Research Statement
Tagyoung Chung

My research focuses on statistical approaches to natural language processing, especially machine translation and natural language understanding. My research interests also include computational morphology and machine learning for NLP tasks. My previous works fall into four broad categories: Improving synchronous context free grammar (SCFG)-based machine translation, using latent information to improve NLP tasks, improving optimization in natural language processing, and improving natural language understanding systems.

Contributions

My contributions in machine translation extend to all stages of machine translation pipeline. Starting from tokenization [10], alignment [8], rule extraction [1] [4] [6] [7], decoding [6], and all the way to parameter tuning [2] [3].

• SCFG-based machine translation

Although SCFG with rich representation can be a powerful modeling tool for translation, decoding SCFG suffers from many practical issues such as managing complexity. I have worked with collaborators to make SCFG-based machine translation to be more efficient. I found that enforcing SCFG rules to always have source-side lexical items by modifying rule extraction improves translation accuracy [6]. In the same publication, I discuss several practical solutions that make SCFG decoding feasible. I also found that explicitly considering lexical items when factorizing SCFG leads to better and faster search during decoding [5]. With collaborators, I have developed a novel sampling algorithm, which efficiently samples tree fragments from forests, which in turn are used as rules for translation [1]. This algorithm enables more principled extraction of rules that are more compact and accurate from aligned parallel data.

• Using latent information for NLP tasks

In syntactic theory, empty categories are covert nouns that have no surface form. I found that explicitly considering empty categories by recovering them through various methods is advantageous to the task of parsing [9], and machine translation [7]. I have also found that annotating SCFG with latent annotation learned through hierarchical clustering improves the
accuracy and the speed of decoding [4]. In languages written without segmentation or morphologically rich languages, tokenization is a nontrivial problem and the first step in the machine translation pipeline. I investigated an unsupervised learning of the source-side language tokenization from the target-side language where token sizes are explicitly controlled [10]. The method presented in the paper provides effective tokenization for languages with limited resources and for some language pairs, it outperforms supervised methods.

• Optimization for NLP tasks

I have investigated formulating the word alignment problem as an explicit optimization problem [8]. I found that the word alignment problem (IBM model 1) can be solved with traditional optimization algorithms such as LBFGS instead of EM. I also found that by formulating the word alignment problem as a matrix factorization problem, the problem can be solved with less time complexity. In addition, word fertility is learned without any additional computation. Parameter tuning for machine translation is the important last step of machine translation training pipeline. With collaborators, I found that formulating the parameter optimization as linear regression makes it converge faster and decoding with the resulting parameters yields better translation results [3]. During an internship at Microsoft Research, I worked on using direct search to tune decoder-related parameters and non-linear features in conjunction with standard linear model parameters [2]. The method can effectively tune parameters that cannot be easily tuned using common parameter tuning methods and improves translation quality.

• Improving natural language understanding systems

In the industry, I was mostly involved in improving natural language understanding systems. I worked on research and development of system that enables rapid development of data-driven natural language understanding systems. Research topics included intent clustering and labeling, using syntactic parsing and semantic role labeling to improve mention detection and classification, and adapting systems to work in different languages. Cumulative work resulted in patent applications.

Research Goal

The demand for natural language processing technology is increasing for all types of languages, including languages with limited resources. I am interested in working with such languages where linguistic differences should be carefully observed and noted to build successful systems. I am also interested in unsupervised learning since such languages unavoidably require unsupervised learning methods to solve the problem of limited resources. Above all, I am
interested in practical solutions that are applicable to real world natural language processing problems. I hope to continue contributing to natural language processing along these lines.

References


