Abstract

Knowledge extraction is typically the result of the combination of different NLP techniques, implemented in different tools, extracting various kind of information: e.g., maned entities, semantic frames, temporal expressions, entity types. As this information is independently derived, contradicting content may be produced from the same piece of text. Hence, it is crucial to research and develop techniques to coherently select/harmonize/complement the content returned by the various NLP tools used, in order to improve the overall quality of the knowledge extracted from text. In this research project we investigate a probabilistic model, derived from the ontological knowledge extracted from existing knowledge bases (Yago and DBpedia) and an annotated corpus (AIDA-CONLL), to maximize the joint probability of NERC and entity linking candidate annotations on the same textual mention. After being tested on gold-standard, our model was found to improve significantly the performance of Entity Linking significantly, while the improvement for NERC is relatively not obvious. This project is carried on within the context of a state-of-the-art knowledge extraction framework: PIKES.