Slav Petrov - Research Statement

My research is motivated by the prospect of computer systems that can understand natural language. To develop this capability we need to automatically analyze the syntactic and semantic structure of text. This analysis is an extremely complex inferential process, which, like recognizing a face or walking, is effortless to humans. When we hear an utterance, we will be aware of only one, or at most a few sensible interpretations, as in Fig. 1. However, for a computer there will be many possible analyses. In the figure, "book" might be interpreted as a verb rather than a noun, and "read" could be a verb in different tenses, but also a noun. This pervasive ambiguity leads to combinatorially many analyses, most of which will be extremely unlikely. Manually devised rules are not sufficient to provide coverage to handle the complex structure of natural language, necessitating a system that can automatically learn from examples. To handle the flexibility of natural language, we use a statistical approach, where probabilities are assigned to the different readings of a word and the plausibility of grammatical constructions. We may learn that the probability of "book" being a verb is moderately high in general, but very small when it is preceded by "the." Similarly, we would like to learn that the two noun phrases (NP) in Fig. 1 are not interchangeable, as it is not possible to substitute the subject NP ("She") for the object NP ("the book"). We encode these phenomena in a grammar, which models a distribution over all possible interpretations of a sentence, and then search for the most probable interpretation.

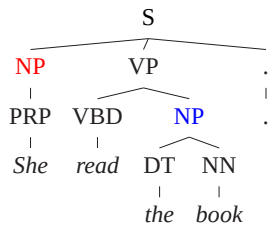

Figure 1: Syntactic parse trees model grammatical relationships.

In my dissertation, I introduce *hierarchical latent-variable grammars* for automatically learning rich linguistic structures with little or no human supervision. Starting from an extremely simple initial grammar, we use a latent variable approach to automatically learn a broad coverage grammar. Our final grammar provides an accurate and compact model, and exhibits many linguistically interpretable patterns, despite being automatically induced. This is in contrast to previous work, which often relied on manually encoding linguistic intuitions. Unfortunately, grammars that are sufficiently complex to handle the grammatical structure of natural language are challenging to work with in practice because of their size. I therefore develop an approximate *coarse-to-fine* inference procedure that gives tremendous speed-ups over exact inference. The resulting parsing system improves the state-of-the-art in both accuracy (above 90% for an array of languages), and processing speed (only 200ms per sentence), enabling the deployment of parsers in large scale natural language processing systems. I have also applied hierarchical latent-variable models and coarse-to-fine inference to a variety of other natural language processing tasks, ranging from machine translation to speech recognition.

Learning Latent-Variable Grammars

In order to enable NLP applications like machine translation, question answering, and information extraction, systems must analyze the syntactic structure of input text. The task in syntactic parsing is to learn a grammar from example parse trees like the one shown in Fig. 1, and then to use the grammar to predict the syntactic structure of previously unseen sentences. Unfortunately, the provided syntactic annotation is not sufficient for modeling the true underlying processes. For example, the annotation standard uses a single noun phrase (NP) category, but the characteristics of NPs depend highly on the context. Fig. 2 shows that NPs in subject position have a much higher probability of being a single pronoun than NPs in object position. Similarly, there is a single pronoun label (PRP), but only nominative case pronouns can be used in subject position, and ac-

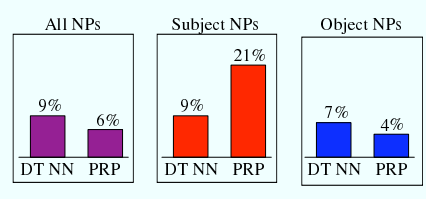

Figure 2: Distribution of the internal structure of noun phrase (NP) constructions. Subject NPs use pronouns (PRPs) more frequently.

cusative case pronouns in object position. Classical approaches have attempted to encode these linguistic phenomena by creating semantic subcategories in various ways. Unfortunately, building a highly articulated model by hand is error prone and labor intensive; it is often not even clear what the exact set of refinements ought to be.

