Interpolated Word and Class Bigram Models for Spanish Conversational Speech Recognition

Pierre Dupont

France Telecom - CNET/LAA/TSS/RCP
dupont@lannion.cnet.fr

Abstract

Introduction

This study investigates issues that arise when dealing with sparse data problems in developing language models. The practical problem under consideration is Spanish conversational speech recognition over the telephone. The primary data source, the Call Home Spanish corpus, is small in comparison with the English Switchboard corpus. With this problem in mind we explore alternative smoothing techniques.

As decribed below, one of these techniques relies on a word-to-class mapping and an associated class bigram model [3]. Future extensions of this approach may allow for learning of more complex languages models, e.g. general stochastic regular grammars, at the class level or serve as constraints for language model adaptation within the maximum entropy framework.

The Katz Model versus Absolute Discounting

The Katz model, described in [1], may be summarized as follows. Given an initial maximum likelihood estimation of an N-gram, the model is smoothed by discounting a certain probability mass from the seen events and redistributing this mass on unseen events according to their back-off, i.e. N-1 order probability estimates.

In the case of a bigram, a back-off model may be expressed as follows:

$$\hat{p}(w_i|w_{i-1}) = \begin{cases} \frac{C(w_{i-1},w_i)-d}{C(w_{i-1})} & \text{if} \quad C(w_{i-1},w_i) > 0 \\ \alpha.\frac{C(w_i)}{C^*} & \text{if} \quad C(w_{i-1},w_i) = 0 \end{cases}$$

where the $C(.)$ represent the event counts, d is the discounting parameter,

$\alpha = \sum_{w_i:C(w_{i-1},w_i)>0} \frac{d}{C(w_{i-1})}$, and $C^* = \sum_{w_i:C(w_{i-1},w_i)=0} C(w_i)$.

In the Katz model, the discounting parameter d depends on the frequency of counts and the discounting is usually not applied for high counts while, in the Absolute Discounting [2] model, this parameter is fixed for all the events independently of their actual counts.

Our experiments illustrate that, with a small amount of training data, i.e. 17,375 utterances (141,718 words) from a vocabulary of 10,551 words, the absolute discounting performs better than the Katz model. More precisely, the perplexity computed on an independent test set of 1070 utterances goes from 220.0 to 204.6.

Class Bigram Model

Given a deterministic word-to-class mapping[1] $g : w \to g(w)$, a class bigram is defined as follows :

$$p(w_i|w_{i-1}) = p(w_i|g(w_i)) \cdot p(g(w_i)|g(w_{i-1}))$$

where $p(w_i|g(w_i))$ represents the class membership probability and $p(g(w_i)|g(w_{i-1}))$ is the bigram probability at the class level. The actual estimates of these probabilities are smoothed according to the back-off model with absolute discounting.

Let k denote the number of classes. Note that two extreme cases arise when either k is equal to 1, i.e. the class model is equivalent to a word unigram model, or k is equal to the lexicon size, i.e. the class model reduces to the word bigram model.

The class estimation problem may be stated as follows. Given a fixed k and some initial word to class mapping, reestimate the mapping so as to minimize the perplexity on the training set. The value of k is subsequently chosen according to the optimal cross-validation set perplexity. We exploit the greedy algorithm for mapping reestimation described in [3].

Interpolated Models

Because of the back-off, the class bigram (referred to as Model 0) depends on the class unigram probability. However, due to the limited amount of training data, it might be interesting to make additional use of the lower order statistics by linearly interpolating both estimates. This gives rise to Model 1, in which the interpolation parameter λ_c has been estimated by cross-validation and is typically equal to 0.2. Model 2 results from a linear interpolation between the class bigram Model 1 and the word bigram previously described. Our experimental results show that the test set perplexities are 229.9, 217.9 and 185.1, respectively.

Model 0 : $p_{C_0}(w_i|w_{i-1}) = p(w_i|g(w_i)) \cdot p(g(w_i)|g(w_{i-1}))$

Model 1 : $p_{C_1}(w_i|w_{i-1}) = p(w_i|g(w_i)) \cdot [(1-\lambda_c) \cdot p(g(w_i)|g(w_{i-1})) + \lambda_c \cdot p(g(w_i))]$

Model 2 : $p_{C_2}(w_i|w_{i-1}) = [(1-\lambda) \cdot p_{C_1}(w_i|w_{i-1}) + \lambda \cdot p_W(w_i|w_{i-1})]$

Initial word error rate performance figures for Model 2 and the Katz baseline have been generated for the cross-validation test data. Additional experiments are under way on the test data.

References

[1] S. M. Katz, *Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer*, IEEE Trans. on ASSP, Vol. 35, No. 3, pp. 400-401, 1987.

[2] H. Ney and U. Essen, *Estimating 'Small' Probabilities by Leaving-One-Out*, Proc. of EUROSPEECH, pp. 2239-2242, 1993.

[3] H. Ney and R. Knesser, *Improved Clustering Techniques for Class-based Statistical Language Modelling*, Proc. of EUROSPEECH, pp. 973-976, 1993.

[1] We assume here that each word w belongs exactly to one class $g(w)$.