Minimum Bayes Risk Decoding for BLEU

Nicola Ehling and Richard Zens and Hermann Ney

Human Language Technology and Pattern Recognition
Lehrstuhl für Informatik 6 – Computer Science Department
RWTH Aachen University, D-52056 Aachen, Germany
{ehling,zens,ney}@cs.rwth-aachen.de

Abstract

We present a Minimum Bayes Risk (MBR) decoder for statistical machine translation. The approach aims to minimize the expected loss of translation errors with regard to the BLEU score. We show that MBR decoding on N-best lists leads to an improvement of translation quality.

We report the performance of the MBR decoder on four different tasks: the TC-STAR EPPS Spanish-English task 2006, the NIST Chinese-English task 2005 and the GALE Arabic-English and Chinese-English task 2006. The absolute improvement of the BLEU score is between 0.2% for the TC-STAR task and 1.1% for the GALE Chinese-English task.

1 Introduction

In recent years, statistical machine translation (SMT) systems have achieved substantial progress regarding their perfomance in international translation tasks (TC-STAR, NIST, GALE).

Statistical approaches to machine translation were proposed at the beginning of the nineties and found widespread use in the last years. The "standard" version of the Bayes decision rule, which aims at a minimization of the sentence error rate is used in virtually all approaches to statistical machine translation. However, most translation systems are judged by their ability to minimize the error rate on the word level or n-gram level. Common error measures are the Word Error Rate (WER) and the Position Independent Word Error Rate (PER) as well as evaluation metric on the n-gram level like the BLEU and NIST score that measure precision and fluency of a given translation hypothesis.

The remaining part of this paper is structured as follows: after a short overview of related work in Sec. 2, we describe the MBR decoder in Sec. 3. We present the experimental results in Sec. 4 and conclude in Sec. 5.

2 Related Work

MBR decoder for automatic speech recognition (ASR) have been reported to yield improvement over the widely used maximum a-posteriori probability (MAP) decoder (Goel and Byrne, 2003; Mangu et al., 2000; Stolcke et al., 1997).

For MT, MBR decoding was introduced in (Kumar and Byrne, 2004). It was shown that MBR is preferable over MAP decoding for different evaluation criteria. Here, we focus on the performance of MBR decoding for the BLEU score on various translation tasks.

3 Implementation of Minimum Bayes Risk Decoding for the BLEU Score

3.1 Bayes Decision Rule

In statistical machine translation, we are given a source language sentence $f_1^J = f_1 \dots f_j \dots f_J$, which is to be translated into a target language sentence $e_1^I = e_1 \dots e_i \dots e_I$. Statistical decision theory tells us that among all possible target language sentences, we should choose the sentence which minimizes the Bayes risk:

$$\hat{e}_1^{\hat{I}} = \underset{I, e_1^I}{\operatorname{argmin}} \left\{ \sum_{I', {e'}_1^{I'}} Pr({e'}_1^{I'} | f_1^J) \cdot L(e_1^I, {e'}_1^{I'}) \right\}$$

Here, $L(\cdot, \cdot)$ denotes the loss function under consideration. In the following, we will call this decision rule the MBR rule (Kumar and Byrne, 2004).

Proceedings of the ACL 2007 Demo and Poster Sessions, pages 101–104,
Prague, June 2007. ©2007 Association for Computational Linguistics

Although it is well known that this decision rule is optimal, most SMT systems do *not* use it. The most common approach is to use the MAP decision rule. Thus, we select the hypothesis which maximizes the posterior probability $Pr(e_1^I|f_1^J)$:

$$\hat{e}_1^{\hat{I}} = \operatorname*{argmax}_{I,e_1^I} \left\{ Pr(e_1^I|f_1^J) \right\}$$

This decision rule is equivalent to the MBR criterion under a 0-1 loss function:

$$L_{0-1}(e_1^I, {e'}_1^{I'}) = \begin{cases} 1 & \text{if } e_1^I = {e'}_1^{I'} \\ 0 & \text{else} \end{cases}$$

Hence, the MAP decision rule is optimal for the sentence or string error rate. It is *not* necessarily optimal for other evaluation metrics as for example the BLEU score. One reason for the popularity of the MAP decision rule might be that, compared to the MBR rule, its computation is simpler.