In contrast, our latent-variable approach to grammar learning is much simpler and fully automated. We model the annotated corpus as a coarse trace of the true underlying processes. Rather than devising linguistically motivated features or splits, we use latent variables to refine each label into unconstrained subcategories. Learning proceeds in an incremental way, resulting in a hierarchy of increasingly refined grammars. We are able to automatically learn not only the subject/object distinction shown in Fig. 2, but also many other linguistic effects. Fig. 3 shows how our algorithm automatically discovers different pronoun subcategories for nominative and accusative case first, and then for sentence initial and sentence medial placement. The final grammars exhibit most of the linguistically motivated annotations of previous work, but also many additional refinements, providing a tighter statistical fit to the observed corpus. Because the model is learned directly from data and without human intervention, it is applicable to any language, and, in fact, produces state-of-the-art accuracies on all languages with appropriate data sets. In addition to English, these include related languages like German and French, but also syntactically divergent languages like Chinese and Arabic.

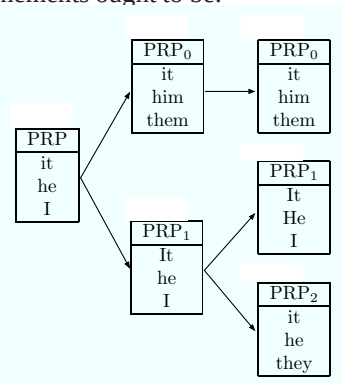

Figure 3: Incrementally learned pronoun (PRP) subcategories for grammatical cases and placement. Categories are represented by the three most likely words.

Latent-variable approaches are not limited to grammar learning. In acoustic modeling for speech recognition, one needs to learn how the acoustic characteristics of phones change depending on context. Traditionally, a decision-tree approach is used, where a series of linguistic criteria are compared. We have shown that a latent-variable approach can yield better performance while requiring no supervision. In the future, I also plan to use latent-variable models for machine translation. Herein the latent variables can be used to more accurately model how translations change depending on grammatical context. In general, I am excited to apply these techniques to other domains that require the estimation of more highly articulated models than human annotation can provide.

Hierarchical Coarse-to-Fine Inference

When working with rich probabilistic models for real-world problems, inference can be prohibitively slow. Coarse-to-fine reasoning is an idea which has enabled great advances in scale, across a wide range of problems in artificial intelligence. The general idea is simple: when a model is too complex to work with, we construct simpler approximations thereof and use those to guide the search. Despite the intuitive appeal of such methods, it was not obvious how they might be applied to natural language processing (NLP) tasks. In NLP, the search spaces are often highly structured and dynamic programming is used to compute probability distributions over the output space. We have developed a principled framework for projecting a complex (heavily refined) search space onto a hierarchy of simpler (coarser) search spaces, where the projections are obtained by hierarchical clustering of the dynamic programming states. Our empirical results show that coarse-to-fine inference outperforms other approximate inference techniques on a range of tasks, because it prunes only low probability regions of the search space and therefore makes very few search errors.

In parsing, the grammars are often large, and inference becomes too slow for practical applications. To use coarse-to-fine parsing, we create a hierarchy of coarser grammars, and then repeatedly re-parse the sentence with increasingly refined grammars. The final grammar then needs to consider only a small fraction of the possible search space as shown in Fig. 4. With coarse-to-fine inference, our parser can process a sentence in less than 200ms (compared to 60sec per sentence for exact search), without a drop in accuracy. This speed-up makes the deployment of a parser in larger natural language processing systems possible. In machine translation, the space of possible translations is very large because natural languages have many words. However, because words are atomic units, there is not an obvious way for resolving this problem. We use a hierarchical clustering scheme to induce latent structure in the search space and thereby obtain simplified languages. We then translate into a sequence of simplified versions of the target language, having only a small number of word tokens and prune away words that are unlikely to occur in the

