Overcoming the Sparseness Problem of Spoken Language Corpora using Other Large Corpora of Distinct Characteristics

Sehyeong Cho† , SangHun Kim‡ , Jun Park‡ , YoungJik Lee‡

† MyongJi University, Department of Computer Science
San 38-2 Yong In, KyungGi, Korea
shcho@mju.ac.kr

‡ Electronics and Telecommunication Research Institute,
Yusong, Daejon, Korea

Abstract. This paper proposes a method of combining two *n*-gram language models, one constructed from a very small corpus of the right domain of interest, the other constructed from a large but less adequate corpus, resulting in a significantly enhanced language model. This method is based on the observation that a small corpus from the right domain has high quality *n*-grams but has serious sparseness problem, while a large corpus from a different domain has more *n*-gram statistics but inadequately biased. Two *n*-gram models are combined by extending the idea of Katz's *backoff*. We ran experiments with 3-gram language models constructed from newspaper corpora of several million to tens of million words together with models from smaller broadcast news corpora. The target domain was broadcast news. We obtained significant improvement (30%) by incorporating a small corpus around one thirtieth size of the newspaper corpus..

1 Introduction

Language modeling is an attempt to capture the regularities and make predictions, and is an essential component of automatic speech recognition. Statistical language models are constructed from large text, or corpus. However, it is not easy to collect enough spoken language text. Other corpora are readily available, but the difference in spoken and written language leads to a poor quality.

This granted, what we need is a way of making use of existing information to help lower the perplexity of the language model. However, simply merging two corpora will not help much, as we shall see later in the next section.

2 Related Work

Linear combination is probably the simplest way of combining two language models: $P_{combined}(w \mid h) = \sum_{k=1..n} \lambda_k P_k(w \mid h)$. Linear interpolation has the advantage of extreme simplicity. It is easy to implement, easy to compute. Linear combination is consistent as far as *n*-gram models are concerned.