Recognition and Normalization of Biomedical Entities Based on Ontologies

André Leal[1], Bruno Martins[2], Francisco M. Couto[1]

[1]LaSIGE, Faculdade de Ciências, Universidade de Lisboa
[2]INESC-ID, Instituto Superior Técnico, Universidade de Lisboa

Clinical notes in the form of textual context occur frequently in Electronic Health Records (EHRs). They are mainly used to describe treatment plans, symptoms, diagnostics, etc. Clinical notes are recorded in narrative language without any structured form and, since each medical professional uses different types of terminologies according to context and to their specialization, these notes are very challenging for their complexity, heterogeneity and contextual need.

Forcing medical professionals to introduce the information in a predefined structure simplifies the interpretation. However, the imposition of such a rigid structure increases not only the time needed to record data, but it also puts some heavy barriers at recording unusual cases.

One possible solution consists on the application of text-mining techniques to the clinical texts, in order to support the recognition and normalization of medical concepts. Together, these techniques can result in the correct and efficient information gathering by information systems.

We developed a system which on a first instance recognizes medical concepts in clinical notes and then normalizes them with a UMLS concept unique identifier (CUI). This system was developed with the intention to overcome some challenges presented in this task, such as the recognition of non-continuous entities and the normalization of ambiguous entities.

For the recognition we use the novel SBIEON encoding which contains a tag to specify words inside recognized entities that are not part of them. We also explore non-annotated clinical notes to generate lower-dimensional representation of the word vocabulary, and therefore reduce the data sparsity. CRF models were generated based on the mentioned features among others, such as domain specific lexicon, token shape,

etc. For normalization we use a rule based approach to normalize the recognized entities and we also take in consideration the information content of each entity for disambiguation. This system was used to participate in SemEval 2015 Task 14, achieving a second place in the competition.

We also intent to explore semantic similarity between entities inside individual clinical notes to improve normalization results. This approach is based on the assumption that entities inside individual clinical notes are similar between them.

Preference for presentation: Oral or Poster (no special preference)

Location: University of Lisbon

Author for Correspondence: leal.andre92@gmail.com