3.2 Baseline System

The posterior probability $Pr(e_1^I|f_1^J)$ is modeled directly using a log-linear combination of several models (Och and Ney, 2002):

$$Pr(e_1^I|f_1^J) = \frac{\exp\left(\sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J)\right)}{\sum\limits_{I', {e'}_1^{I'}} \exp\left(\sum_{m=1}^{M} \lambda_m h_m({e'}_1^{I'}, f_1^J)\right)} \tag{1}$$

This approach is a generalization of the source-channel approach (Brown et al., 1990). It has the advantage that additional models $h(\cdot)$ can be easily integrated into the overall system.

The denominator represents a normalization factor that depends only on the source sentence f_1^J. Therefore, we can omit it in case of the MAP decision rule during the search process. Note that the denominator affects the results of the MBR decision rule and, thus, cannot be omitted in that case.

We use a state-of-the-art phrase-based translation system similar to (Matusov et al., 2006) including the following models: an n-gram language model, a phrase translation model and a word-based lexicon model. The latter two models are used for both directions: $p(f|e)$ and $p(e|f)$. Additionally, we use a word penalty, phrase penalty and a distortion penalty. The model scaling factors λ_1^M are optimized with respect to the BLEU score as described in (Och, 2003).

3.3 BLEU Score

The BLEU score (Papineni et al., 2002) measures the agreement between a hypothesis e_1^I generated by the MT system and a reference translation $\hat{e}_1^{\hat{I}}$. It is the geometric mean of n-gram precisions $\text{Prec}_n(\cdot,\cdot)$ in combination with a brevity penalty $\text{BP}(\cdot,\cdot)$ for too short translation hypotheses.

$$\begin{aligned}
\text{BLEU}(e_1^I, \hat{e}_1^{\hat{I}}) &= \text{BP}(I, \hat{I}) \cdot \prod_{n=1}^{4} \text{Prec}_n(e_1^I, \hat{e}_1^{\hat{I}})^{1/4} \\
\text{BP}(I, \hat{I}) &= \begin{cases} 1 & \text{if } \hat{I} \geq I \\ \exp\left(1 - I/\hat{I}\right) & \text{if } \hat{I} < I \end{cases} \\
\text{Prec}_n(e_1^I, \hat{e}_1^{\hat{I}}) &= \frac{\sum\limits_{w_1^n} \min\{C(w_1^n|e_1^I), C(w_1^n|\hat{e}_1^{\hat{I}})\}}{\sum\limits_{w_1^n} C(w_1^n|e_1^I)}
\end{aligned}$$

Here, $C(w_1^n|e_1^I)$ denotes the number of occurrences of an n-gram w_1^n in a sentence e_1^I. The denominator of the n-gram precisions evaluate to the number of n-grams in the hypothesis, i.e. $I - n + 1$.

As loss function for the MBR decoder, we use:

$$L[e_1^I, \hat{e}_1^{\hat{I}}] = 1 - \text{BLEU}(e_1^I, \hat{e}_1^{\hat{I}}) \, .$$

While the original BLEU score was intended to be used only for aggregate counts over a whole test set, we use the BLEU score at the sentence-level during the selection of the MBR hypotheses. Note that we will use this sentence-level BLEU score only during decoding. The translation results that we will report later are computed using the standard BLEU score.

3.4 Hypothesis Selection

We select the MBR hypothesis among the N best translation candidates of the MAP system. For each entry, we have to compute its expected BLEU score, i.e. the weighted sum over all entries in the N-best list. Therefore, finding the MBR hypothesis has a quadratic complexity in the size of the N-best list. To reduce this large work load, we stop the summation over the translation candidates as soon as the risk of the regarded hypothesis exceeds the current minimum risk, i.e. the risk of the current best hypothesis. Additionally, the hypotheses are processed according to the posterior probabilities. Thus, we can hope to find a good candidate soon. This allows for an early stopping of the computation for each of the remaining candidates.

