Proceedings of HLT-NAACL 2003
Short Papers , pp. 4-6
Edmonton, May-June 2003

Factored Language Models and Generalized Parallel Backoff

Jeff A. Bilmes **Katrin Kirchhoff**
SSLI-LAB, University of Washington, Dept. of Electrical Engineering
{bilmes,katrin}@ssli.ee.washington.edu

Abstract

We introduce factored language models (FLMs) and generalized parallel backoff (GPB). An FLM represents words as bundles of features (e.g., morphological classes, stems, data-driven clusters, etc.), and induces a probability model covering sequences of bundles rather than just words. GPB extends standard backoff to general conditional probability tables where variables might be heterogeneous types, where no obvious natural (temporal) backoff order exists, and where multiple dynamic backoff strategies are allowed. These methodologies were implemented during the JHU 2002 workshop as extensions to the SRI language modeling toolkit. This paper provides initial perplexity results on both CallHome Arabic and on Penn Treebank Wall Street Journal articles. Significantly, FLMs with GPB can produce bigrams with significantly lower perplexity, sometimes lower than highly-optimized baseline trigrams. In a multi-pass speech recognition context, where bigrams are used to create first-pass bigram lattices or N-best lists, these results are highly relevant.

1 Introduction

The art of statistical language modeling (LM) is to create probability models over words and sentences that tradeoff statistical prediction with parameter variance. The field is both diverse and intricate (Rosenfeld, 2000; Chen and Goodman, 1998; Jelinek, 1997; Ney et al., 1994), with many different forms of LMs including maximum-entropy, whole-sentence, adaptive and cache-based, to name a small few. Many models are simply smoothed conditional probability distributions for a word given its preceding history, typically the two preceding words.

In this work, we introduce two new methods for language modeling: *factored language model* (FLM) and *generalized parallel backoff* (GPB). An FLM considers a word as a bundle of features, and GPB is a technique that generalized backoff to arbitrary conditional probability tables. While these techniques can be considered in isolation, the two methods seem particularly suited to each other — in particular, the method of GPB can greatly facilitate the production of FLMs with better performance.

2 Factored Language Models

In a *factored language model*, a word is viewed as a vector of k factors, so that $w_t \equiv \{f_t^1, f_t^2, \ldots, f_t^K\}$. Factors can be anything, including morphological classes, stems, roots, and other such features in highly inflected languages (e.g., Arabic, German, Finnish, etc.), or data-driven word classes or semantic features useful for sparsely inflected languages (e.g., English). Clearly, a two-factor FLM generalizes standard class-based language models, where one factor is the word class and the other is words themselves. An FLM is a model over factors, i.e., $p(f_t^{1:K}|f_{t-1:t-n}^{1:K})$, that can be factored as a product of probabilities of the form $p(f|f_1, f_2, \ldots, f_N)$. Our task is twofold: 1) find an appropriate set of factors, and 2) induce an appropriate statistical model over those factors (i.e., the structure learning problem in graphical models (Bilmes, 2003; Friedman and Koller, 2001)).

3 Generalized Parallel Backoff

An individual FLM probability model can be seen as a directed graphical model over a set of $N+1$ random variables, with child variable F and N parent variables F_1 through F_N (if factors are words, then $F = W_t$ and $F_i = W_{t-i}$). Two features make an FLM distinct from a standard language model: 1) the variables $\{F, F_1, \ldots, F_N\}$ can be heterogeneous (e.g., words, word clusters, morphological classes, etc.); and 2) there is no obvious natural (e.g., temporal) backoff order as in standard word-based language models. With word-only models, backoff proceeds by dropping first the oldest word, then the next oldest, and so on until only the unigram remains. In $p(f|f_1, f_2, \ldots, f_N)$, however, many of the parent variables might be the same age. Even if the variables have differing seniorities, it is not necessarily best to drop the oldest variable first.