Named entity recognition from Arabic Wikipedia

Authors **Mohit Behrang, Kemal Oflazer, Noah Smith**

Institutions Carnegie Mellon University in Qatar, Doha, Qatar
Carnegie Mellon University, Pittsburg, PA, USA

Named Entity Recognition (NER) is the problem of locating mentions to entities such as persons, locations and organizations. The named entity information is helpful for reducing the complexity of monolingual and multilingual processing tasks, such as information extraction, parsing and machine translation. We investigate the Arabic NER problem from the Arabic Wikipedia text. We employ statistical sequence labeling methods for solving the NER task. Previous studies suggest that sequence labeling methods, such as Conditional Random Fields, are the state of the art NER frameworks.

The sequence labeling methods require human labeled training data. Most of the Arabic human labeled data for NER belong to the political news domain and the consequent trained models are biased towards the news domain. In contrast, our target test data (Arabic Wikipedia articles) has a very diverse set of topics. The domain mismatch between the train and test data results in poor NER performance.

In order to reduce the coverage problem, we present three techniques: (1) we use the Wikipedia network structure to collect additional information about the text. Information such as monolingual and cross-lingual hyperlinks and text formatting lead us to use new features of the Wikipedia text in NER models. Moreover, we use cross-lingual projection to collect named entity information from English Wikipedia. (2) We use a domain adaptation technique to shift the model from the baseline political domain to domains relevant to our test data. Our model adaptation uses a small set of in-house-labeled Arabic Wikipedia articles. (3) We use self-training to port from a fully supervised to a semi-supervised learning framework: we collect a large volume of unlabeled Arabic Wikipedia articles to expand the underlying NER domain to new text domains. Our model expansion is gradual and iterative. In each iteration we add a new set of unlabeled articles to the training and use the current model to label and construct a larger model.

Our NER evaluations are based on the standard precision and recall metrics. We evaluate our proposed framework in four different text domains of Arabic Wikipedia.