Multilingual linkage of clinically relevant text

Aurélie Névéol¹, Cyril Grouin¹, Kevin B. Cohen¹,², Sophie Rosset¹, Pierre Zweigenbaum¹

¹CNRS, UPR 3251, LIMSI
91403 Orsay, France
{firstname.lastname}@limsi.fr

²University of Colorado School of Medicine
Aurora, Colorado, USA
kevin.cohen@gmail.com

We suggest the use of the QUAERO French Medical Corpus to link annotated clinically relevant text in French (MEDLINE titles and drug inserts) to similarly annotated text in English. The French corpus comprises manual annotations for 10 types of clinically relevant entities. The French texts are aligned with their English version at the sentence level. Automatic annotations for 9 of the entities are available for the English version of the texts. We challenge BLAH participants to create links between French and English annotations at the entity level.

Data Set: the QUAERO French Medical Corpus

The proposed dataset consists of biomedical text in French, including MEDLINE titles and EMEA drug information leaflets. The text is annotated by 10 types of clinical entities for mentions mapped to concepts in the Unified Medical Language System (UMLS). The data set is called QUAERO French Medical Corpus. It has been developed as a resource for named entity recognition and normalization in 2013 (Névéol et al. 2014). It is distributed under the GNU Free Documentation license (GFDL) http://en.wikipedia.org/wiki/Wikipedia

The data set has been created in the wake of the 2013 CLEF-ER challenge, with the purpose of creating a gold standard set of normalized entities for French biomedical text. A selection of the MEDLINE titles and EMEA documents used in the 2013 CLEF-ER challenge were selected for human annotation and will be used in this challenge. The annotation process was guided by concepts in the Unified Medical Language System (UMLS):

1. 10 types of clinical entities, as defined by the following UMLS Semantic Groups (Bodenreider and McCray 2003) were annotated: Anatomy, Chemical and Drugs, Devices, Disorders, Geographic Areas, Living Beings, Objects, Phenomena, Physiology, Procedures
2. The annotations were made in a comprehensive fashion, so that nested entities were marked, and entities could be mapped to more than one UMLS concept. See (Névéol et al. 2014) for additional details, and corpus excerpts.

Suggested tasks for the QUAERO French Medical Corpus

1. Creating a linkable version of the dataset: the dataset is currently available in standoff annotation format compatible with the BRAT rapid annotation tool (http://brat.nlplab.org/). To increase the usability of the dataset and its ability to be linked to other datasets, an important task will be to convert the dataset in additional formats such as BioC and RDF.
2. Linking gold annotations in French to silver annotations in English: the CLEF-ER 2013 challenge provided silver-standard annotations on the English version of the documents in the QUAERO French Medical Corpus. Links are already available between the parallel texts, at the sentence level. An interesting task would be to create links at the entity level between the entities in the French gold standard and entities in the English silver standard. The results would provide a valuable multilingual resource; it could validate some of the silver standard annotations, and provide material for a contrasted study of biomedical text in French vs. English.

References