A web-based platform to support text mining of clinical reports for public health surveillance

Annie T. Chen[1]*, Wendy Chapman[2], Mike Conway[2] and Brian Chapman[2]

[1]University of North Carolina at Chapel Hill, Chapel Hill, NC, USA; [2]University of California, San Diego, La Jolla, CA, USA

Objective

We describe the development of a web based application—PyConTextKit—to support text mining of clinical reports for public health surveillance.

Introduction

PyConTextKit is a web-based platform that extracts entities from clinical text and provides relevant metadata—for example, whether the entity is negated or hypothetical—using simple lexical clues occurring in the window of text surrounding the entity. The system provides a flexible framework for clinical text mining, which in turn expedites the development of new resources and simplifies the resulting analysis process. PyConTextKit is an extension of an existing Python implementation of the ConText algorithm (1), which has been used successfully to identify patients with an acute pulmonary embolism and to identify patients with findings consistent with seven syndromes (2).

Public health practitioners are beginning to have access to clinical symptoms, findings and diagnoses from the EMR. Making use of these data is difficult, because much of it is in the form of free text. Natural language processing techniques can be leveraged to make sense of this text, but such techniques often require technical expertise. PyConTextKit provides a web-based interface that makes it easier for the user to perform concept identification for surveillance.

Methods

PyConTextKit's annotation lexicon can be derived from existing lexicons or ontologies and then used to extract concepts relevant to a particular domain or syndrome. In this case, the symptoms from the Syndromic Surveillance Ontology (3) and the Extended Syndromic Surveillance Ontology (ESSO) (4) have been imported into PyConTextKit. Users can create their own text classifier by porting concepts from ESSO and by adding new concepts. Concepts are ultimately mapped to standardized vocabularies like the Unified Medical Language System.

PyConTextKit currently supports the following six features: view of documents to be annotated,management of a lexicon, document annotation using the lexicon,view of annotation results, document classification based on the annotations and summary statistics generation.

PyConTextKit allows the user to manage a lexicon for extraction targets, such as symptoms. It also allows the user to manage a lexicon for modifiers, such as negation cues (e.g., 'no' and 'absence of') and temporality cues (e.g., 'history of'). The modifiers are applied to the targets by pyConTextKit during the annotation phase, and the user can determine the criteria for extraction of a target from a report based on the modifiers. For example, the user may only want to extract symptoms that occurred recently and not historically. The document classification feature identifies documents containing the targets and modifiers specified by the user. For instance, the user may want to identify documents with recent and nonnegated instances of respiratory symptoms and diagnoses.

Finally, PyConTextKit also enables the user to view summary statistics such as the number of documents in the dataset meeting the specified criteria. If the application were run on a dataset involving patients from a particular population, for example, the user could view the number of patients meeting the criteria in that population.

Conclusions

PyConTextKit is aimed at a clinical audience attempting to apply NLP to clinical reports. The strength of PyConTextKit lies in its flexibility in incorporating new knowledge, its hopefully intuitive interface and the sophistication of its document-level analysis.

Keywords

Natural language processing; ontology; text mining

References

1. Chapman WW, Chu D, Dowling JN. ConText: An algorithm for identifying contextual features from clinical text. Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing. Prague. Czech Republic 2007:81–8.
2. Chapman BE, Lee S, Kang, HP, Chapman WW. Document-level classification of CT pulmonary angiography reports based on an extension of the ConText algorithm. J Biomed Inform. 2011;44:728–37. DOI: 10.1016/j.jbi.2011.03.011.
3. Okhamatovskaia A, Chapman WW, Collier N, Espino J, Buckeridge D. SSO: the syndromic surveillance ontology. Proceedings of the International Society for Disease Surveillance, Miami, USA; 2009.
4. Conway M, Dowling J, Chapman WW. Developing an application ontology for mining free text clinical reports: the Extended Syndromic Surveillance Ontology. Proceedings of the Workshop on Health Document Text Mining and Information Analysis. Bled, Slovenia; 2011:75–82.

***Annie T. Chen**
E-mail: atchen@email.unc.edu

Citation: Emerging Health Threats Journal 2011, **4**: 1120 - DOI: 10.3402/ehtj.v4i0.11120