Delivering Clinical Information Extraction Tools to Practitioners

Laurianne Sitbon – Queensland University of Technology
Mahnoosh Kholghi – Queensland University of Technology
Guido Zuccon – Queensland University of Technology
Anthony Nguyen – The Australian e-Health Research Centre (CSIRO)
Bevan Koopman – The Australian e-Health Research Centre (CSIRO)
Michael Lawley – The Australian e-Health Research Centre (CSIRO)

While the machine learning community is demonstrating great performances of automated methods (models) to extract clinical concepts (such as symptoms, medications or tests) from clinical documents such as discharge summaries, the methods that they use are only highly successful for the documents they were developed for, that is the same type of documents and the same use of clinical language (a given dialect of English, a given level of detail in the documents, or a given specialist's perspective). When these models are used for other types of documents, their effectiveness sensibly decreases, thus making them unreliable in actual clinical settings. The machine learning community is investigating this issue of model adaptation, also called transfer learning, but no robust technique has yet transpired. The current practice instead prescribes to provide a somewhat large set of manually annotated data to train a new model in a given context. This often requires large annotation costs and requires people with clinical knowledge that can then be trained to annotate in an appropriate manner.

An alternative to this costly and large annotation process is to consider active learning, i.e. an iterative process where at each iteration only small subsets of informative data are annotated according to a sampling strategy (the query strategy). In this presentation we demonstrate that active learning frameworks can reduce annotation costs by 75%. Such a reduction in annotation costs means that it becomes more affordable to create such high quality customised information extraction models.

The use of an active learning strategy, however, does not fully remove the need for annotation. Instead, it introduces the need for a dynamic annotation process: one where annotator and machine iteratively dialogue, with the algorithm that at each iteration provides data to annotate and the annotator that provides ground truth labels. This implies a commitment from the annotators over a period of time.

An in-depth cost analysis is required to ultimately assess the cost reduction in data annotation provided by the active learning strategy when compared to the traditional supervised learning approach. In clinical information extraction, costs may include the cost for creating an annotation, the cost for examining/revising an automatically provided annotation, the cost for correcting an automatically- provided (but incorrect) annotation, the cost for training annotators. We believe that such a cost analysis would lead to best-practice guidelines and interfaces to support the customization of machine learning based models for clinical information extraction in real-world settings.