Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 25 (2011) 802–828

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Minimum Bayes Risk decoding and system combination based on a recursion for edit distance

Haihua Xu [a], Daniel Povey [b,*], Lidia Mangu [c], Jie Zhu [a]

[a] *Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China*
[b] *Microsoft Research, Redmond, WA, USA*
[c] *IBM T.J. Watson Research Center, Yorktown Heights, NY, USA*

Received 6 May 2010; received in revised form 27 January 2011; accepted 2 March 2011

Abstract

In this paper we describe a method that can be used for Minimum Bayes Risk (MBR) decoding for speech recognition. Our algorithm can take as input either a single lattice, or multiple lattices for system combination. It has similar functionality to the widely used Consensus method, but has a clearer theoretical basis and appears to give better results both for MBR decoding and system combination. Many different approximations have been described to solve the MBR decoding problem, which is very difficult from an optimization point of view. Our proposed method solves the problem through a novel forward–backward recursion on the lattice, not requiring time markings. We prove that our algorithm iteratively improves a bound on the Bayes risk.

Keywords: Speech recognition; Minimum Bayes Risk decoding

1. Introduction

Speech recognition systems generally make use of the maximum a posteriori (MAP) decoding rule:

$$\begin{aligned} W^* &= \mathbf{argmax}_W P(W|\mathcal{X}) \\ &= \mathbf{argmax}_W P(W)p(\mathcal{X}|W), \end{aligned} \tag{1}$$

where W is the word-sequence, $\mathcal{X}$ is the acoustic observation sequence, $P(W)$ is the language model probability and $p(\mathcal{X}|W)$ is the acoustic likelihood (ignoring likelihood scaling for now).

It can be shown that under assumptions of model correctness, Eq. (1) gives the Minimum Bayes Risk estimate with respect to the *sentence* error, i.e. it minimizes the probability of choosing the wrong sentence. It is not clear that this is the best approach; the standard error metric for speech recognition systems is the word error rate (WER), computed

[*] Corresponding author. Tel.: +1 425 7062667; fax: +1 425 7067329.
E-mail addresses: haihua_xu@sjtu.edu.cn (H. Xu), dpovey@microsoft.com (D. Povey), mangu@us.ibm.com (L. Mangu), zhujie@jtu.edu.cn (J. Zhu).

0885-2308/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2011.03.001

for sentences $n = 1 \ldots N$ as:

$$\mathrm{WER}(W_1 \ldots W_N | R_1 \ldots R_N) = \frac{\sum_{n=1}^{N} L(W_n, R_n)}{\sum_{n=1}^{N} |R_n|}, \tag{2}$$

where $L(A, B)$ is the Levenshtein edit distance (Levenshtein, 1966) between sequence A and B, W_n and R_n are the n'th transcribed sentence and reference sentence, respectively, and $|A|$ is the number of symbols in sequence A. WER is usually expressed as a percentage.

A substantial amount of work has previously been done on decoding methods that minimize a WER-like risk measure, based on a lattice of alternative outputs from a speech recognizer. These methods all fall under the general category of Minimum Bayes Risk (MBR) decoding. Note that MBR decoding is an ambiguous term because it is defined only with respect to a particular measure of risk. For speech recognition (including this paper) we generally have in mind the Levenshtein edit distance, but in the machine translation literature, N-gram counting methods related to the BLEU score (Papineni et al., 2002) are generally used. In this paper we introduce a technique for MBR decoding (w.r.t. the Levenshtein edit distance) that is simpler and has a clearer theoretical basis than the most widely used method, known as Consensus (Mangu et al., 2000). The core of it is a two-dimensional recursion that in one dimension is like a forward–backward algorithm on a lattice and in the other is like the Levenshtein edit distance recursion.

In Section 2 we introduce the concept of Minimum Bayes Risk decoding and describe previous work in this area. In Section 3 we give a more detailed overview of our approach and describe how it relates to previous work. In Section 4 we describe the Levenshtein edit distance and give an algorithm to compute it, which will motivate our lattice-based algorithm. In Section 5 we discuss lattices and introduce our notation for them. In Section 6 we describe our method for approximating the edit distance between a lattice and a word sequence. In Section 7 we explain how we optimize our hypothesis with respect to this metric. In Section 8 we describe our experimental setup, and in Section 9 we present our experiments. In Section 10 we conclude. In Appendix A we prove that our approximated edit distance is an upper bound on the true edit distance, and in Appendix B we prove that our algorithm decreases the approximated edit distance on each iteration until convergence. Appendix C describes extensions of the algorithm to handle alternative lattice formats and to compute time alignments.

2. Minimum Bayes risk decoding methods

The Bayes Risk with respect to the Levenshtein distance may be written as:

$$\mathcal{R}(W) = \sum_{W'} P(W'|\mathcal{X}) L(W, W'), \tag{3}$$

and minimizing this is equivalent to minimizing the expected Word Error Rate (given the assumption of model correctness). The general aim of Minimum Bayes Risk (MBR) decoding methods is to compute the W that minimizes (3) as exactly as possible, i.e. to compute

$$W^* = \mathbf{argmin}_W \sum_{W'} P(W'|\mathcal{X}) L(W, W') \tag{4}$$

where the values of W and W' are generally constrained to a finite set covered by an N-best sentence list or a lattice. In order to motivate the problem at this point, we will give a simple example where the output differs between this method and the MAP formula (see Fig. 1). Fig. 1(a) shows the probabilities assigned by our model to different sentences; in this example it only assigns nonzero probabilities to three different sentences. Fig. 1(b) shows the expected sentence-level error metric and word-level error metric, respectively, if we output the string given in the corresponding table row, and assuming the modeled probabilities are accurate. There is no reference sentence here; the assumption is that our model accurately models ambiguity in the data. The check mark in each column is next to the hypothesis with the lowest "expected error" for that metric; this would be the output of the corresponding decoding algorithm. The MBR estimate "A D C" has a lower expected word error, but a higher expected sentence error, than the MAP estimate "A B C" (which corresponds to minimizing the expected sentence-level error rate). In general MBR decoding will increase the sentence error rate, since it exploits the difference between string-level and word-level error rates. This should be borne in mind when evaluating the appropriateness of MBR decoding to a particular situation, as one runs the risk of "over-tuning" to a particular metric such as Word Error Rate, which may only be a proxy for an underlying risk that