

Maximum Entropy Language Modeling with Non-Local Dependencies

Jun Wu

A dissertation submitted to the Johns Hopkins University in conformity with the
requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

2002

Copyright © 2002 by Jun Wu,

All rights reserved.

Abstract

Stochastic language models are an important component in many natural language processing applications, such as automatic speech recognition and machine translation. A language model is a probability measure on word-sequences in a language. The most widely used models are N-gram models, which treat word sequences as a Markov process and predict the next word from the preceding N-1 words. For reasons of data sparseness, N is typically 2-4. N-gram models successfully “learn” local lexical dependencies but fail to capture syntactic well-formedness in sentences and semantic coherence within and across sentences.

To improve the performance of language models, two critical problems must be solved: first, deciding what kinds of long-range dependence should be used in language models, and second, determining how dependencies from different sources can be incorporated in a sound model. In this dissertation, a new language model is presented that overcomes some of the shortcomings of N-gram models by combining collocational dependencies with two sources of important long-range dependence: the syntactic structure of a sentence and the topic of a discourse. Maximum entropy techniques, which are particularly well suited for modeling diverse sources of statistical dependence, are used.

Previously known parameter estimation procedures for maximum entropy models have a computational cost that makes them impractical for large-scale applications, including the two language modeling tasks examined in this dissertation. Some fundamental algorithmic improvements in the parameter estimation procedure for maximum entropy models are presented. The computational complexity of the model parameter estimation is reduced by 2-3 orders of magnitude.

Significant improvements due to the new language model over a trigram model are demonstrated in perplexity and in word error rate for the Switchboard and the

Broadcast News tasks. Experimental results show that topic information is most helpful on content-bearing words, and syntactic structure is more useful when meaningful predicting words cannot be captured by N-grams. Experimental results also show that the topic dependence and the syntactic dependence are complementary and the gains from modeling them are nearly additive. A comparison of maximum entropy models with other models proposed in the literature is provided throughout the dissertation.