

Natural language processing: algorithms and tools to extract computable information from EHRs and from the biomedical literature

doi:10.1136/amiajnl-2013-002214

Lucila Ohno-Machado, *Editor-in-chief*, Prakash Nadkarni, Kevin Johnson

The increasing adoption of electronic health records (EHRs) and the corresponding interest in using these data for quality improvement and research have made it clear that the interpretation of narrative text contained in the records is a critical step. The biomedical literature is another important information source that can benefit from approaches requiring structuring of data contained in narrative text. For the first time, we dedicate an entire issue of *JAMIA* to biomedical natural language processing (NLP), a topic that has been among the most cited in this journal for the past few years. We start with a description of a contest to select the best performing algorithms for detection of temporal relationships in clinical documents (*see page 806*), followed by a general review of significance and brief description of commonly used methods to address this task (*see page 814*). Top performing approaches are featured in seven articles from five different countries—Canada (*see page 843*), China (*see page 849*), France (*see page 820*), Serbia (*see page 859*), and the US (*see page 828, 836, 867*).

The field of biomedical NLP is quickly evolving. However, an important bottleneck for NLP research is the availability of annotated samples for building and testing new algorithms. Savova (*see page 922*) describes the construction of a large open-access corpus of annotated clinical narratives with high inter-annotator agreement to promote this NLP research. Because training and employing annotators is expensive, solutions that minimize the need for accumulating a large number of annotated documents are needed. Xu (*see page 1001*) shows that active learning (a technique to

select which items should be annotated first to maximize classifier performance) can reduce the amount of labeling by almost 40% without significant performance degradation. Developing new NLP algorithms and approaches and applying them effectively to real clinical problems is the next step.

Contemporary NLP is built by a combination of rule-based and machine learning systems that frequently employ support vector machines and conditional random fields. These and related methods are illustrated in this special issue as being very valuable in addressing tasks as diverse as longitudinal analysis of pain in metastatic prostate cancer (*see page 898*), finding falls in ambulatory care documents (*see page 906*), identifying attributes for medications in clinical notes and clinical trial announcements (*see page 915*), linking medications and their indications (*see page 954*), discovering health information technology incidents (*see page 980*), and identifying adverse drug events (*see page 947*) and allergies (*see page 962*).

Adaptation of general NLP algorithms and tools to the clinical domain is often necessary. NLP techniques can improve on existing processes for concept identification for disease normalization (*see page 876*). Ferraro (*see page 931*) quantifies the increased performance when general part-of-speech tagging algorithms are specifically adapted to the clinical domain, and Wright (*see page 887*) describes adequate cross-institutional performance of an NLP-based system that categorizes clinical notes. Specialized articles in this special issue focus on specific NLP tasks such as word sense disambiguation

(*see page 882*) and co-reference resolution (*see page 891*) in clinical text.

NLP is also used to extract information from biomedical literature. An example is an application to support clinician information needs (*see page 995*). The next step is to harmonize the extracted concepts using standards. For example, allergy information terminologies are evaluated by Goss (*see page 969*), and LOINC mapping is described by Vandenburg (*see page 940*). These standardized concepts are then used within frameworks that enable interoperability (*see page 986*).

The biomedical informatics field has grown rapidly in the past few years, and an established sub-specialty of biomedical NLP now exists, consisting of a vibrant group of professionals that includes researchers and practitioners. While NLP at large is widely recognized by the computer science community, it is important that the biomedical informatics community also recognizes the efforts of those who helped this sub-specialty mature and establish itself as a critical component of biomedical informatics. Dr Carol Friedman, a pioneer and active researcher in biomedical NLP, has trained many of these professionals. Some of these individuals and their teams are represented in this issue, and several others had their articles published in recent issues of the journal. We thank the biomedical NLP community for past, present, and future contributions to *JAMIA*. We look forward to editing another special issue on NLP next year.

Competing interests None.

Provenance and peer review Commissioned; internally peer reviewed.



Natural language processing: algorithms and tools to extract computable information from EHRs and from the biomedical literature

Lucila Ohno-Machado, Prakash Nadkarni and Kevin Johnson

J Am Med Inform Assoc 2013 20: 805

doi: 10.1136/amiajnl-2013-002214

Updated information and services can be found at:

<http://jamia.bmj.com/content/20/5/805.full.html>

These include:

Email alerting service

Receive free email alerts when new articles cite this article. Sign up in the box at the top right corner of the online article.

Notes

To request permissions go to:

<http://group.bmj.com/group/rights-licensing/permissions>

To order reprints go to:

<http://journals.bmj.com/cgi/reprintform>

To subscribe to BMJ go to:

<http://group.bmj.com/subscribe/>