Natural language processing and machine learning for detecting and characterizing patient phenotypes from their clinical records

Xavier Tannier
xavier.tannier@sorbonne-universite.fr

November 29, 2019

1 Context

Clinical data warehouses (CDWs) are becoming widespread in France and in the world, bringing together vast amounts of data on patients' journeys in the hospital (currently 50 million reports at AP-HP, partner of the project).

Despite a desire to structure patients' records at the source, more than 80% of hospital data are collected in the form of texts, mainly in medical reports (clinical notes)[6]. These documents, written in natural language, by humans and for humans, are still very difficult to analyze and therefore to value. This is due to the variation of language in general, but also the technical nature of documents, whose vocabulary varies strongly from one medical specialty to another.

There is a major challenge to extract an exploitable informational value from these texts, such as personal and family history, lifestyle, symptoms, signs, diagnoses, acts, results of biological or imaging analyzes, drug or non-drug treatments. Once extracted, another challenge is to mix them with the available structured data, in order to obtain a comprehensive representation of the patient.

Finally, a last subject is to query these concepts and representations in order to find patients with given characteristics (phenotyping) or retrieving similar medical cases.

This is not only a research question for natural language processing (NLP), but also a hindrance to the development of data-driven AI approaches in e-health, since much of the information about patients is contained in these textual notes. Once structured and integrated into a CDW, this data could reinforce many approaches already used in statistics, data mining, machine learning, AI in general. This concerns tasks in medical research (phenotyping, pharmacovigilance, epidemiology, prediction, phenome-genome association studies, clinical research...), care (visualization of the care pathway, similar patient retrieval), medico-economic management (value-based healthcare)... Efforts in this direction are currently very language-dependent [7, 1, 2, 5], and even hospital-dependent, because of the variety of medical terminologies used in the hospitals, and of the confidentiality of the data, preventing from sharing labeled resources.

In this context, we propose several possible subjects for a 4-6-month internship around two axes of NLP research within LIMICS laboratory and Sorbonne Center for Artificial Intelligence (SCAI):

- Apply current state of the art NLP methods to medical problems brought by our partners from different hospitals (2.1).
- Explore more general, weakly supervised approaches to information extraction and representation learning from medical reports (2.2, 2.3).

2 Subjects

2.1 Epifractal project: detecting patients with a low-energy fracture

The project aims to detect elderly patients who have had a low-energy fracture due to osteoporosis problems, with the aim of offering them personalized follow-up and reducing the risk of relapse. We will build a semi-automatic system for selecting these patients, based of their hospitalization or emergency room textual reports. A gold standard of 1,500 patients have been constituted for training. The objective is to maximize the recall of this system, in order to propose a prefiltering system to the human expert, who will make the final decision.

This will involve supervised machine learning, probably with a little help of a rule-based preprocessing. This project requires frequent contacts with the medical experts.

2.2 Active learning for medical concept characterization

The task of medical concept extraction includes:

- concept mention identification (or named entity recognition, i.e. find the text spans in a clinical record that are mentions of a medical concept)
- concept normalization (or entity linking, i.e. link the textual mentions to existing concepts in a terminology or an ontology)
- mention characterization, i.e. determine whether the concepts, in the context of the patient record, are negated, uncertain, of high or low intensity, or relate to a person other than the patient (family member or third party). These aspects can obviously fundamentally change the understanding of the text.

The third part (characterization) has only been handled with rule-based systems so far [3, 4]. The use of classical active learning approaches will here be beneficial and will greatly reduce the necessary annotation time.

2.3 Patient-level representation learning

Vector representations (embeddings) of words, sentences and documents are now massively used in NLP and information retrieval. These dense representations, obtained without supervision, allow to reduce the dimensions of the objects while implicitly conveying their meaning (the semantically close words will be close in the vector space) as well as the relationships they maintain between them. As for medical reports, the question of the representation of the patient (patient embedding) has a particular interest. Information about hospitalized patients are very heterogeneous: structured data such as lab results; time series produced by continuous monitoring;

textual notes written by the health care team. Tasks using data mining or statistical learning, which require collecting variables of interest on patients, would benefit from an aggregated representation that is easy to manipulate. This is the case, for example, with similar patient retrieval, automatic indexing, selection of cohorts of patients sharing characteristics (phenotyping), or prediction of the patient's disease or outcome, the quantification of risk... these tasks being common research topics in medical informatics.

We will first perform a supervised model aimed at optimizing the representation of a patient for diagnosis prediction, as already explored for English data; the code for diagnosis is available for each stay and will be used for training and evaluation. We will then explore an unsupervised approach that we will compare with the supervised one. The final objective (out of the scope of this internship) is to use structured data jointly with textual data. Indeed, another benefit of using dense vectors to represent objects is that it is possible to embed features of heterogeneous natures into the same vector space.

References

[1] Alan R. Aronson and Franois-Michel Lang. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association : JAMIA*, 17(3):229–236, 2010.

[2] Chlo Cabot, Romain Lelong, Julien Grosjean, Lina F. Soualmia, and Stfan J. Darmoni. Retrieving Clinical and Omic Data from Electronic Health Records. *Studies in health technology and informatics*, 221:115, 2016.

[3] W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, 34:301–310, October 2001.

[4] Wendy W Chapman, Dieter Hillert, Sumithra Velupillai, Maria Kvist, Maria Skeppstedt, Brian E Chapman, Mike Conway, Melissa Tharp, Danielle L Mowery, and Louise Deleger. Extending the NegEx lexicon for multiple languages. *Studies in health technology and informatics*, 192:677–681, 2013.

[5] Georg Dietrich, Jonathan Krebs, Georg Fette, Maximilian Ertl, Mathias Kaspar, Stefan Strk, and Frank Puppe. Ad Hoc Information Extraction for Clinical Data Warehouses. *Methods of information in medicine*, 57:22–29, May 2018.

[6] Preethi Raghavan, James L. Chen, Eric Fosler-Lussier, and Albert M. Lai. How essential are unstructured clinical narratives and information fusion to clinical trial recruitment? *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, 2014:218–223, 2014.

[7] Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association : JAMIA*, 17(5):507–513, 2010.