ISDS

Emerging Health Threats Journal 2011, **4**:s68. **doi:** 10.3134/ehtj.10.068

www.eht-journal.org

ABSTRACT

Challenges in adapting an natural language processing system for real-time surveillance

WW Chapman, M Conway, JN Dowling, F-C Tsui, Q Li, LM Christensen, H Harkema, T Sriburadej, and JU Espino

Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA
E-mail: wendy.w.chapman@gmail.com

Objective
Adapt an existing natural language processing (NLP) system to be a useful component in a system performing real-time surveillance.

Introduction
We are developing a Bayesian surveillance system for real-time surveillance and characterization of outbreaks that incorporates a variety of data elements, including free-text clinical reports. An existing NLP system called Topaz is being used to extract clinical data from the reports. Moving the NLP system from a research project to a real-time service has presented many challenges.

Methods
We describe the approaches we are taking to address the challenges, along with results of the current implementation on findings relevant for Shigellosis and influenza surveillance.

Modeling relevant knowledge
Which diseases should we monitor, what clinical information provides evidence for those diseases, how can we represent the knowledge so that collaborators in public health, clinical medicine, knowledge engineering, and software development have access to consistent information? We are developing an application for ontology/thesaurus-relating diseases of interest to public health practitioners to relevant findings potentially described in clinical reports.[1] The thesaurus provides a single point of reference used to manually design a Bayesian case detection model and represent target concepts the NLP system will extract.[2,3]

Creating an efficient web service for the NLP system
We are receiving real-time HL-7 messages from the University of Pittsburgh Medical Center and must process reports as they become available; however, the NLP system was much too slow for real-time processing. We increased efficiency by only loading a small portion of the UMLS Metathesaurus and by replacing the MetaMap indexing module with an implementation of IndexFinder, which is substantially faster. Moreover, we lacked a communication mechanism between the NLP system and the Bayesian case detector. We created an xml schema based on the concepts in the thesaurus for standardized NLP output that can be parsed by the case detector.

Building a framework for tracking performance and diagnosing errors
Assessing errors and updating the NLP system's lexicon and rules were initially accomplished through a painstaking process involving translating output to tab-delimited files, calculating outcome measures in Excel, identifying reports with errors from the Excel files, and examining sentences causing the errors in separate text documents. We are building a framework to simplify results review that automatically calculates agreement between two xml files (using the schema described previously). To quickly identify errors, a user can click on a cell in the contingency table and find the reports contributing to the cell's number. Clicking on a report brings up the text of the report and shows the annotations made by the human reference standard and by the NLP system.

Results
Using the results review analysis tool we have iteratively identified errors on 53 target concepts and made changes to the NLP system using a development set of 26 laboratory-verified Shigellosis cases and 20 laboratory-verified influenza cases. Positive predictive value (PPV) and sensitivity are 84 and 47% for Shigellosis concepts and 89 and 67% for influenza concepts. Sensitivity is consistently lower than PPV, indicating an incomplete lexicon. We are developing a lexicon-building tool to mine synonyms from clinical documents and the UMLS Metathesaurus.

Conclusions
Implementing an NLP system for a real-time surveillance application is more difficult than we expected it to be, and the

challenges are caused less by NLP technicalities than they are by the lack of a robust environment for NLP engineering, adaptation, and improvement. We are developing processes for knowledge engineering and modeling, communication standards for applications using NLP system output, and a dashboard-like interface for performing results review, error analysis, and results tracking. We hope the tools and standards will be generalizable to other surveillance projects and can be used for NLP system adaptation in other domain areas.

Acknowledgements

The project is funded by CDC grants P01 HK000086 and 1U38 HK000063-01. This paper was presented as an oral presentation at the 2010, International Society for Disease Surveillance Conference, held in Park City, UT, USA on 1–2 December 2010.

References

1 Conway M, Dowling J, Chapman W. *Developing a Biosurveillance Application Ontology for Influenza-Like-Illness*. Proceedings of the OntoLex Workshop: Beijing, 2010.
2 Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp* 2001; 17–21. Available at http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2243666/.
3 Zou Q, Chu WW, Morioka C, Leazer GH, Kangarloo H. IndexFinder: a method of extracting key concepts from clinical texts for indexing. *AMIA Annu Symp Proc* 2003; 763–7. Available at http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1480259/.