Semantics Isn't Easy:
Thoughts on the Way Forward

Nancy Ide*, Rebecca Passonneau, Collin Baker†, Christiane Fellbaum††**

*Vassar College
Poughkeepsie, New York USA

** Columbia University
New York, New York USA

†International Computer Science Institute
Berkeley, California USA

††Princeton University
Princeton, New Jersey USA

E-mail: ide@cs.vassar.edu, becky@cs.columbia.edu, collinb@icsi.berkeley.edu, fellbaum@princeton.edu

We argue that linguistic knowledge acquisition by means of semi/un- supervised statistical methods from huge unannotated corpora is on its own inadequate to grapple seriously with semantic analysis. We do not argue against the approach *per se*, but rather, we urge that this work be undertaken in close collaboration and coordination with work in the field that attempts to get at the more fluid and dynamic aspects of language by performing (semi-)manual (and therefore necessarily smaller) analyses. Our position is that both approaches are not only essential, but can be best leveraged if performed in close coordination.

It is necessary to acknowledge the limitations of a methodology relying on semi/un-supervised learning. In its extreme, it assumes that there is a (single) set of semantic facts (meanings, relations, etc.) that is stable and discoverable, upon which humans can agree (at least, 90% of the time, as recently posulated [Hovy, et al., 2006]), and that all we need do is train the machine to do the same. However, unlike morphological and syntactic properties of language, word and phrasal meaning is fluid, dynamic and, to some degree, generative (Pustejovsky, 1995; Nunberg, 1995). We are far from even a rudimentary understanding of the ways in which this fluidity affects meaning, and we have not undertaken a serious exploration of alternative means to represent and process language that takes it into account. We have also done little to explore fundamentals such as the inter-relations among different levels/types of semantic analysis, which could significantly impact current processing methods. While assembling massive bodies of (static) semantic facts may get us 60%—or even 80%—of language understanding "accuracy", the elusive and far more impenetrable 20% remains the major obstacle to broadly usable language processing capability.

It is also necessary to acknowledge that the data upon which these analyses are based (presumably drawn from the web) has distributional features that are fundamentally unknown. Even if we accept that sheer size of the data will overcome the many sources of noise and account for all or most varieties of and situations for language use and eliminate faulty information, we are nonetheless left with a linguistic black box, in which variations due to differing text types and genres, regional differences and dialects, situations (e.g. written vs .spoken transcript), contexts, and date of production cannot be discovered (and are, indeed, conflated).

Finally, it is necessary to recognize and embrace the need to couple the resources such an undertaking strives to create in a form and format that can interoperate with annotations and other resources already developed and used in the community. The NLP community does not need yet-another-independent-resource that is either difficult or impossible to use in conjunction with other resources because of orthogonality of design and (more basically) incompatible format.

Our comments are offered in the context of our work on development of a Manually Annotated Sub-Corpus (MASC) of the Open American National Corpus (OANC), which includes language data representative of a wider range of genres and language varieties than has been treated in most NLP analyses. MASC incorporates manually validated annotations for a wide variety of phenomena, including word and sentence boundaries, part of speech, noun and verb chunks, named entities, and WordNet senses and FrameNet frames for a subset of lexical items. In addition, MASC includes validated annotations contributed by projects including Penn TreeBank, PropBank, NomBank, and TimeBank for a portion of the texts. All of the annotations are represented in a common stand-off format, and an API is provided that enables merging and comparison of any or all annotations, as well as output in a variety of formats suitable for use with NLP software (NLTK, UIMA, etc.). WordNet sense annotations are assigned through a collaborative, iterative process, whereby initial sense assignments are used to refine/modify the WordNet sense inventory, followed by re-annotation using the refined sense set. WordNet and FrameNet annotations are also being done via a process of mutual, iterative refinement in order to move towards harmonization of the two resources. We are also developing new methods for measuring agreement among annotators that are sensitive to relations among word senses, which will be published along with the corpus, and using criteria derived from our analyses of inter-annotator agreement to evaluate and refine automatic annotation tools. Because it enables merging annotations at different linguistic levels, MASC will facilitate a deeper investigation of interactions among linguistic phenomena than has been possible in the past, including the learnability of one phenomenon given features derived from another, which will contribute to better understanding of the workings of language at the semantic level.

Hovy, E.H., M. Marcus, M. Palmer, S. Pradhan, L. Ramshaw, and R. Weischedel. 2006. OntoNotes: The 90% Solution. Short paper. *Proceedings of the Human Language Technology / North American Association of Computational Linguistics conference (HLT-NAACL 2006)*. New York, NY

Nunberg, G. (1995), Transfers of Meaning, *Journal of Semantics*, 12 (2), 109.

Pustejovsky, J. (1995). *The Generative Lexicon: A Theory of Computational Lexical Semantics*. Cambridge, MA: The MIT Press