Panel: Clinical Natural Language Processing in Languages Other Than English

Aurélie Névéol¹, PhD, Hercules Dalianis², PhD, Guergana Savova³, PhD, Pierre Zweigenbaum¹, PhD

¹ CNRS, LIMSI – rue John von Neumann, 91400 Orsay, France
² DSV, Stockholm University, Forum 100, 164 40 Kista, Sweden
³ Children's Hospital Boston and Harvard Medical School, Boston, Massachusetts, USA

{neveol;pz}@limsi.fr, hercules@dsv.su.se, Guergana.Savova@childrens.harvard.edu

Abstract

Natural Language Processing (NLP) of clinical free-text has received a lot of attention from the scientific community. Clinical documents are routinely created across health care providing institutions and are generally written in the official language(s) of the country these institutions are located in. As a result, free-text clinical information is written in a large variety of languages. While most of the efforts for clinical NLP have focused on English, there is a strong need to extend this work to other languages, for instance in order to gain medical information about patient cohorts in geographical areas where English is not an official language. Furthermore, adapting current NLP methods developed for English to other languages may provide useful insight on the generalizability of algorithms and lead to increased robustness. This panel aims to provide an overview of clinical NLP for languages other than English, as for example French, Swedish and Japanese and discuss future methodological advances of clinical NLP in a context that encompasses English as well as other languages.

General Description of the Panel

The goal of this panel is to engage the medical informatics and clinical Natural Language Processing community in a discussion about ways to advance research through languages other than English. We will provide an overview the current state of clinical NLP in a variety of European and non-European languages as well as focused reports on French, Swedish and Bulgarian. We will motivate the need for developing clinical NLP in languages other than English by the potential for methodological and medical advances. Finally, we will propose strategies to contribute to advance work on languages other than English and integrate it in a state-of-the-art platform.

Clinical NLP in languages other than English

Natural Language Processing (NLP) of clinical free-text has received a lot of attention from the scientific community, demonstrating its potential to provide the means to analyze large quantities of documents rapidly and accurately (Demner-Fushman et al. 2010). Prime clinical applications for NLP include assisting healthcare professionals with retrospective studies and clinical decision making. The ability to analyze clinical text in languages other than English opens access to important medical data concerning cohorts of patients who are treated in countries where English is not the official language. Recently, Kohane et al. (2012) also showed the impact of methods allowing an aggregated exploitation of clinical data. In this context, data extracted from clinical texts in languages other than English adds another dimension to data aggregation.

As the importance of clinical NLP gains recognition, clinical corpora become available to researchers in languages other than English, prompting work that sometimes builds on methods validated for English. Adapting systems that work well for English to another language is a difficult task that may be carried out with varying level of success depending on the task and language (Grouin et al., 2009; Velupillai et al. 2014; Täckström et al., 2012). For non-European languages, approaches that account for entirely different word and sentence structures sometimes need to be developed (Shinohara et al. 2013), and cultural differences between clinical narrative styles accounted for (Wu et al. 2013). Access to terminologies and corpora in languages other than English can also be challenging (Schulz et al. 2013; Xu et al. 2013). These experiments prompt a reflection on how to carry out clinical NLP in a more global context: should methods be developed for one language and then ported to other languages? Can the source language method benefit from the porting? Can algorithms be more robust if they are designed with a multi-language perspective from the start?
French is widely spoken around the world and benefits from one of the largest coverage in the UMLS. Automatic de-identification is becoming quite advanced for French (Grouin & Névéol, 2013), leading to good results for targeted clinical information extraction tasks (Deléger et al. 2010; Grouin et al. 2011). Recent efforts from the French biomedical Informatics community have addressed rules and regulations to improve the access of NLP researchers to clinical corpus. Furthermore, the success of initiatives such as that reported by Grouin et al. (2011) increased the awareness of the potential implication of clinical NLP in clinical practice and contributed to making the timing ripe for making clinical corpus available for annotation and NLP tool development. On-going efforts currently address the annotation of clinical corpora for entity, modality and relations. Tools are being designed for information extraction as well as semantic indexing, information retrieval and clinical data visualization.

Much of the research in Swedish clinical NLP has used the Stockholm EPR Corpus, (Dalianis 2012), that contains more than one million patient records encompassing the years 2006-2010, from over 550 clinical units origin from Karolinska University Hospital. Part of this corpus has been manually annotated for Protected Health Information, negations, uncertainty levels, symptoms, diseases, drugs, body parts and abbreviations. The annotated corpora have been used both for training of machine learning systems and evaluation. Some applications are explorative as comorbidity networks, warning and reporting systems detecting hospital acquired infections or adverse drug events, but also work on text simplification of patient record content for the layman patient, (Dalianis 2012). Tools that have been developed for this is an adaptation of NegEx for Swedish (Skeppstedt 2012), a system for classifying terms into six levels of assertion levels pyConTextSwe, (Velupillai et al. 2014), abbreviation detection, (Isenius et al. 2012) and machine learning system based on CRF++ that recognizes named clinical entities as symptoms, diseases, drugs and body, (Skeppstedt et al. 2014).

Integrating languages other than English in Apache cTAKES
Apache cTAKES (ctakes.apache.org) has been quite successful in assembling and sustaining a global community of developers and users of state-of-the-art English language clinical NLP. Because these techniques involve computational machine learning methods, datasets from the targeted language are needed to train and evaluate the algorithms on. We will discuss what types and size of data were used to build the various cTAKES components – sentence boundary detector, tokenizer, part of speech tagger, syntactic parser, event and temporal expression detector, temporal relation modules, general relation module. We will also discuss what types of gold standard labels (and how much of each type) are needed to port cTAKES components to other language within the light of some use cases such as porting the temporal expression discovery and normalization module originally developed for English (Bethard, 2013) to Swedish. We will outline available resources in other languages such as Swedish, Finnish, Bulgarian. This is a step towards globalization of information extraction from the clinical narrative.

Panelists
Prof. Hercules Dalianis (Professor at Stockholm University, Sweden) will present on work building on the Stockholm EPR Corpus, a major resource for Swedish clinical NLP.
Dr. Aurélie Névéol (staff scientist at LIMSI-CNRS, France) will act as a moderator and will present the medical and methodological benefits of clinical NLP in languages other than English. Dr. Névéol has been leading a project addressing the automatic understanding of French clinical narratives for translational research.
Dr. Guergana Savova (Assistant Professor at Harvard Medical School) will talk about integrating clinical NLP in different languages. Dr. Savova has been leading the development of the core Clinical Text Analysis and Knowledge Extraction System, now part of the Apache Software Foundation (cTAKES; ctakes.apache.org).
Dr. Pierre Zweigenbaum (principal investigator at LIMSI-CNRS, France) will discuss research efforts for clinical NLP in French, many of which he has initiated and coordinated over the past decade.

List of Discussion Points
After the introductory presentations, the moderator will ask questions as well as solicit questions from the audience, to prompt discussion among the panelists. Potential topics and questions include:
- Describe problems and experiences regarding NLP work on clinical text. How language-dependant is the work you are familiar with? What are the specificities of each language?
- Describe instances of a successful NLP application in a language other than English that yielded interesting new medical knowledge
- Discuss the methodological challenges to bringing NLP in languages other than English to the level of state-of-the-art for English. Are some specific languages riper than others? Which languages? Why?
- Describe the technical and organizational challenges that must be overcome to integrate several languages other than English in an NLP platform such as cTAKES

**References**


**Statement of Participation**

The first author affirms that all panel participants have agreed to participate and have contributed to the preparation of this document (as of August 1, 2014)