

Learning to Extract Relations from the Web and Biomedical Corpora

Razvan Bunescu

Univ. of Texas at Austin

<http://www.cs.utexas.edu/users/razvan/>

Tuesday February 15, 2007

2:30-4:00pm,

TECH Center 111

Abstract:

Automatically identifying semantic relationships between entities mentioned in text documents is an important task in natural language processing. The set of relevant relationships can be very diverse, ranging from company acquisitions mentioned in web documents to interactions between human proteins as mentioned in biomedical articles. In this talk I will describe two approaches to learning relation extractors that differ in the type of supervision required. In the first, traditional approach, a machine learning algorithm is given a collection of documents that have been manually annotated for the relation of interest. A subsequence kernel is trained on this type of sentence-level supervision, resulting in a relation extraction system that is effective at mining protein-protein interactions from biomedical abstracts. In the second approach, the amount of supervision is significantly reduced to only a handful of pairs of entities known to exhibit or not exhibit a particular relationship. Bags of sentences containing the pairs are extracted from the web, and an existing relation extraction system is extended to handle this weaker form of supervision. I will present experimental results demonstrating that the new approach can reliably extract relations from web documents.