Why Doesn't Natural Language Come Naturally?

Richard Schwartz
BBN Technologies
Cambridge, MA 02138
schwartz@bbn.com

Abstract

We have seen great success over the past 15 years in speech recognition. This success is due, largely, to the broad acceptance of Hidden Markov Models (HMMs) at the beginning of that period, which then facilitated rapid and steady improvements in speech recognition that are still continuing today. Although no one believes speech is produced by an HMM, the model affords a rich framework in which to improve the model in a rigorous and scientific manner.

Could we create the same environment for a uniform probabilistic paradigm in NL? It requires several ingredients:

- A uniform notational system to express meanings,
- A statistical model that can represent the associations between meanings and words,
- A training program that estimates parameters from annotated examples,
- An understanding program that finds the most likely meaning given a word sequence, and
- A substantial corpus with meanings annotated and aligned to the words.

These problems are fundamental. In speech recognition, we can all agree that the desired output is a sequence of orthographic words. But in understanding, we lack agreement as to the meaning of meaning. And it gets harder from there, since the structures we must look at are not sequences, but rather trees or more complex structures. Still the goal is a worthwhile one.

We attempt to formulate several different language understanding problems as probabilistic pattern recognition problems. In general, our goal is to rely heavily on corpus based methods and learning techniques rather than on human generated rules. At the same time, it is essential that we be able to incorporate our intuitions about the problem into the model. We choose probabilistic methods as our preferred form of learning technique because they have several desirable properties. First, if we can accurately estimate the posterior probability of our desired result, then we know a decision based on this posterior probability will minimize the error rate. Second, we have a large inventory of techniques for estimation of robust probabilities from finite data. Third, in contrast to classical pattern recognition problems, language deals almost exclusively with sequences (of sounds, phonemes, characters, words, sentences, etc.) Our goal is not to recognize or understand each of these independently, but rather to understand the sequence. Probability theory provides a convenient way to combine several pieces of evidence in making a decision.

We present several language problems for which we have developed probabilistic methods that achieve accuracy comparable to that of the best rule-based systems. In each case we developed a model that is (somewhat) appropriate for the problem. These problems include Topic Classification, Information Retrieval, Extracting Named Entities, and Extracting Relations.