Natural Language Processing and Clinical Outcomes: The Promise and Progress of NLP for Improved Care

Save to myBoK

By Hilary Townsend, MSI

The emergence of electronic health records (EHRs) has necessitated the use of innovative technologies to facilitate the transition from paper-based records for healthcare providers. Natural language processing (NLP) is one technology providers are turning to for improving clinical outcomes and simplifying data entry.

Structured vs. Unstructured Data

Within the EHR, data is captured in one of four ways:

- Entering data directly-including templates
- Scanning documents
- Transcribing text reports created with dictation or speech recognition
- Interfacing data from other information systems such as laboratory systems, radiology systems, blood pressure monitors, or electrocardiographs

Clinical data is represented in structured and unstructured form. Structured data is created through constrained choices in the form of data entry devices including drop-down menus, check boxes, and pre-filled templates. This type of data is easily searchable and aggregated, can be analyzed and reported, and is linked to other information resources. However, it is not always sufficient in allowing individualization of the EHR.

Unstructured clinical data exists in the form of free text narratives. Provider and patient encounters are commonly recorded in free-form clinical notes. Free text entries into the patient's health record give the provider flexibility to note observations and concepts that are not supported or anticipated by the constrained choices associated with structured data. However, unstructured text narrative must be transformed into structured data if HIM professionals want to analyze the data and use it to improve care. This is one useful role for natural language processing.

NLP and Word Sense Disambiguation

NLP is a technology that extracts data from free text. In the clinical setting, NLP converts providers' notes and narratives into structured, standardized formats. This is valuable to HIM professionals because directly processing text with computer applications allows organizations to use clinical documentation data to improve communication between caregivers, reduce the cost of working with clinical documentation, and automate the coding and documentation improvement processes.[1]

Clinical text poses a challenge to NLP. This text is often ungrammatical, consists of "bullet point" telegraphic phrases with limited context, and lacks complete sentences. Clinical notes make heavy use of acronyms and abbreviations, making them highly ambiguous. The Unified Medical Language System (UMLS) Metathesaurus is a repository of over 100 biomedical vocabularies. Within the Metathesaurus, terms across vocabularies are grouped together based on meaning, forming concepts. Each individual concept is assigned multiple semantic type categories from the UMLS Semantic Network. Within UMLS, 33.1 percent of abbreviations have multiple meanings. The presence of abbreviation ambiguity is even higher in clinical notes, with a rate of 54.3 percent.[2]

Word sense disambiguation poses a challenge in extracting meaningful data from unstructured text. Clinical notes often contain terms or phrases that have more than one meaning. For example, "discharge" can signify either bodily excretion or release from a hospital; "cold" can refer to a disease, a temperature sensation, or an environmental condition. Similarly, the abbreviation "MD" can be interpreted as the credential for "Doctor of Medicine" or as an abbreviation for "mental disorder."

The ability to infer the intended meaning of words makes it much easier for computers to find useful patterns in mountains of data. Word sense ambiguity is a pervasive problem in the noisy text of clinical notes. Extracting structured patient information regarding symptoms, tests, and procedures is dependent on being able to assign correct interpretations to the relevant words.[3]

Researchers at the Massachusetts Institute of Technology's (MIT) Computer and Science and Artificial Intelligence Laboratory presented a new system for disambiguating the senses of words used in doctors' clinical notes at the annual American Medical Informatics Association's (AMIA) symposium in November 2012. The system presents a fundamentally alternative approach to word sense disambiguation and promises to significantly reduce the burden on human effort required to develop more accurate systems.

The MIT research employs an area of research known as topic modeling, which seeks to automatically identify the topics of documents by inferring relationships among prominently featured words. Data-driven approaches involving the development of algorithms used to infer patterns require a learning phase, which can be supervised or unsupervised. Supervised learning requires each item in training data to be labeled with the correct answer, while unsupervised learning processes try to recognize patterns automatically.

Topic modeling requires limited human oversight, allowing researchers to continually refine and revise their algorithm and incorporate more features.[4] One such feature is integrating listings from the UMLS into the algorithm, along with other linguistic features and word associations established by the National Institutes of Health's paper on medical subject headings, available at www.nlm.nih.gov/mesh/meshrels.html. The new word sense disambiguating system is on average 75 percent accurate in disambiguating words with two senses.

NLP Uses Reach Beyond Structured Data

There are significant implications for improved accuracy in word sense disambiguation for analytics and patient outcomes. Enhanced language processing ability improves HIM processes like clinical decision support and clinical documentation improvement programs through an increased ability to accurately mine clinical documents and fill gaps and ambiguities in documentation.[5]

Decreased ambiguity in clinical data allows for more insightful extraction of data, and the ability to infer words' intended meanings makes it much easier for computers to find useful patterns in mountains of data. Identifying such patterns can facilitate meaningful analytics, like those used by IBM's Watson Supercomputer technology.

IBM's Patient Care and Insights software offering, for example, is using advanced analytics to surface new insights and early, personalized intervention opportunities and is turning these insights into action with accountable case-based patient-centered care management.

The content analytics collects and analyzes structured and unstructured data, while similarity analytics uses NLP and machine learning technologies from the Watson project to analyze thousands of variables for a patient's condition and medical history, and to generate comparison to others with similar conditions and potential outcomes.[6]

MIT's new system for word sense disambiguation has the potential to impact clinical data analytics through an increased ability to accurately infer meanings of extracted data. Not only does this hold value for improved patient outcomes, it can support more accurate billing and improved clinical workflows through enhanced clinical documentation improvement programs.

Notes

1. Wolniewicz, Richard. "Auto-Coding and Natural Language Processing." 3M Health Information Systems, 2011. http://www.himss.org/content/files/AutoCodingandNaturalLanguageProcessing(WhitePaper).pdf.
2. Youngjun, Kim et al. "Using UMLS Lexical Resources to Disambiguate Abbreviations in Clinical Text." AMIA Annual Symposium Proceedings, 2011. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3243121/.
3. Chasin, Rachel et al. "Using UMLS for Word Sense Disambiguation in Clinical Notes." AMIA Annual Symposium Proceedings, 2012. http://proceedings.amia.org/2p9act/.

4. Hardesty, Larry. "Mining physicians' notes for medical insights." *MITnews*. http://web.mit.edu/newsoffice/2012/digital-medical-records-offer-insights-1031.html.

5. Dooling, Julie A. "Advancing Technology Connects Transcription and Coding: The Developing Role of NLP, NLU, and CAC in HIM." *Journal of AHIMA* 83, no.7 (July 2012): 52-53.

6. Goldberg, Michael. "IBM Makes New Health Care Push with Predictive Analytics, Process Management." *Data Informed*. http://data-informed.com/ibm-makes-new-health-care-push-with-predictive-analytics-process-management/.

Hilary Townsend (hilary.townsend@ehealthinitiative.org) is policy analyst at eHealth Initiative.

Article citation:
Townsend, Hilary. "Natural Language Processing and Clinical Outcomes: The Promise and Progress of NLP for Improved Care" *Journal of AHIMA* 84, no.2 (March 2013): 44-45.

Driving the Power of Knowledge