Combining Knowledge Sources for Automatic Semantic Tagging

Douglas Jones and Boyan Onyshkevych
Natural Langauge Processing Research Branch
Department of Defense, Attn: R525
9800 Savage Road
Ft. Meade, MD 20755-6000 USA
{daj3,baonysh}@afterlife.ncsc.mil

1 Goal of the Session

In this working session, we will discuss methods which could plausibly be used for combining evidence for assigning semantic tags to words in a text. We will discuss methods that apply at knowledge acquisition time to produce a single static knowledge source to be used by a single, complete, semantic tagger, as well as methods for dynamically combining outputs of a set of independent, possibly incomplete, semantic taggers.

A variety of evidence-combination and parallel-hypothesis-selection mechanisms will be considered for the dynamic case, including Dempster-Shafer and Bayesian approaches, best-evidence, and chart management methods. For the static case (at knowledge acquisition time), we will consider the difficulties of merging unrelated knowledge sources.

2 Issues for the Workshop

The main issue for discussion will be the advantages of various methods of combining evidence.

Other issues that could be discussed include:

- Do we assume there is always a single correct tag, or do we allow a set of equally correct tags?
- Do we rank or assign probabilities for all senses?
- Do we tag phrases/collocations/idioms (or just individual tokens)? If so, this complicates evidence combination.
- What preprocessing do we assume as input to the taggers in the dynamic scenario?
- Do we approach evidence combination for homograph distictions differently than for polysemy? Are there other types of differences among senses that might affect evidence combination?
- What are the implications of a sequential combination of evidence vs. a paralel approach for the dynamic scenario?
- How do we map word senses/semantic tags from multiple knowledge sources into a single set in the static knowledge acquisition scenario?

Possible sources of evidence that could be considered for dynamic combination include: domain tags (e.g., LDOCE box codes), collocational and corpus co-occurrence approaches, frequency (domain-specific or domain-independent), selectional restrictions, decision trees, part of speech and subcategorization, Lesk et al dictionary approaches, semantic distance approaches over ontologies, spreading activation/marker passing over semantic nets, scripts/MOPs, word experts.

Possible sources for static combination include: MRD entries, WordNet, Levin verb classes, corpus statistics, and other lexical resources.

In order to constrain the discussion, we will make the following assumptions: Senses for each word have been pre-enumerated (Compare, for example, Pustejovsky or Nunberg, and the references cited in these works which point out difficulties in enumerating senses.) In the dynamic case, we are combining compatible knowledge sources, i.e., they share the a semantic tagset. (Contrast work in combining WordNet and Levin Classes.)

3 Organization of the Session

As preparation for the workshop, participants are encouraged to consider what kinds of knowledge may be combined using these methods and also to consider which methods of combination may be preferable.

If any of our assumptions are too restrictive, potential participants can send their ideas to either of the working session leaders.

All participants who wish to discuss their position regarding the above-mentioned problems are invited to make a very short presentation of their work and how it offers answers to those problems (or why it fails to do so); please send the working session leaders a brief discussion of your position if you intend to make a presentation.

If we get an indication of sufficient interest from workshop participants beforehand, we will have separate periods of discussion for the static and the dynamic scenarios, each starting with these brief position statements.

Prior to the workshop, participants are encouraged to submit discussions of their preferred approach to this issue for distribution to all other working session participants. Participants are also encouraged to submit discussions and examples of the types of evidence to be combined.