中央研究院　資訊科學研究所
Institute of Information Science, Academia Sinica

Eduard Hovy

Information Science Institute, University of Southern California

Tuesday, Marh 13, 2012

Auditorim 106 at New IIS Building

14:00 ~ 15:30 A New Semantics: Merging Propositional and Distributional Information

Despite hundreds of years of study on semantics, theories and representations of semantic content—the actual meaning of the symbols used in semantic propositions—remain impoverished. The traditional extensional and intensional models of semantics are difficult to actually flesh out in practice, and no large-scale models of this kind exist. Recently, researchers in Natural Language Processing (NLP) have increasingly treated topic signature word distributions (also called 'context vectors', 'topic models', 'language models', etc.) as a de facto placeholder for semantics at various levels of granularity. This talk argues for a new kind of semantics that combines traditional symbolic logic-based proposition-style semantics (of the kind used in older NLP) with (computation-based) statistical word distribution information (what is being called Distributional Semantics in modern NLP). The core resource is a single lexico-semantic 'lexicon' that can be used for a variety of tasks. I show how to define such a lexicon, how to build and format it, and how to use it for various tasks. Combining the two views of semantics opens many fascinating questions that beg study, including the operation of logical operators such as negation and modalities over word(sense) distributions, the nature of ontological facets required to define concepts, and the action of compositionality over statistical concepts.

PowerPoint download at http://www.iis.sinica.edu.tw/public/workshop/2012/12mar-ExtSem-v1.pdf

16:00 ~ 17:30 Text Harvesting and Ontology Construction using a Powerful New Method

People build databases and metadata structures/ontologies to collect, systematize, and make available to users knowledge in a consistent and hopefully trustworthy form. But the largest data collection today, the web, is not systematic, consistent, or trustworthy, and the access techniques we use are provably inadequate. Over the past decade, various researchers have developed web harvesting methods to extract information from the web and organize it in various ways. Various different methods have been tried, but none has had much success; inconsistencies, knowledge gaps, the need for manual intervention, the lack of gold standard material to evaluate against, and other problems plague the automated harvesting methods. Focusing on unstructured text, I describe a method to extract information from the web, organize it, and form both a knowledge base and its taxonomic term ontology/metadata. The method is competitive to or outperforms existing large-scale information harvesting from the web, and is very simple to implement. In the talk, I also describe some of the deep problems fundamental to ontology building, as they are made apparent in this work.
This is joint work with Dr. Zornitsa Kozareva (USC Information Sciences Institute).

PowerPoint download at http://www.iis.sinica.edu.tw/public/workshop/2012/12mar-LearningOntol.pdf