RESEARCH STATEMENT
Kristina Toutanova

Many natural language (NL) understanding problems can be formulated as problems of disambiguation. In recent years, we have seen that statistical machine learning approaches can go a long way toward solving some isolated disambiguation problems, such as finding the correct part of speech of a word in context, or identifying named entities in text. For many of these problems, standard machine learning techniques for independent and sequence classification have yielded high accuracy. My work has focused on advancing natural language processing technologies to address problems such as question answering and information extraction, which require disambiguation at the level of whole sentences and larger units. In order to achieve such "deeper" understanding, we need to model more complex language phenomena. My research has aimed at building and studying novel probabilistic models for sophisticated linguistic representations – for example, phrase structure parse trees, richer unification-based grammar parse trees, and semantic functional structure. I have done this in two ways: (*i*) by choosing representations and models which are more effectively able to encode domain knowledge and (*ii*) by developing machine learning algorithms more suitable for NL disambiguation tasks, which are characterized by sparse training data and large, structured spaces of hidden labels.

Representations and Models for Linguistic Structures

Syntactic parsing, which involves deciding amongst the possible grammatical structures for a sentence, can be considered a classification task where each possible parse is a class. The data representation expected by many classification machine learning algorithms is vectors of real numbers, but parse trees are not readily representable in this form. What features represent a syntactic tree and what defines a similarity between trees makes for most of the difference in performance among systems. The problem of how to break a parse tree into parts which are predictive and which are not too large is a challenging one. In previous work parse tree representations have been centered on local subtrees of the parse tree. I defined a representation of parse trees that captures useful context for disambiguation by increasing the connection of the features of the tree with the words of the sentence [EMNLP 04]. This representation is not based on local subtrees but rather on paths from words to the root of the tree, which can be seen as "the tree from the point of view of a particular word". It also paves the way for exploring kernels between trees based on string kernels between these word paths. I showed that significant performance gains over previous models were possible by applying existing string kernels in this context and devising new kernels more suitable for modeling word paths.

Another example of this line of research is building models for semantic parsing, or labeling of semantic roles of verbs such as recipient, location and time of an action [in preparation]. Solving this task makes it possible to answer basic questions about a sentence, such as "Who did what to whom?" The problem of semantic role labeling can be viewed as a multi-class classification problem for nodes in a syntactic parse tree, but since the classification decisions within a sentence are not independent, the space of possible joint labelings can be extremely complex. The challenge is to come up with models that capture important dependencies, but still generalize well and have time-space complexity that is not prohibitive. Previous work has primarily built independent classifiers for every node of the parse tree and incorporated dependence among the decisions in limited ways. The model I proposed achieves an error reduction of more than 50% by more sophisticated modeling of these dependencies. I devised dynamic programming and re-ranking algorithms to make training efficient. Other high performing systems I developed are for part-of-speech tagging [NAACL 03], alignment for machine translation [EMNLP 02], spelling correction [ACL 02] and syntactic parse disambiguation [JLAC 05]. The part-of-speech tagger is freely available and has been used in labs in several top universities.

Machine Learning for Natural Language Disambiguation

A major part of my research has been developing machine learning algorithms suitable for dealing with some of the unique problems of NL disambiguation. For example, one important problem is estimating how likely it is for two words to occur in certain relationships of interest (e.g. *verb-subject, verb-object, sequence in a sentence*), which is tantamount to learning distributions for pairs of words. Such estimates are a key component of many disambiguation systems such as syntactic parsers and speech recognizers. The problem in solving this task is the extreme sparsity of annotated data from which one could derive statistics. I developed a method for incorporating many additional sources of information such as morphology, hand-built knowledge bases specifying synonymy and "isa"-type relationships, and statistics derived from a large number of un-annotated sentences. This is achieved by learning a *random walk* model, where the states correspond to words, such that its stationary distribution is a good model for the specific word-distribution modeling task. This model combines multiple knowledge sources while learning how much to trust each of them, and chains inferences together. I showed that this model leads to large gains in a disambiguation task requiring such word distributions [ICML 04].

Another characteristic problem in NL processing is classification where the space of possible labels is complex, structured and very large. An example of this problem is syntactic parse disambiguation where the labels are the possible parse trees for a sentence. I studied empirically and theoretically classes of models applied to the task of syntactic disambiguation. I analyzed the properties of generative history-based parse tree models, which are obtained by chaining together models of small pieces of structure. I found interesting patterns in the relationship between joint likelihood, conditional likelihood and accuracy achieved by such compound models, suggesting that heavier smoothing is needed to optimize accuracy [ECML 03, best student paper runner-up]. I also compared experimentally the behavior of generative-discriminative pairs on the task of parsing for a rich unification-based grammar, varying both training set size and complexity of the hypothesis space [JLAC 05].

Future Plans

I intend to do research in the general direction of solving harder NL disambiguation tasks. I am interested in the pure machine learning aspect of the problem and also in practical applications.

On the practical side, I will center my efforts on the task of question answering, which is knowledge-rich and requires semantic processing of text. As a first step in the context of such real world application, I will focus on integration and evaluation of existing disambiguation modules as components of a working system. In particular, I believe it is useful to develop a better understanding of the value of the linguistic structures that are assigned to language units (by humans or learned classifiers) for the NL application and how much the imperfection of induced classifiers affects the performance of the end-user application. It is very important to know how useful part-of-speech tagging, syntactic parsing, semantic role labeling, and other disambiguation models are and what accuracy is needed for these components, since current and future work in the field will most likely continue to concentrate on them. As a second step, I will explore learning models for the major component that I think is missing from question answering systems – a textual entailment component. Building upon my semantic parsing work, and my unification grammar disambiguation work, I want to develop models of semantic entailment for verb and noun frames. For instance, from "John gave Mary a book" conclude "Mary received a book from John". Finally, I want to continue working on improving the component models by, for example, developing kernels for syntactic and semantic parsing or more sophisticated models for coreference resolution.

On the theoretical side, some questions I want to explore are: What are the properties of learning algorithms for nominal and highly sparse contexts? How should multiple sources of knowledge be combined – hand specified knowledge and learned probabilistic models? How should models for different components based on different corpora be combined? How should uncertainty be most effectively and efficiently propagated across component models? Lastly, my interest in probabilistic modeling extends to other domains that have common characteristics to the NL domain; for instance, I am interested in pursuing cross-disciplinary research in the bioinformatics area.