Name:	Alexandre Yahi
Affiliation:	Department of Biomedical Informatics, Columbia University
Email:	alexandre.yahi@columbia.edu
Presentation type (select one):	Poster

Natural Language Processing in Clinical and Translational Research: Integrating Context-Dependent Modifier Combinations Across Diverse Note Types

Alexandre Yahi, MS[1], Ning Shang, PhD[1], Nicholas P. Tatonetti, PhD[1], Noémie Elhadad, PhD[1], George Hripcsak, MD, MS[1]

[1]Department of Biomedical Informatics, Columbia University, New York, NY, USA

Abstract

Phenotyping is a central problem in clinical and translational studies using the Electronic Health Records (EHR). With more complex and comprehensive approaches, integrating structured and unstructured clinical data has become critical. As the Common Data Model (CDM) maintained by the Observational Health Data Sciences and Informatics (OHDSI) is evolving to represent Natural Language Processing (NLP) outputs in its schema, the integration of this data requires a careful semantical characterization to fully leverage the wealth of information contained in narrative clinical notes. In this study, we focused on terms related to diseases and symptoms for 1,685 patients across 38,110 notes parsed with cTAKES, an open-source, state-of-the-art tool for concept identification and normalization from notes. We demonstrated that modifier combinations need to be used in concert with note sections to capture what is true for the patient at the time of the note. We also suggested that NLP would benefit from a non-relational data structure for novel approach in machine learning.

Introduction

The secondary use of the Electronic Health Records (EHR) has enabled researchers to use data science in unprecedented way for clinical and translational studies. However, the medical characterization, or phenotyping, of patients or cohort of patients to support these studies needs to be accurate and reliable.

Therefore, phenotyping has become a central question, with efforts made to build clinical algorithms to stratify patients[1] and distribute them on platforms such as PheKB[2]. As the desired granularity increases, more complex and heterogeneous sources of data need to be integrated together, constituting a major challenge.[3] The Clinical Data Model (CDM) [4] is an illustration of the importance of standards to represent structured clinical data. With common data structure and tools, the Observational Health Data Sciences and Informatics (OHDSI)[5] fosters the ambition to distribute suite of applications and share clinical data to generate evidence in healthcare. One of the most recent successes of this international initiative is the recent study on treatment pathways for type 2 diabetes mellitus, hypertension and depression by Hripcsak et al. across the whole OHDSI network[6].

As the CDM is evolving to integrate the output of Natural Language Processing in its schema, it is important to evaluate the challenges of the integration of NLP outputs into research workflow. To do so, we evaluated if the modifiers provided by NLP tools are enough to identify medical terms that represent the truth for the patient at the time of the note. More specifically, we studied the impact of the additional semantic layer provided by note sections on diseases and symptoms modifiers.

Data and Methods

We focused on the clinical notes of 1,685 patients at Columbia University Medical Center/New York Presbyterian (CUMC/NYP) from August 1999 to July 2014. These patients belong to a pilot cohort consisting of individuals with available genotyping data. Therefore, they represent an example of cohort that could be used for precision medicine research studies. It represents 38,110 documents across 114 note types. We parsed these clinical notes using the clinical Text Analysis and Knowledge Extraction System (cTAKES)[7]. cTAKES is an open-source NLP tool relying on existing open-source technologies such as the Unstructured Information Management Architecture (UIMA) framework and OpenNLP toolkit. This tool's named entity recognition (NER) annotator implements a terminology-agnostic dictionary look-up algorithm and maps each concept to terminologies including SNOMED CT and RxNORM. The modifiers annotated by cTAKES were: polarity, uncertainty, conditional, generic, subject, historyOf. For the section identification, we manually curated note sections based on the paragraph headers and classified them into 20 categories. These 20 section header types were based on the most frequent section headers found in the 114 notes types studied. We grouped some headers together (e.g., "PMH psychology" in "PMH") in order to have

sections non-specific to hospital services but reflecting semantic properties in terms of modifiers. We also grouped section together when the outcome would not have an important impact on the term modifiers, or is a synonym (e.g. “Diagnosis” grouped with “Assessment”).

