Collective Approaches to Named Entity Disambiguation

Ayman A. Alhelbawy

A thesis presented for the degree of

Doctor of Philosophy

Department of Computer Science

University of Sheffield

United Kingdom

July, 2014

Abstract

Internet content has become one of the most important resources of information. Much of this information is in the form of natural language text and one of the important components of natural language text is named entities. So automatic recognition and classification of named entities has attracted researchers for many years. Named entities are mentioned in different textual forms in different documents. Also, the same textual mention may refer to different named entities. This problem is well known in NLP as a disambiguation problem. Named Entity Disambiguation (NED) refers to the task of mapping different named entity mentions in running text to their correct interpretations in a specific knowledge base (KB). NED is important for many applications like search engines and software agents that aim to aggregate information on real world entities from sources such as theWeb. The main goal of this research is to develop new methods for named entity disambiguation, emphasising the importance of interdependency of named entity candidates of different textual mentions in the document.

The thesis focuses on two connected problems related to disambiguation. The first is Candidates Generation, the process of finding a small set of named entity candidate entries in the knowledge base for a specific textual mention, where this set contains the correct entry in the knowledge base. The second problem is Collective Disambiguation, where all named entity textual mentions

in the document are disambiguated jointly, using interdependence and semantic relations between the different NE candidates of different textual mentions. Wikipedia is used as a reference knowledge base in this research.

An information retrieval framework is used to generate the named entity candidates for a textual mention. A novel document similarity function (NEBSim) based on NE co-occurrence is introduced to calculate the similarity between two documents given a specific named entity textual mention. NEB-sim is also used in conjunction with the traditional cosine similarity measure to learn a model for ranking the named entity candidates. Na¬ıve Bayes and SVM classifiers are used to re-rank the retrieved documents. Our experiments, carried out on TACKBP 2011 data, show NEBsim achieves significant improvement in accuracy as compared with a cosine similarity approach.

Two novel approaches to collectively disambiguate textual mentions of named entities against Wikipedia are developed and tested using the AIDA dataset. The first represents the conditional dependencies between different named entities across Wikipedia as a Markov network, where named entities are treated as hidden variables and textual mentions as observations. The number of states and observations is huge, and na¬ıvely using the Viterbi algorithm to find the hidden state sequence which emits the query observation sequence is computationally infeasible given a state space of this size. Based on an observation that is specific to the disambiguation problem, we develop

an approach that uses a tailored approximation to reduce the size of the state space, making the Viterbi algorithm feasible. Results show good improvement in disambiguation accuracy relative to the baseline approach, and to some state-of-the-art approaches. Our approach also shows how, with suitable approximations, HMMs can be used in such largescale state space problems.

The second collective disambiguation approach uses a graph model, where all possible NE candidates are represented as nodes in the graph, and associations between different candidates are represented by edges between the nodes. Each node has an initial confidence score, e.g. entity popularity. Page-Rank is used to rank nodes, and the final rank is combined with the initial confidence for candidate selection. Experiments show the effectiveness of using Page-Rank in conjunction with initial confidence, achieving 87% accuracy, outperforming both baseline and state-of-the-art approaches.