Biomedical Informatics Insights

EDITORIAL

Computational Semantics in Clinical Text

Stephen Wu

Mayo Clinic, Rochester, MN. Corresponding author email: wu.stephen@mayo.edu

Introduction

This special issue of Biomedical Informatics Insights presents the full paper proceedings of the first workshop on Computational Semantics in Clinical Text (CSCT), held in 2013. Along with Nigam Shah and Kevin Bretonnel Cohen, my co-organizers, I am grateful for BII's willingness to produce this forward-looking publication.

To the Medical Informaticist

Medical informatics has awakened to the reality that natural language processing (NLP) is here to stay. There is great potential for simplicity and efficiency in using structured data such as billing codes, patient metadata, lab values, and medication orders. However, it cannot be ignored that humans are communicative creatures who deal with a world of nuance, ambiguity, and presupposition, including when they describe a symptom or make a diagnosis. Thus, even the most structured clinical data is described or augmented by non-trivial natural language descriptions, and NLP techniques[1–5,8,9,11] have begun to tap into this crucial source of clinical information for clinical, translational, and public health research.

With this Special Issue we would assert that computational semantics is an indispensable sub-discipline of NLP—and perhaps the primary one—for medical informatics, principally because the semantic meaning embedded in clinical text is what any medically-oriented person is after. While medical professionals rarely have something to say about parse trees from clinical text, they have plenty of intuition of what symptoms were present, during what time frame, and with what degree of certainty. We aim here to highlight some of the analysis and techniques that are possible when viewing clinical text from a rigorous computational semantics perspective. In doing so, we hope to encourage increased adoption of computational semantics techniques and resources.

Biomedical Informatics Insights 2013:6 (Suppl. 1) 3–5

doi: 10.4137/BII.S11847

This article is available from http://www.la-press.com.

To the Computational Linguist

Clinical text is a domain very well suited to both exploration in and application of computational semantics. Medical professionals generate clinical text with respect to constrained pragmatic settings, describing, for example, physician-patient encounters and laboratory tests. This has a pervasive effect on the linguistics of the text,[7] from the scoping and intent of negation (usually, to assert that some named entity is not present) to the frequency of syntactic constructions (eg, sentence fragments and semi-structured text). Furthermore, the semantics of clinical language are largely connected to the real world; many medical entities and events actually correspond to characteristics of a patient. Despite this grounding, a full spectrum of linguistic characteristics is available, including term ambiguity, hedging, and paraphrasing.

Clinical NLP is a relatively well-resourced domain. Numerous ontologies, thesauri, terminologies, code sets, and classifications each encode the curated knowledge of domain experts. The Unified Medical Language System (UMLS) Metathesaurus[6] brings together over 150 of these resources, including the Systematized Nomenclature of Medicine—Clinical Terms (SNOMED-CT), the Medical Subject Headings (MeSH), and the International Classification of Diseases, Ninth Revision (ICD-9). Additionally, annotated corpora have been developed[10,12] that overcome the previously difficult question of patient confidentiality in data access.

Finally, the very clear potential for real world impact attracts many to NLP. Computational semantics models and approaches can be implemented in a very pragmatic context, namely on domain-specific tasks such as cohort discovery (ie, find patients that meet my criteria), patient summarization (ie, show relevant information about this patient across records), and clinical decision support (ie, present the right knowledge to the right person at the right time). These real world tasks often produce concrete results, and we often get the added benefit of measurable extrinsic evaluation for our methods.

Review Process and Articles

In the review process for CSCT 2013 and this special issue, we requested that authors make their work anonymous and completed a double-blind review by 24 qualified scientists with expertise both in natural language processing and biomedical informatics. We took special care to avoid any conflicts of interest both in reviewing and in the assigning of reviewers.

We received 9 submissions and accepted 4 long papers for the CSCT Workshop. Accepted long papers were then extended and revised for inclusion in this Special Issue; these are joined by an invited paper by CSCT's first keynote speaker. The CSCT Workshop included 3 additional short papers that are not part of this Special Issue.

There is range of topics discussed in this issue's articles. Two publications deal with distributional semantics methods. The first, *Investigating Topic Modelling for Therapy Dialogue Analysis,* is by Howes, Purver, and McCabe. Outcomes in psychiatric therapy are often influenced by doctor-patient communication, and this article considers whether the high-level topics predict various patient outcomes. Automatic topics from Latent Dirichlet Allocation (LDA) are compared against manually coded topics for predictiveness.

