Concept Negation in Free Text Components of Vaccine Safety Reports

Herman Tolentino MD [1,2], Michael Matters PhD, MPH [1,3], Wikke Walop PhD [4], Barbara Law MD [3], Wesley Tong [5], Fang Liu MS [6], Paul Fontelo MD, MPH [6], Katrin Kohl MD, PhD, MPH [7], Daniel Payne PhD, MSPH [2]

[1] Public Health Informatics Fellowship Program, Career Development Division, Office of Workforce and Career Development, Centers for Disease Control and Prevention Atlanta GA 30333; [2] Bacterial Vaccine-Preventable Diseases Branch, Epidemiology and Surveillance Division, National Immunization Program, Centers for Disease Control and Prevention, Atlanta GA 30333; [3] Division for Heart Disease and Stroke Prevention, National Center for Chronic Disease Prevention and Health Promotion, Centers for Disease Control and Prevention, Atlanta GA 30341; [4] Immunization & Respiratory Infections Division, Centre for Infectious Disease Prevention & Control, Public Health Agency of Canada, Ottawa, Ontario K1A 0K9; [5] Honours Biology and Pharmacology Programme, McMaster University, Hamilton, Ontario L8S 4L8; [6] Office of High Performance Computing and Communications, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894; [7] Immunization Safety Office, Office of the Chief Science Office, Centers for Disease Control and Prevention, Atlanta GA 30333

Abstract

Large amounts of information are locked up in free text components of clinical reports. Surveillance systems that monitor adverse events following immunizations (AEFI) can utilize these components after concept extraction using natural language processing (NLP). Specifically, our method for the identification and filtering of negated concepts using the Unified Medical Language System (UMLS) potentially improves the quality of AEFI surveillance systems.

Introduction

Clinical narrative reports such as discharge summaries, radiology reports, pathology reports, admission histories and physical examinations contain large amounts of clinical information that are locked up in their free text components [1]. AEFI reports submitted to pharmacovigilance information systems contain similar free text components that can be mapped to domain-specific concepts related to AEFI syndrome classes and disease states. The UMLS Semantic Network (SN) provides consistent categorization of concepts found in the UMLS Metathesaurus (MT) [2]. MT concepts retrieved from free text can potentially be used to classify AEFI reports according to specific syndromes. However, the use of concepts for classification requires their negated forms be identified and filtered out to reduce false positives [3]. This paper describes negation in AEFI reports and the method we applied for its detection with the help of UMLS SN semantic types.

Methods

We formed a collaborative workgroup which includes the CDC Vaccine Analytic Unit, members of the Brighton Collaboration (BC) from the CDC Immunization Safety Office and the Public Health Agency of Canada (PHAC), and National Library of Medicine (NLM) researchers. We then created a vaccine safety corpus using de-identified AEFI surveillance reports from the PHAC. To correct misspelled words and to tag free text with UMLS MT concepts we used previously developed algorithms. The outputs of these algorithms became inputs for the negation detection algorithm we developed. Subsequently, we compared the performance of the algorithm against human negation detection.

Results

The algorithm we developed using freely available tools is a combined rule-based and finite state machine algorithm. The following key words were detected if they modified the assertion status of concept-tagged free-text terms classified under certain UMLS SN semantic types: *no*, *neither/nor*, *ruled out*, *denies* and *without*. Using 41 documents with a total of 4,969 concepts, we compared algorithmic with human detection and calculated the algorithm's precision to be 94% and its recall to be 89%.

Conclusion

Concept negation is an important issue that must be addressed for accurate classification of adverse events in AEFI surveillance systems that make use of free text inputs. Preliminary analysis suggests that our algorithm can detect negated concepts reliably for enhancing AEFI surveillance systems. Future enhancements focusing on deeper level semantic parsing can potentially increase the performance of this negation algorithm.

References

1. Hripcsak G, Friedman C, Alderson PO, DuMouchel W, Johnson SB, Clayton P: Unlocking clinical data from narrative reports: a study of natural language processing. Ann Intern Med 1995, 122:681-688.
2. The Unified Medical Language System. URL: http://www.nlm.nih.gov/research/umls/. Accessed on March 2006.
3. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform 2001, 34:301-310.

Acknowledgements

This project was supported in part by an appointment to the Public Health Informatics Fellowship Program sponsored by the Office of Workforce and Career Development at the Centers for Disease Control and Prevention (CDC) administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and CDC.