Unsupervised Relation Extraction with General Domain Knowledge

Mirella Lapata
University of Edinburgh

Information extraction (IE) is becoming increasingly useful as a form of shallow semantic analysis. Learning relational facts from text is one of the core tasks of IE and has applications in a variety of fields including summarization, question answering, and information retrieval. Previous work has traditionally relied on extensive human involvement (e.g., hand-annotated training instances, manual pattern extraction rules, hand-picked seeds). Standard supervised techniques can yield high performance when large amounts of hand-labeled data are available for a fixed inventory of relation types, however, extraction systems do not easily generalize beyond their training domains and often must be re-engineered for each application.

In this talk I will present an unsupervised approach to relational information extraction which could lead to significant resource savings and more portable extraction systems that require less engineering effort. The proposed model partitions tuples representing an observed syntactic relationship between two named entities (e.g., “X was born in Y” and “X is from Y”) into clusters corresponding to underlying semantic relation types (e.g., BornIn, Located). Our approach incorporates general domain knowledge which we encode as First Order Logic rules. Specifically and automatically combine with we combine a topic model developed for the relation extraction task with automatically extracted domain relevant rules, and present an algorithm that estimates the parameters of this model. Evaluation results on the ACE 2007 English Relation Detection and Categorization (RDC) task show that our model outperforms competitive unsupervised approaches by a wide margin and is able to produce clusters shaped by both the data and the rules.

(Joint work with Oier Lopez de Lacalle)