Injecting Linguistics into NLP through Annotation

Eduard Hovy
USC/ISI
4676 Admiralty Way
Marina del Rey, CA 90292
USA
hovy@isi.edu

Over the past 20 years, the size of the *L* in Computational Linguistics has been shrinking relative to the size of the *C*. The result is that we are increasingly becoming a community of uninformed but sophisticated engineers, applying to problems very complex machine learning techniques that use very simple (simplistic?) analyses/theories. (Try finding a theoretical account of subjectivity, opinion, entailment, or inference in publications surrounding the associated competitions of the past few years.)

When we grow tired of embarrassing ourselves, what should we do? Fortunately, injecting some linguistic (and other) sophistication into our work is not that complicated. The key is annotation: by using a theoretically informed set of choices rather than a bottom-up naive one, we can have annotators tag corpora with labels that reflect some underlying theories. While the large-C contingent of our community will not care, researchers interested in investigating language rather than processing will be able to find new ways to connect with Corpus Linguists, Psycholinguists, and even Ontologists.

It turns out that many of our surrounding academic communities – Linguists, Political Scientists, Biocurators, etc. – have been performing annotation for years in order to build and prove their theories. They have however been largely unaware of the power of NLP technology and the benefits we can bring to them. There is a natural marriage – several, actually – waiting to happen.

What is the benefit to us? What's wrong with simply continuing to use half-baked annotation schemes to train our machine learning systems on? Several things:

- half-baked schemes generally fail in the long run-that's why more-sophisticated ones are developed
- there are dozens to hundreds of graduate students and young researchers in surrounding communities eager to help build corpora by running annotation efforts and using the problems uncovered while annotating to drive further theory formation
- because they're generally more 'correct', more-sophisticated annotations allow stacking of multiple phenomena upon the same material with fewer internal inconsistencies and problems.

Such stacking eventually enables multi-phenomenon analysis and mutual disambiguation in ways that an incommensurately annotated corpus does not.

Proceedings of the 2010 Workshop on NLP and Linguistics: Finding the Common Ground, ACL 2010, page 79,
Uppsala, Sweden, 16 July 2010.