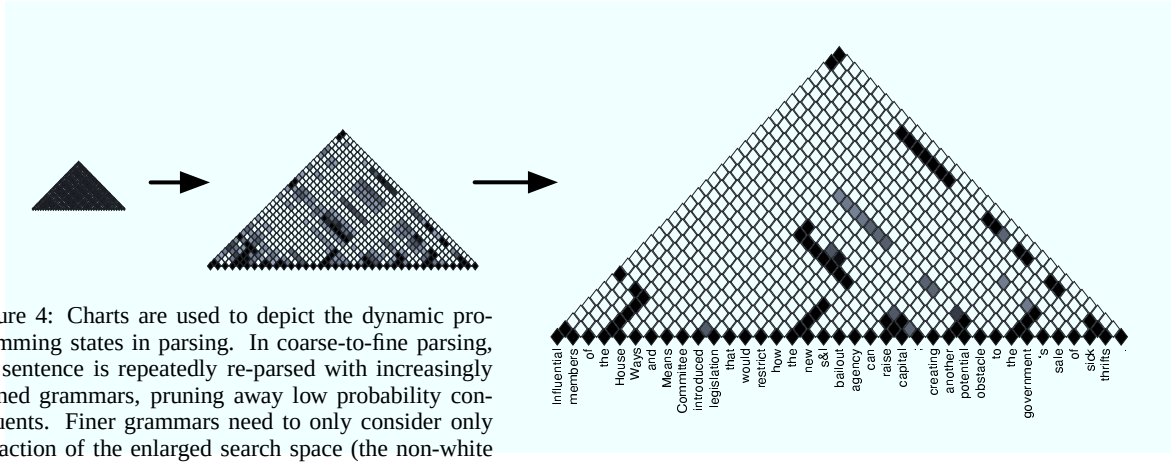

Figure 4: Charts are used to depict the dynamic programming states in parsing. In coarse-to-fine parsing, the sentence is repeatedly re-parsed with increasingly refined grammars, pruning away low probability constituents. Finer grammars need to only consider only a fraction of the enlarged search space (the non-white chart items).

translation. This results in 50-fold speed-ups at the same level of accuracy, alleviating one of the major bottlenecks in machine translation. Alternatively, one can obtain significant improvements in translation quality at the same speed.

Deep Analysis for NLP Applications

To date, NLP applications have typically avoided deep structural and semantic analysis of the input text for two reasons: computational limitations and lack of accurate models of deeper linguistic phenomena. The time is ripe for a change as computers are powerful and cheap, and parsers for many languages are now readily available. We have released a software tool for syntactic parsing so that our research results can be of direct use to other researchers in NLP and related fields. It is exciting to see that our parser has been downloaded several hundred times, and is a central component in multiple state-of-the-art translation systems, including the winner of the 2008 NIST machine translation competition. Of course, machine translation is just one example application where syntactic information has already been shown to lead to improved performance. Other downstream applications that could benefit from syntactic analysis are document understanding, information extraction, or question answering. I believe that syntactic analysis will eventually be used in most, if not all, NLP applications.

However, I also believe that we will see even more benefits when the analysis component is more closely integrated into the final application system. Rather than viewing parsing as a standalone task performed by a separate module, the analysis should be performed with a specific task in mind, as very different analyses might be required for different applications. For information extraction, where we only want to extract particular facts, analyzing the relationships between different objects might be sufficient. For machine translation, in contrast, where want to preserve the meaning as closely as possible, a very rich semantic representation might be required. A closer integration within the final application will enable deeper and more appropriate analysis that goes beyond pure syntactic structure and involves lexical semantics and meaning representation.

Besides working on natural language processing, I am interested in machine learning more generally, and also in other related fields which face similar problems and share similar techniques. In the past, I have worked on computer vision problems for soccer playing robots and also designed a video classification system, and I would be excited to collaborate with computer vision and robotics colleagues to bridge the gap between our fields. Video analysis and interactive robots are exciting domains requiring a fundamental understanding of several fields and I would be excited to contribute my NLP expertise in such a multi-disciplinary endeavor.

Throughout my research, I am particularly interested in designing elegant, streamlined models that are easy to understand and analyze, but nonetheless maximize accuracy and efficiency. My work so far has advanced the state-of-the-art in a number of NLP domains, but is just a small step towards the ultimate goal of designing systems that allow us to interact with computers in the same way that we do with humans, which in turn would enable a plethora of new opportunities.