Emerging Health Threats Journal 2011, **4**:s76. **doi:** 10.3134/ehtj.10.076
© 2011 G Divita *et al.*; licensee Emerging Health Threats Journal.
www.eht-journal.org

ABSTRACT

Standardization to aid interoperability between NLP systems

G Divita[1,2], QZ Treitler[1,2], SM Meystre[1,2], B South[1,2], S Shen[1,2], R Cornia[1,2], J Garvin[1], M Samore[1,2], J Nebeker[1,2], S DuVall[1,2], J Potter[1,2], S Lee[2], P Haug[2], B Bray[2], W Chapman[3], H Harkema[4], L D'Avolio[5], P Elkin[6], M Tuttle[6], G Savova[7], A Coden[8], M Tanenblatt[8], I Sominsky[8], C Clark[9], J Friedlin[10], D Demner-Fushman[11], TC Rindflesch[11], A Aronson[11], AC Browne[11], M Fiszman[11], and O Bodenreider[11]

[1]*VA Salt Lake City Health Care System, Salt Lake City, UT, USA;* [2]*University of Utah, Salt Lake City, UT, USA;* [3]*University of California San Diego, San Diego, CA, USA;* [4]*University of Pittsburgh, Pittsburgh, PA, USA;* [5]*VA Boston Healthcare System, Boston, MA, USA;* [6]*Mount Sinai, New York, NY, USA;* [7]*Harvard Medical School, Boston, MA, USA;* [8]*IBM T.J. Watson Research Center, Hawthorne, NY, USA;* [9]*Mitre Corporation, Boston, MA, USA;* [10]*Regenstrief Institute, Indianapolis, IN, USA; and* [11]*NLM, Bethesda, MD, USA*
E-mail: guy.divita@hsc.utah.edu

Objective

The Consortium for Healthcare Informatics Research, a Department of Veterans Affairs (VA) Office of Research and Development is sponsoring the development of a standard ontology and information model for Natural Language Processing interoperability within the biomedical domain.

Introduction

There are a number of Natural Language Processing (NLP) annotation and Information Extraction (IE) systems and platforms that have been successfully used within the medical domain.[1] Although these groups share components of their systems, there has not been a successful effort[2] in the medical domain to codify and standardize either the syntax or semantics between systems to allow for interoperability between annotation tools, NLP tools, IE tools, corpus evaluation tools and encoded clinical documents. There are two components to a successful interoperability standard: an information and a semantic model.

Methods

Platform-specific information models, such as the UIMA CAS, GATE annotation graphs, and to a lesser degree, Protégé frames and Knowtator,[1] have been adapted to serve as the information model. However, each includes complexity and/or verbosity that has hindered wide adoption. A GATE-lite syntax is under development through this effort, with an *Annotation* at the center of the syntax.

Design principles have emerged around the representation of each *Annotation*: decouple the message syntax from the semantics; design for (space) efficiency; use a standoff annotation model. Create a model is simple enough to be adopted by a wide community, yet be expressive enough to encode a clinical document, a named entity, relationships between entities, and be able to represent temporal features. This model should be usable within annotation tasks, information extraction tasks, and document and corpus evaluation tasks.

Each *Annotation* is intrinsically typed with a tag name or category from a tag set. The naming convention for tags and their semantics are the focus of the other component of the interoperability standard. Semantics have been a stumbling block of other efforts. Under specification of entities such as *Sentence*, *Phrase* and *Token* has led to incomplete integrations at best. This effort includes definition and naming conventions from several established standards to define document structure and named entity semantics. Clinical Document Architecture[3] standard is being used to define clinical document structure, and to represent named entities at the annotation level. This is being augmented with tags from the Penn Treebank parsing and tagging guidelines[4] to define types of phrases, and part-of-speech tags. The model that has emerged includes components such as *Token* and *Term* that have otherwise not been defined within adoptable standards. Elements of these standards have been augmented to allow for relationships between named entities, and for post-coordinated concept coverage within a named entity.

Discussion

A draft should be available for discussion in September 2010. A reference version 0.0 should be available shortly thereafter. This effort is not just an academic exercise for Consortium for Healthcare Informatics Research. The Veteran's Informatics and Computing Infrastructure initiative within

Veterans Affairs has a real need to have an interlingua between external NLP, annotations and IE systems, as well as for in-house development.

Acknowledgements

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 Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? *J of Biomed Inform. Biomedical Natural Language Processing* 2009;**42:**760–72.
2 Ide NM, Suderman K. Bridging the Gaps: Interoperability for GrAF, GATE, and UIMA. *Proceedings of the Third Linguistic Annotation Workshop, ACL-IJCNLP* 2009, pp 27–34. www.aclweb.org/anthology/W/W09/W09-3004.pdf.
3 Dolin R, Alschuler L, Boyer S, Beebe C, Behlen FM, Biron PV, *et al.* HL7 clinical document architecture, release 2. *J Am Med Inform Assoc* 2006;**13:**30–9.
4 Penn Treebank Guidelines. http://www.cis.upen.edu/~treebank.