Research Statement

Jinho D. Choi

My research interests are primarily in the area of natural language processing (NLP). My research goal is aimed at the practical side of NLP: how we can improve NLP components to work well in real applications, what kind of NLP components we need for diverse disciplines, or how NLP can be used to bring more innovation to other fields. Accomplishing this goal requires joint work between many research fields as well as combined efforts between multiple departments. I have developed several NLP components such as a part-of-speech tagger, a dependency parser, a semantic role labeler, etc., showing state-of-the-art performance for both accuracy and speed (Choi and Palmer, 2011a,b, 2012a; Choi, 2012; Choi and McCallum, 2013). These components are optimized for robustness across different domains and scalability on large data. All of my NLP components are implemented in an open source project,\(^1\) which has been adopted by many academic and industrial organizations. To understand how NLP is applied to real applications, I spent six months at IPsoft Inc. as a full-time researcher building a human-computer dialog system that uses various NLP techniques for question-answering, information extraction, summarization, etc. This industry experience helped me realize the gap between academic and industrial research and broadened my research interests.

My current research interests break down into two parts: improving NLP components so they can be reliably used for other research, and applying these NLP components to create more innovative intelligent systems.

Improving NLP Components

There are three ways of improving NLP components that I have approached: designing better algorithms for the NLP tasks, developing more advanced machine learning techniques, and creating more data using linguistic knowledge. For dependency parsing\(^2\) and semantic role labeling\(^3\), I designed several algorithms that significantly improved processing speed yet still showed state-of-the-art accuracy (Choi and Nicolov, 2009; Choi and Palmer, 2011b; Choi, 2012; Choi and McCallum, 2013). My latest dependency parsing algorithm using ‘selectional branching’ dynamically adjusts its search space during runtime such that it determines if a bigger search space is needed for the input to be correctly classified by measuring the confidence score of a classifier using a smaller search space (Choi and McCallum, 2013). This technique allows my parser to be faster than most other systems while maintaining state-of-the-art accuracy in multiple languages. My latest semantic role labeling algorithm using ‘higher-order argument pruning’ halves its search space by taking linguistic knowledge into account; in this way, it performs twice as fast as the previous state-of-the-art system while keeping similar accuracy for multiple genres in English (Choi, 2012, Chap. 6). The key aspect of these algorithms is approximation: I believe exhaustive search is often not necessary in NLP because of many natural linguistic constraints, and smart approximation performs faster and as accurately as exhaustive search for many tasks in NLP. As we apply these NLP components to different domains (e.g., medical data, twitter, online chat) or languages, this kind of approximation becomes more important because we can always design a more efficient algorithm by customizing it to the input data.

Most state-of-the-art NLP research use statistical machine learning approaches, and good performance gains can be achieved by adapting more advanced machine learning techniques. I have adapted several machine learning techniques to work well with NLP tasks. For part-of-speech tagging\(^4\), I introduced ‘dynamic model selection’ that allows the tagger to choose between distinct statistical models for different input data during decoding by measuring similarity between the training and the input data (Choi and Palmer, 2012a). This

---

\(^1\)ClearNLP: [http://www.clearnlp.com](http://www.clearnlp.com)

\(^2\)Dependency parsing: a task of finding dependency relations between words.

\(^3\)Semantic role labeling: a task of finding semantic representations of sentences.

\(^4\)Part-of-speech tagging: a task of finding grammatical categories of words.
technique significantly improved tagging accuracy for out-of-domain data, and performed faster and more accurately than most other state-of-the-art systems. For dependency parsing, I introduced ‘bootstrapping’ that narrows the discrepancies between features generated during training and decoding by bootstrapping dynamic features during training (Choi and Palmer, 2011a); it can be applied to any sequence classification where previously predicted labels are used as features to predict upcoming labels. This technique significantly improved parsing accuracy, and allowed my greedy parser to show comparable accuracy against other non-greedy parsers. The key aspect of these approaches is anticipation of future data. The dynamic model selection anticipates future data by training multiple models that work well for different domains, while the bootstrapping anticipates it by generating training instances that are more likely to occur during decoding. This kind of anticipation is important when our training and evaluation data are restricted to certain corpora; given the limited amount of resources we have, it is crucial to improve the robustness of these NLP components to ensure good performance for data from many different domains and languages.

For statistical learning, good performance gains can also be achieved by supplying more annotated data. I have been deeply involved with various annotation projects providing high-quality data in multiple domains and languages (Vaidya et al., 2011; Verspoor et al., 2012; Albright et al., 2013). These projects required me to learn more about syntax and semantics in Linguistics. Moreover, I developed interactive tools for annotating semantic representations, which are currently in use by several organizations (Choi et al., 2010a,b). These annotation projects helped me understand how much and what kind of data we should annotate to build high performance NLP components for new domains. I have also done extensive work on the conversion between constituent and dependency structures (Choi and Palmer, 2010a,b, 2012b). This conversion enabled me to use a large amount of data annotated with constituent structures for dependency parsing and dependency-based semantic role labeling, which improved the robustness of my dependency parser and semantic role labeler across many different domains. I am currently working on this kind of conversion in several languages such as Arabic, Chinese, and Korean, for the benefit of data-driven NLP components.

Applying NLP Components

I am very interested in the applications of NLP to other research. I am currently working on a human-computer dialog system funded by IPsoft Inc.\footnote{IPsoft webpage: \url{http://www.ipsoft.com}}, which makes heavy use of my NLP components. My goal for this research is to build a dialog-based question-answering system, which will potentially replace customer services provided by human agents. Most customer service operations come with genetic guidelines that human agents use for answering questions. Our idea is to feed these guidelines into a learning system, which will automatically find answers for customers by analyzing the information in them. I have developed a system that constructs an ontology interconnecting information between sentences from different documents by using output from many NLP components. This system also finds answers for questions by matching syntactic and semantic structures between the questions and the ontology. Moreover, the system uses a summarization technique to generate brief and informative answers. Our study has shown that this system can cover up to 80% of customer questions, saving a significant amount of human labor. This task is different from information extraction; it is easier in a sense that the domains are more restricted, but harder in a sense that the extracted information needs to be much more precise and also be generated in a human-comprehensible form.

I am also interested in emotion detection on user reviews, social network data, emails, etc., which is different from sentiment analysis detecting only positive and negative emotions; I want to detect finer-grained emotions (e.g., happy, anger, respect) given input texts that grow dynamically. This is an inter-disciplinary research involving multiple departments, at the same time, it is an under-explored task that can benefit many fields. By detecting emotion, we can advance our intelligent systems greatly. For instance, we can build a search engine finding documents with certain emotions: finding quotes that would make you happy, finding emails made your friend angry, finding posts showing respect to my work, etc. This is just a step towards building a machine reading system, which is the ultimate research goal I want to pursue.
References


