Toward Semantic Corpora: Creating Concepts from Words via Senses, and Storing them in an Ontology

Eduard Hovy

Information Sciences Institute – University of Southern California
4676 Admiralty Way, Marina del Rey, CA 90292 - USA
hovy@isi.edu

Abstract

Recent work in several computational linguistics (CL) applications (especially question answering) has shown the value of semantics (in fact, many people argue that the current performance ceiling experienced by so many CL applications derives from their inability to perform any kind of semantic processing). But the absence of a large semantic information repository that provides representations for sentences prevents the training of statistical CL engines and thus hampers the development of such semantics-enabled applications. This talk refers to recent work in several projects that seek to annotate large volumes of text with shallower or deeper representations of some semantic phenomena. It describes one of the essential problems—creating, managing, and annotating (at large scale) the meanings of words, and outlines the Omega ontology, being built at ISI, that acts as term repository. The talk illustrates how one can proceed from words via senses to concepts, and how the annotation process can help verify good concept decisions and expose bad ones. Much of this work is performed in the context of the OntoNotes project, joint with BBN, the Universities of Colorado and Pennsylvania, and ISI, that is working to build a corpus of about 1M words (English, Chinese, and Arabic), annotated for shallow semantics, over the next few years.

It can be argued that the current performance ceilings experienced by so many language technology applications stem from their inability to perform semantic processing: speech recognition has not made significant recognition rate improvements for several years; information retrieval remains stuck at around 50% recall and precision in controlled experiments; text summarization systems do not perform materially better on newspaper text than the simple first-paragraph baseline; etc. But recent work in other applications has shown the value of even a limited form of semantics. For example, the open-domain question answering system built by LCC (Harabagiu et al., 2001), that includes some shallow semantic parsing and inference to perform controlled term expansion and answer plausibility ranking, significantly outperforms all other QA systems in evaluations year after year.

Can computer programs be built to produce semantic analyses of arbitrary text automatically, with reliable quality? Experience over the past decade has shown fairly convincingly that methodologies that require humans to build rules by hand tend to fail, while methodologies that employ machine learning algorithms to induce rules from annotated corpora tend to succeed, albeit with variable quality performance; one need mention only part of speech tagging, syntactic parsing, and wordsense disambiguation as examples. So if a large enough semantic corpus were to exist, then it is likely that semantic analyzers could be built; the question is how well they would perform on each semantic phenomenon.

What is required to build a large semantic corpus? Several rather difficult questions must be addressed, including:

Which (aspects of) semantics should be represented? Some phenomena, such as word sense, negation, understanding of numbers and dates, etc., have long been studied in Computational Linguistics (CL), and are immediate candidates for inclusion. Others, such as entailment and discourse structure, are much less understood, and require more study. But not all aspects are equally useful: some phenomena occur in every sentence; others occur perhaps once per discourse.

Which semantic resources would be most useful? Traditional resources in CL include ontologies, semantic zones within lexicons, and sets of rules for handling quantifiers, negation, aspect, mood, etc. But the theories underlying these resources are many and varied, and little consistency exists, despite attempts over a decade to produce standards, for example for the uppermost regions of ontologies. In order to train semantic analyzers, one would require at least a corpus of text, in which each word or sentence is annotated with appropriate semantic information, with high consistency.

How large should the semantic resources be to be effective? Given the complexity of semantics, the training resource will have to be large. In addition, the semantic term 'lexicon' of primitive symbols should contain more than a few hundred symbols. The LCC QA system mentioned above uses a set of some 120,000 axioms, most of them derived semi-automatically from WordNet definitions; perhaps this—the size of an educated person's lexicon in English—could serve as a rough upper bound for at least some kinds of semantic symbol set?

How can one actually go about building a semantically annotated corpus in a systematic way so as to satisfy the competing requirements of broad coverage, high consistency, and interestingly deep semantics? Clearly, this cannot be an automated process, for then we could simply include the process in the CL applications from the start, so it has to involve human effort. But humans are notorious slow, expensive, and inconsistent.

Despite these difficulties, there are encouraging signs of progress. In the past few years, several projects have addressed the task of producing text corpora annotated with limited kinds of (very shallow) semantic information (often in conjunction with syntactic and other information). These projects include the Prague Dependency Treebank (Hajic et al., 2001), the TIGER/SALSA corpus (Burchardt et al., 2006), the Interlingua Annotation for Machine Translation (IAMT) project (Reeder et al., 2004; also see Rambow et al. in this

proceedings), and the OntoNotes project (Ramshaw et al., 2006).

The author has been fortunate enough to be a member of both the IAMT and OntoNotes projects. The IAMT project, a collaboration of six partner institutions (Universities of New Mexico State, Maryland, Columbia, CMU, MITRE, and ISI), lasting just over a year, focused on requirements for the representation of semantics as illustrated by a comparison of seven languages (Arabic, English, French, Hindi, Japanese, Korean, and Spanish). This research aimed at representational depth but compromised on the size of coverage. In contrast, the OntoNotes project, a collaboration of BBN, the Universities of Colorado and Pennsylvania, and ISI, is building a large annotated corpus (eventually, 1 million words) of English, Arabic, and Chinese text, aiming for coverage and compromising on depth.

This talk outlines the general approach taken in both these projects. It outlines the various tasks comprising such annotation efforts, the role of an ontology in the work, and the difficulty of evaluation. It concludes with a discussion of the potential for resource construction to be systemized into a 'science', as well as the problem facing resource builders in writing papers that are sufficiently 'hard' to be accepted in most conferences.

References

Burchardt, A., K. Erk,A. Frank, A. Kowalski, S. Pado, and M. Pinkal. 2006. Consistency and Coverage: Challenges for Exhaustive Semantic Annotation. *Proceedings of DGfS-06*, Bielefeld.

Hajic, J., B. Vidová-Hladká, P. Pajas. 2001. The Prague Dependency Treebank: Annotation Structure and Support. *Proceeding of the IRCS Workshop on Linguistic Databases*, pp. 105–111.

Harabagiu, S., D. Moldovan, M. Pasca, R. Mihalcea, M. Surdeanu, R. Bunescu, R. Girju, V. Rus, P. Morarescu. 2001. The Role of Lexico-Semantic Feedback in Open-Domain Textual Question-Answering. *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL-2001)*, 274–281.

Ramshaw, L., E.H. Hovy, M. Marcus, M. Palmer, S. Pradhan, R. Weischedel. 2006. The 90% Solution. Proceedings of the Human Language Technology conference of the North American Association of Computational Linguistics (HLT-NAACL).

Reeder, F., B. Dorr, D. Farwell, N. Habash, S. Helmreich, E.H. Hovy, L. Levin, T. Mitamura, K. Miller, O. Rambow, A. Siddharthan. 2004. Interlingual Annotation for MT Development. *Proceedings of the AMTA conference.*