCaster Ontology (BCO). The use of ontologies has been shown to be particularly effective in specialized knowledge domains, including biosurveillance. An important example in this regard is the *BioCaster Ontology* (BCO) (Collier et al. 2006). This multilingual ontology, covering several European and East Asian languages, is a central component of the *BioCaster* project: an international effort to build a text-mining system for outbreak surveillance (particularly in East Asia) lead by the National Institute of Informatics in Tôkyô (Collier et al. 2008).

BCO has a relatively shallow hierarchy. At the top level, it has a foundation ontology consisting of very general classes common to many general-purpose ontologies. The mid level consists of target entity classes such as *Disease*, *Syndrome* and *Symptom*. Relations associate diseases with symptoms, pathogens with organs, pathogens with transmission modes, etc. All of this is encoded in a single OWL file, freely available from `http://biocaster.nii.ac.jp/`.

Protégé. A number of semantic editors are available to work with OWL ontologies such as BCO. Our choice is *Protégé*, a popular free and open-source platform to create,

*Supported in part by NSF Grant CCF-0722137.

visualize and manipulate ontologies in various formats, including OWL (Musen and others 2009). Most importantly, we have chosen Protégé for its rich functionality, in particular, its ability to support plug-ins and Java-based APIs for building knowledge-based applications. The choice is also partly motivated by the fact that BCO itself has been developed by Protégé. To navigate and manage BCO's OWL file, we have used the *Protégé-OWL* editor.

Basic architecture

Our overall goal is to design an ontology-based application to analyze online news reports concerning potential infectious disease outbreaks. The application's main task is to augment the knowledge of a news article by (i) comparing the article's words against the ontology's terms and then (ii) assigning scoring metrics for relevant diseases and syndromes based on the comparison. Ultimately, such metrics will collectively provide Riff an ability to identify inherent clustering of seemingly unrelated news reports.

A high-level description of the information flow through our application is given in Figure 1. The application ingests plaintext documents (i.e., news articles) one at a time. It also loads an OWL ontology (e.g., BCO) chosen by the user.

The first step of document analysis is to parse a given document and assign frequencies to each word of the document. We then match these words against the ontology's terms that describe *symptoms*. For each matched symptom, we navigate through the ontology to determine all *diseases* and *syndromes* that have this particular symptom. If a matched symptom is very specific to a certain disease (resp. syndrome), then it is easy to determine the most likely disease (resp. syndrome). In most cases, however, a single symptom (such as *fever*) is associated with many diseases and syndromes. A disease or syndrome with the most matched symptoms receives the highest score in the final output.

Simple symptoms are often defined by single words (such as cough), but there are also many symptoms defined by phrases (such as stuffy nose). To identify multi-word symptoms, the prototype is capable of performing n-*gram* matching of input phrases against an ontology's multi-word terms (for important related work, see (Conway et al. 2008)).

All of this was implemented in Java using Eclipse.

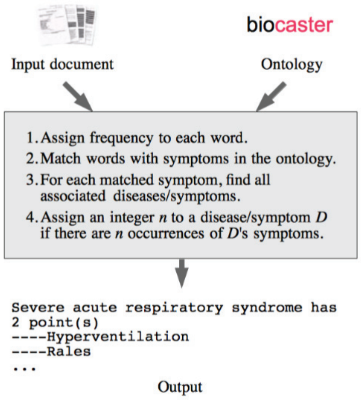
Future work

As the natural starting point, our present prototype focuses on deriving diseases and syndromes from symptoms found in given documents. Naturally, BCO also has a rich collection of terms for *diseases* and *syndromes*. Our immediate plan is to extend our implementation to improve our scoring mechanism by also performing direct matching of input words against an ontology's disease and syndrome terms.

Another area of improvement is to add an ability to recognize stemmed words (such as coughing). We also wish to test our implementation with other OWL ontologies.

Acknowledgement

The authors would like to acknowledge Trishan de Lanerolle for technical assistance and fruitful discussions.


biocaster
Input document
Ontology
1. Assign frequency to each word.
2. Match words with symptoms in the ontology.
3. For each matched symptom, find all associated diseases/symptoms.
4. Assign an integer n to a disease/symptom D if there are n occurrences of D's symptoms.
Output

```
Severe acute respiratory syndrome has
2 point(s)
----Hyperventilation
----Rales
...
```

Figure 1: Overview of information flow

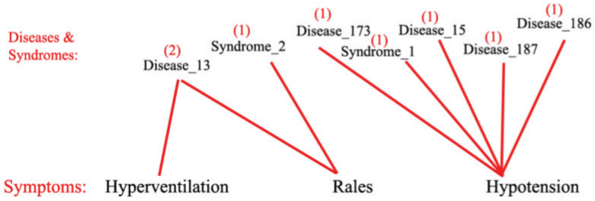


Figure 2: Matching symptoms with diseases and syndromes

References

Center for Disease Control and Prevention. 2009. *BioSense*. Atlanta. (`http://www.cdc.gov/BioSense/`).

Collier, N.; Kawazoe, A.; Jin, L.; Shigematsu, M.; Dien, D.; Barrero, R. A.; Takeuchi, K.; and Kawtrakul, A. 2006. A multilingual ontology for infectious disease surveillance: rationale, design and challenges. *Language Resources and Evaluation* 40:405–413.

Collier, N.; Doan, S.; Kawazoe, A.; Goodwin, R. M.; Conway, M.; Tateno, Y.; Ngo, Q.-H.; Dien, D.; Kawtrakul, A.; Takeuchi, K.; Shigematsu, M.; and Taniguchi, K. 2008. Biocaster: detecting public health rumors with a web-based text mining system. *Bioinformatics* 24:2940–2941.

Conway, M.; Doan, S.; Kawazoe, A.; and Collier, N. 2008. Classifying disease outbreak reports using n-grams and semantic features. In Salakoski, T.; Rebholz-Schuhmann, D.; and Pyysalo, S., eds., *Proceedings of the Third International Symposium on Semantic Mining in Biomedicine (SMBM 2008), Sept. 1–3, 2008, Turku*, 29–36. Turku: Turku Centre for Computer Science.

Kass-Hout, T., and di Tada, N. 2009. *Riff*. InSTEDD, Palo Alto, Calif. (`http://instedd.org/evolve/`).

Musen, M., et al. 2009. *Protégé,* ver. 4.0. Stanford Center for Biomedical Informatics Research, School of Medicine, Stanford University, Stanford, Calif. (`http://protege.stanford.edu/`).