Emerging Health Threats Journal 2011, **4**:s19. **doi:** 10.3134/ehtj.10.019

www.eht-journal.org

ABSTRACT

Finding medically unexplained symptoms within VA clinical documents using v3NLP

G Divita[1,2], and QZ Trietler[1,2]

[1] *Salt Lake City VA Health Care System, Salt Lake City, UT, USA; and* [2] *University of Utah, Salt Lake City, UT, USA*
E-mail: guy.divita@hsc.utah.edu

Objective

Pro-WATCH (protecting war fighters using algorithms for text processing to capture health events), a syndromic surveillance project, includes a task to identify medically unexplained symptoms. The v3NLP entity extraction tool is being customized to identify symptoms, then to assign duration assertions to address part of this project. The v3NLP tool was recently enhanced to find problems, treatments, and tests for the i2b2/VA challenge. The problem capability is being further refined to find symptoms. Machine learning models will be developed using an annotated corpus currently in development to find duration assertions.

Introduction

Pro-WATCH (protecting war fighters using algorithms for text processing to capture health events), a syndromic surveillance project for veterans of operation enduring freedom (OEF)/operation Iraqi freedom (OIF), includes a task to identify medically unexplained symptoms (MUS). The v3NLP entity extraction tool is being customized to identify symptoms within VA clinical documents, and then refined to assign duration. The identification of medically unexplained symptoms and the aggregation of this information across documents by patient's is not addressed here.

Methods

The v3NLP tool, previously known as HITEXt,[1] includes the capability to identify medical statements from the notes sections within VA clinical documents. The v3NLP is built using GATE[2] pipelines. It includes cTAKES[3] POS tagger and noun phrase parser. The v3NLP recently adopted NLM's MetaMap[4] phrase to concept mapping tool to map to Unified Medical Language System (UMLS) concepts (http://www.nlm.nih.gov/research/umls/). The v3NLP includes a section identification component.

The tool has been further developed to address the 2010 i2b2/VA NLP challenge. This challenge called for the identification of problems, treatments and tests within clinical documents. The challenge also called for the assignment of given assertions associated with medical problems and the identification of relationships between the problem, treatment and tests. The tool was augmented with a statistical machine learning component to improve its performance at identifying problems, treatments and tests, trained on annotated text. The features used included the presence of frequently occurring salient words, concepts and semantic types returned by MetaMap, generalizations of the semantic types, the parts of speech and words around medical concepts, document type and section headings.

For this Pro-WATCH task, the problems found using v3NLP will be further refined down to symptoms, partially by semantic type assignment, but also possibly by the use of an additional machine learning model, trained using annotations from a training corpus currently under development.

A statistical machine learning model will be developed to address the duration component of the symptom identification. The training corpus is expected to have duration assertions. The presence of salient duration words and concepts will be considered some of the features to learn from.

Conclusion

The VINCI team is leveraging the knowledge gained, the methodology, and the software components from the i2b2/VA challenge to address the ProWATCH MUS task. It will further extend v3NLP's capabilities based on machine learning from human annotations.

Acknowledgements

This paper was presented as a poster at the 2010 International Society for Disease Surveillance Conference, held in Park City, UT, USA on 1–2 December 2010.

References

1 Zeng QT, Goryachev S, Weiss S, Sordo M, Murphy SN, Lazarus R, *et al.* Extracting principal diagnosis, co-morbidity and smoking

status for asthma research: evaluation of a natural language processing system. *BMC Med Inform Decis Mak* 26 July 2006;**6**:30.

2 Cunningham H, Maynard D, Bontcheva K, Tablan V. *GATE: a framework and graphical development environment for robust NLP tools and applications*. Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics 2002.

3 Savova GK, Kipper-Schuler K, Buntrock JD, Chute CG. UIMA-based clinical information extraction system. LREC 2008: towards enhanced interoperability for large HLT systems: UIMA for NLP; 2008; Marrakech, Morocco; 2008.

4 Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp* 2001, pp. 17–21.