Research Statement

Tahira Naseem

My interest in the area of Natural Language Processing (NLP) is motivated by two factors: First, progress in this area will lead towards automation of numerous tasks involving human-human and human-computer communication, ranging from question-answering to translation and text summarization. Second, the vast amount of textual data available today, combined with the ever increasing computational power, can change the foundation of linguistic research; it can open up new ways of evaluating and establishing linguistic phenomena.

Until recently, the research in the area of NLP was focused primarily on English, assuming the availability of annotated texts in order to learn language structure. Over the time, it became increasingly clear that the process of standardizing and developing text annotations is extremely expensive and time consuming. Today, we have annotated text resources for only a handful of languages. As a result, thousands of languages without such resources are beyond the reach of NLP technology. This limited ability of language processing techniques to reuse annotations cross-lingually stands in striking contrast to the unified treatment of languages in linguistics. Linguistic theories have managed to explain a vast variety of languages in terms of general linguistic phenomena. My work has focused on developing multilingual models from the perspective of this unified view. By utilizing relations between languages, these models can transfer information across a diverse set of languages. In addition to algorithmic contributions, this line of work has the potential to unveil novel connections among human languages, complementing the qualitative analysis common in comparative linguistics.

Semi-supervised Dependency Parsing

My recent research focuses on development of probabilistic parsing models that do not require syntactic annotations in the language of interest. In joint work with my collaborators, I designed and developed models that exploit alternative, readily available sources of information. In particular we explored three types of knowledge sources: 1) language universal, 2) syntactic annotation from other languages and 3) partial semantic mark-up.

Syntactic Learning with Language Universals: Despite surface differences, human languages exhibit significant similarity in fundamental aspects of linguistic structure. For instance, in all languages Verbs act as predicate in a sentence that require Nouns as their arguments (subject, object etc.). This property can be expressed as a universal dependency relation between Verbs and Nouns. This type of universals are widely recognized in linguistics but have been tangential to multilingual parsing.

In joint work with H. Chen, M. Johnson and my advisor R. Barzilay, we demonstrate that unsupervised parsing models can greatly benefit from universal linguistic rules. The modeling challenge is to treat universal rules as high level tendencies leaving enough room for learning language specific behaviors from the text. We incorporate a small set of universal syntactic rules as constraints on top of a standard unsupervised parsing model. These

constraints do not affect the model design, instead they guide the model by constraining the search space during parameter learning. Our work demonstrates that infusing universal rules into unsupervised parsing models in this manner yields significant performance gains across multiple languages (EMNLP 2010).

Multilingual Parsing via Selective Sharing: In joint work with A. Globerson and my advisor R. Barzilay, we developed a novel algorithm for multilingual dependency parsing that uses annotations from a diverse set of source languages to parse a new unannotated language (ACL 2012). In contrast to previous approaches, the algorithm can utilize annotations from languages that exhibit significant syntactic divergences from the target language.

The main point of departure is factorization of the dependency model according to linguistic theory that distinguishes between universal syntactic properties (i.e, dependent selection) and language-specific properties (i.e, dependent ordering). For instance, the fact that Nouns take Adjectives and Determiners as dependents is universal across languages. However, the order of these dependents with respect to their parent varies between languages. The ordering factor of our model is learned as a function of typological properties of a language as established by linguistic theories. This allows selective sharing of ordering information between source language are target languages.

We have demonstrated that this model delivers substantial gains over previous state-of-the-art transfer approaches. More importantly, the gains are particularly large when the target language is phylogenetically unrelated to the source languages. This property of the model broadens the benefits of multilingual learning to languages that differ substantially in syntactic structure from existing resource-rich languages.

An interesting finding of this work is that if typological features are treated as hidden variables learned by the model, most of the performance gains are still retained. This opens up the possibility of using such a model for discovering language properties and crosslingual connections.

Using Semantic Cues to Learn Syntax: Semantic annotations can provide useful clues about the underlying syntactic structure of a sentence. Moreover, partial semantic annotations are readily available in many domains (e.g., info-boxes and HTML markup), thus offering an alternative form of supervision for the languages lacking in syntactically annotated resources. In collaboration with my advisor R. Barzilay (AAAI 2011), I developed a probabilistic model that jointly explains the syntactic and semantic structures of a sentence, thereby constraining the syntactic variations based on the observed semantic annotations. Our results demonstrate that even a small amount of incomplete semantic annotations greatly improve the accuracy of learned dependencies.

Unsupervised Multilingual Learning

In my earlier work done in collaboration with B. Snyder and my advisor R. Barzilay, we focused on fully unsupervised models that induce underlying linguistic structure jointly for multiple languages from raw parallel data. Our approach is based on the insight that different languages exhibit different patterns of ambiguity - what is difficult for an automated system in one language may be a straightforward task in another language. For

example, the word "fish" in English may function as a noun or a verb, whereas many other languages, such as Hebrew, express these different uses with two distinct words. The key idea of multilingual learning is that combining natural cues from multiple languages makes the structure of each individual language more readily apparent. Our primary contribution is the development of models and algorithms that infer individual linguistic structure jointly for multiple languages. Flexible parameterization of these models encourages effective learning of cross-lingual regularities while permitting language-specific idiosyncrasies. Since the scope of shared structure differs for any given combination of languages, we employ Bayesian non-parametric models which allow the data to dictate the degree of cross-lingual correspondence.

We have applied unsupervised multilingual learning to the fundamental NLP tasks of POS tagging (EMNLP 2008b, NAACL 2009b, JAIR 2009c), and parsing (ACL 2009b). In both tasks, our multilingual learners consistently outperform their monolingual counterparts by a large margin. In fact, these methods substantially reduce the gap between unsupervised and supervised performance: for parsing by one third, and for morphology and part-of-speech tagging by one half.

Agenda for Future Research

In future, I plan to keep working on development of probabilistic models of language structure that are efficient, linguistically sound and effective for low-resource languages. In particular, I am interested in two research directions: **1)** Design and development of language processing models and applications, that can exploit available sources of linguistic knowledge to compensate for lack of direct supervision. **2)** My second research goal is to use computational models of text to help further the linguistic research on less studied languages.