The State of the Art in Language Modeling

Joshua Goodman
Microsoft Research

Abstract

This tutorial will cover the state-of-the-art in language modeling. Language models give the probability of word sequences, i.e. "recognize speech" is much more probable than "wreck a nice beach." While most widely known for their use in speech recognition, language models are useful in a large number of areas, including information retrieval, machine translation, handwriting recognition, context-sensitive spelling correction, and text entry for Chinese and Japanese or on small input devices. Many language modeling techniques can be applied to other areas or to modeling any discrete sequence. This tutorial should be accessible to anyone with a basic knowledge of probability.

The most basic language models -- n-gram models -- essentially just count occurrences of words in training data. I will describe five relatively simple improvements over this baseline: smoothing, caching, skipping, sentence-mixture models, and clustering. I will talk a bit about the applications of language modeling and then I will quickly describe other recent promising work, and available tools and resources.

I will begin by describing conventional-style language modeling techniques.

- Smoothing addresses the problem of data sparsity: there is rarely enough data to accurately estimate the parameters of a language model. Smoothing gives a way to combine less specific, more accurate information with more specific, but noisier data. I will describe two classic techniques -- deleted interpolation and Katz (or Good-Turing) smoothing -- and one recent technique, Modified Kneser-Ney smoothing, which is the best known.
- Caching is a widely used technique that uses the observation that recently observed words are likely to occur again. Models from recently observed data can be combined with more general models to improve performance.
- Skipping models use the observation that even words that are not directly adjacent to the target word contain useful information.
- Sentence-mixture models use the observation that there are many different kinds of sentences. By modeling each sentence type separately, performance is improved.
- Clustering is one of the most useful language modeling techniques. Words can be grouped together into clusters through various automatic techniques; then the probability of a cluster can be predicted instead of the probability of the word. Clustering can be used to make smaller models or better performing ones. I will talk briefly about clustering issues specific to the huge amounts of data used in language modeling (hundreds of millions of words) to form thousands of clusters.

I will then talk about other language modeling applications, with an emphasis on information retrieval, but also mentioning spelling correction, machine translation, and entering text in Chinese or Japanese.

I will briefly describe some recent successful techniques, including Bellegarda's work using latent semantic analysis and Wang's SuperARV language models. Finally, I will also talk about some practical aspects of language modeling. I will describe how freely available, off-the-shelf tools can be used to easily build language models, where to get data to train a language model, and how to use methods such as count cutoffs or relative-entropy techniques to prune language models.

Those who attend the tutorial should walk away with a broad understanding of current language modeling techniques, and the background needed to build their own language models, and choose the right language model techniques for their applications.