In their article *Using Empirically Constructed Lexical Resources for Named Entity Recognition*, Jonnalagadda et al. also explore distributional semantics methods, but do so on a lexical level rather than on a topic level. They automatically create vector semantics-based lexical resources—a pseudo-lexicon, word clusters, and a pseudo-thesaurus—which they then employ toward the named entity recognition task, with promising results. Similar to Howes et al, they compare these methods to more traditional, manually curated lexical resources.

Similarly, Zweigenbaum et al.'s invited paper, *Combining an expert-based medical entity recognizer with a machine-learning system: Methods and a case study*, provides detailed analyses of named entity recognition methods. This empirical comparison tests the frequently asked question of whether (and how) expert knowledge and data-driven approaches can be combined to improve performance on the NER task.

Towards Converting Clinical Phrases into SNOMED CT Expressions takes a different approach in addressing the connection between a manually curated resource and named entities present in clinical text. To codify out of vocabulary phrases from a corpus of clinical text, Kate formulates the novel task of relation identification, which is distinct from relation extraction; phrases are automatically defined within

SNOMED CT by connecting them to other SNOMED CT concepts via SNOMED CT relationships.

Instead of relating named entities to other ontological concepts, Sohn et al.'s *Analysis of Cross-Institutional Medication Description Patterns in Clinical Narratives* deals with the attributes associated with medication named entities (such as dosage, frequency, and duration) that are observable in nearby context. In this CSCT Best Paper, the authors perform a corpus analysis detailing the semantic patterns that are used to describe medications in multiple institutions, suggesting that the controlled sublanguage of medication descriptions may be extracted with high precision.

I trust that you will enjoy these high-quality articles and the way in which they explore the possibilities for computational semantics in clinical text.

Funding

This work was supported in part by National Science Foundation ABI:0845523, National Institute of Health 5R01LM009959, and NIH Roadmap Grant U54 HG004028.

Competing Interests

Author(s) disclose no potential conflicts of interest.

Disclosures and Ethics

As a requirement of publication the author has provided signed confirmation of compliance with ethical and legal obligations including but not limited to compliance with ICMJE authorship and competing interests guidelines, that the article is neither under consideration for publication nor published elsewhere, of their compliance with legal and ethical guidelines concerning human and animal research participants (if applicable), and that permission has been obtained for reproduction of any copyrighted material.

References

1. Aronson A. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In Proceedings of the AMIA Symposium, page 17. American Medical Informatics Association, 2001.
2. Aronson AR, Lang FM. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*. 2010;71(3):229–36.
3. Denny JC, Irani PR, Wehbe FH, Smithers JD, Spickard A. The knowledgemap project: development of a concept-based medical school curriculum database. In Proceedings of the AMIA Symposium, pages 159–9. American Medical Informatics Association, 2003.
4. Friedman C. A broad-coverage natural language processing system. In Proceedings of the AMIA Symposium, pages 334–44. American Medical Informatics Association, 1998.
5. Friedman C, Hripcsak G. Evaluating natural language processors in the clinical domain. *Methods in Information in Medicine*. 1998;37(4–5): 334–44.
6. Lindberg D, Humphreys B, McCray A. The unified medical language system. *Methods of information in Medicine*. 1993;32(4):281.
7. Meystre S, Savova G, Kipper-Schuler K, Hurdle J. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform*. 2008;3:128–44.
8. Sager N, Lyman M, Bucknall C, Nhan N, Tick LJ. Natural language processing and the representation of clinical data. *Journal of the American Medical Informatics Association*. 1994;1(2):142–60.
9. Savova G, Masanz J, Ogren P, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*. 2010;17(5):507.
10. Uzuner O, South BR, Shen S, DuVall SL. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*. 2011;18(5):552–6.
11. Zeng QT, Goryachev S, Weiss S, Sordo M, Murphy SN, Lazarus R. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. *BMC Med Inform Decis Mak*. 2006;6:30.
12. Albright D, Lanfranchi A, Fredriksen A, et al. Towards syntactic and semantic annotations of the clinical narrative. *Journal of the American Medical Informatics Association*. 2013;0:1–9.