

## Abstract

Language model plays an important role in statistical machine translation systems. It is the key knowledge source to determine the right word order of the translation. Standard  $n$ -gram based language model predicts the next word based on the  $n - 1$  immediate left context. Increasing the order of  $n$  and the size of the training data improves the performance of the LM as shown by the suffix array language model and distributed language model systems. However, such improvements narrow down very fast after  $n$  reaches 6. To improve the  $n$ -gram language model, we also developed dynamic  $n$ -gram language model adaptation and discriminative language model to tackle issues with the standard  $n$ -gram language models and observed improvements in the translation qualities.

The fact is that human beings do not reuse long  $n$ -grams to create new sentences. Rather, we reuse the structure (grammar) and replace constituents to construct new sentences. Structured language model tries to model the structural information in natural language, especially the long-distance dependencies in a probabilistic framework. However, exploring and using structural information is computationally expensive, as the number of possible structures for a sentence is very large even with the constraint of a grammar. It is difficult to apply parsers on data that is different from the training data of the treebank and parsers are usually hard to scale up.

In this thesis, we propose  $x$ -gram language model framework to model the structural information in language and apply this structured language model in statistical machine translation. The  $x$ -gram model is a highly lexicalized structural language model. It is a straight-forward extension of the  $n$ -gram language model. Trained over the dependency trees of very large corpus,  $x$ -gram model captures the structural dependencies among words in a sentence. The probability of a word given its structural context is smoothed using the well-established smoothing techniques developed for  $n$ -gram models. Because  $x$ -gram language model is simple and robust, it can be easily scaled up to larger data.

This thesis studies both semi-supervised structure and unsupervised structure. In semi-supervised induction, a parser is first trained over human labeled treebanks. This parser is then applied on a much larger and unlabeled corpus. Tree transformation is applied on the initial structure to maximize the likelihood of the tree given the initial structured language model. When the treebank is not available, as is the case for most of the languages, we propose the "dependency model 1" to induce the dependency structure from the plain text for language modeling as unsupervised learning.

The structured language model is applied in the SMT  $N$ -best list reranking and evaluated by the structured BLEU metric. Experiments show that the structured language model is a good complement to the  $n$ -gram language model and it improves the translation quality especially on the fluency aspect of the translation. This work of modeling the structural information in a statistical framework for large-scale data opens door for future research work on synchronous bilingual dependency grammar induction.