Results

We were able to associate a note section to the following count of terms: past medical history and past surgical history (P/P) 19865, family history (FHX) 26317, problem list (PROBLEM) 32965, summary before the first section (GENERAL) 199151, social history (SHX) 16564, physical examination (PE) 192246, review of systems (ROS) 67937, past medical history (PMH) 81262, chief complain (CC) 23749, studies (STUDY) 20343, past social history (PSH) 10590, no section identified (N/A) 245289, laboratory tests (LABS) 91702, history of present illness (HPI) 117904, vital signs (VITAL) 40343, assessment and plan (A/P) 198810, allergies (ALL) 31909, health care maintenance (H/M) 7549, medications (MEDS) 86698, immunization (IMMU) 2047.

In terms of modifiers combinations captured by cTAKES, the most frequent tuple was “positive, non-conditional, certain, non-generic, for the patient, in the present” 1,247,997 times, followed by “negative, non-conditional, certain, non-generic, for the patient, in the present” 215,330 times.
However, these modifiers are not enough to interpret the complete semantic of the terms capture by the NLP engine. In the section relative to history (i.e., PSH, P/P, PMH, SHX, FHX), the modifier “historyOf” was triggered only 4.65% of the time. Conversely, for the family history section, the modifier “family_member” was only captured 15.35% of the time. By taking into account the note sections into the filtering process to keep what is true for the patients at the time of the note, we went from 1,247,997 to 816,836 terms, representing a 34.5% decrease.

Conclusion

In conclusion, we demonstrated that the combination of modifiers identified by NLP engines is not enough to classify parsed terms. They should be used along with note sections, in particular to modulate modifiers relative to past/present and patient/family. In this study, limitations came from both the NLP pipeline used for the term matching and form the manual curation of sections. This latter process could be performed in a supervised manner as demonstrated by Li et al. using Hidden Markov Model (HMM)[8] and using the SecTag terminology proposed by Denny et al.[9] to identify and classify section headers.

Given the format of the NLP outputs usually presented in nested structure such as XML or JSON, and the organization by document, we can also wonder if current relational structures in the CDM are appropriate for this type of information. Non-relational solution would have the advantage to be closer to the format of NLP engine outputs. Moreover, as novel approaches like Deep Patient by Miotto et al. [10] start using deep learning to abstract clinical information, non-relational structure could ease the process of structuring NLP data into tensors for neural networks.

References

1. Conway M, Berg RL, Carrell D, Denny JC, Kho AN, Kullo IJ, et al. Analyzing the heterogeneity and complexity of Electronic Health Record oriented phenotyping algorithms. AMIA Annu Symp Proc. American Medical Informatics Association; 2011;2011:274–83.
2. Kirby JC, Speltz P, Rasmussen LV, Basford M, Gottesman O, Peissig PL, et al. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. J Am Med Inform Assoc. 2016 Mar 28;:ocv202.
3. Liao KP, Cai T, Savova GK, Murphy SN, Karlson EW, Ananthakrishnan AN, et al. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. BMJ. British Medical Journal Publishing Group; 2015 Apr 24;350(apr24 11):h1885–5.
4. Reisinger SJ, Ryan PB, O'Hara DJ, Powell GE, Painter JL, Pattishall EN, et al. Development and evaluation of a common data model enabling active drug safety surveillance using disparate healthcare databases. Journal of the American Medical Informatics Association. The Oxford University Press; 2010 Nov;17(6):652–62.
5. Hripcsak G, Duke JD, Shah NH, Reich CG. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. Studies in health …. 2015.
6. Hripcsak G, Ryan PB, Duke JD, Shah NH, Park RW, Huser V, et al. Characterizing treatment pathways at scale using the OHDSI network. Proc Natl Acad Sci USA. National Acad Sciences; 2016 Jun 6;:201510502.
7. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc. The Oxford University Press; 2010 Sep 1;17(5):507–13.
8. Li Y, Lipsky Gorman S, Elhadad N. Section classification in clinical notes using supervised hidden markov model. the ACM international conference. New York, New York, USA: ACM; 2010. 7 p.
9. Denny JC, Spickard A, Johnson KB, Peterson NB, Peterson JF, Miller RA. Evaluation of a Method to Identify and Categorize Section Headers in Clinical Documents. J Am Med Inform Assoc. The Oxford University Press; 2009 Nov 1;16(6):806–15.
10. Miotto R, Li L, Kidd BA, Dudley JT. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. Sci Rep. Nature Publishing Group; 2016 May 17;6:26094.