3.5 Global Model Scaling Factor

During the translation process, the different submodels $h_m(\cdot)$ get different weights λ_m. These scaling factors are optimized with regard to a specific evaluation criteria, here: BLEU. This optimization describes the relation between the different models but does not define the absolute values for the scaling factors. Because search is performed using the maximum approximation, these absolute values are not needed during the translation process. In contrast to this, using the MBR decision rule, we perform a summation over all sentence probabilities contained in the N-best list. Therefore, we use a global scaling factor $\lambda_0 > 0$ to modify the individual scaling factors λ_m:

$$\lambda'_m = \lambda_0 \cdot \lambda_m \ \ , m = 1, ..., M.$$

For the MBR decision rule the modified scaling factors λ'_m are used instead of the original model scaling factors λ_m to compute the sentence probabilities as in Eq. 1. The global scaling factor λ_0 is tuned on the development set. Note that under the MAP decision rule any global scaling factor $\lambda_0 > 0$ yields the same result. Similar tests were reported by (Mangu et al., 2000; Goel and Byrne, 2003) for ASR.

4 Experimental Results

4.1 Corpus Statistics

We tested the MBR decoder on four translation tasks: the TC-STAR EPPS Spanish-English task of 2006, the NIST Chinese-English evaluation test set of 2005 and the GALE Arabic-English and Chinese-English evaluation test set of 2006. The TC-STAR EPPS corpus is a spoken language translation corpus containing the verbatim transcriptions of speeches of the European Parliament. The NIST Chinese-English test sets consists of news stories. The GALE project text track consists of two parts: newswire ("news") and newsgroups ("ng"). The newswire part is similar to the NIST task. The newsgroups part covers posts to electronic bulletin boards, Usenet newsgroups, discussion groups and similar forums.

The corpus statistics of the training corpora are shown in Tab. 1 to Tab. 3. To measure the translation quality, we use the BLEU score. With exception of the TC-STAR EPPS task, all scores are computed case-insensitive. As BLEU measures accuracy, higher scores are better.

Table 1: NIST Chinese-English: corpus statistics.

		Chinese	English
Train	Sentences	9 M	
	Words	232 M	250 M
	Vocabulary	238 K	412 K
NIST 02	Sentences	878	
	Words	26 431	24 352
NIST 05	Sentences	1 082	
	Words	34 908	36 027
GALE 06 news	Sentences	460	
	Words	9 979	11 493
GALE 06 ng	Sentences	461	
	Words	9 606	11 689

Table 2: TC-Star Spanish-English: corpus statistics.

		Spanish	English
Train	Sentences	1.2 M	
	Words	35 M	33 M
	Vocabulary	159 K	110 K
Dev	Sentences	1 452	
	Words	51 982	54 857
Test	Sentences	1 780	
	Words	56 515	58 295

4.2 Translation Results

The translation results for all tasks are presented in Tab. 4. For each translation task, we tested the decoder on N-best lists of size N=10 000, i.e. the 10 000 best translation candidates. Note that in some cases the list is smaller because the translation system did not produce more candidates. To analyze the improvement that can be gained through rescoring with MBR, we start from a system that has already been rescored with additional models like an n-gram language model, HMM, IBM-1 and IBM-4.

It turned out that the use of 1 000 best candidates for the MBR decoding is sufficient, and leads to exactly the same results as the use of 10 000 best lists. Similar experiences were reported by (Mangu et al., 2000; Stolcke et al., 1997) for ASR.

We observe that the improvement is larger for

Table 3: GALE Arabic-English: corpus statistics.

		Arabic	English
Train	Sentences	4 M	
	Words	125 M	124 M
	Vocabulary	421 K	337 K
news	Sentences	566	
	Words	14 160	15 320
ng	Sentences	615	
	Words	11 195	14 493

Table 4: Translation results BLEU [%] for the NIST task, GALE task and TC-STAR task (S-E: Spanish-English; C-E: Chinese-English; A-E: Arabic-English).

	TC-STAR S-E	NIST C-E		GALE A-E		GALE C-E	
decision rule	test	2002 (dev)	2005	news	ng	news	ng
MAP	52.6	32.8	31.2	23.6	12.2	14.6	9.4
MBR	52.8	33.3	31.9	24.2	13.3	15.4	10.5

Table 5: Translation examples for the GALE Arabic-English newswire task.

Reference	the saudi interior ministry announced in a report the implementation of the death penalty today, tuesday, in the area of medina (west) of a saudi citizen convicted of murdering a fellow citizen.
MAP-Hyp	saudi interior ministry in a statement to carry out the death sentence today in the area of medina (west) in saudi citizen *found guilty of killing* one of its citizens.
MBR-Hyp	*the* saudi interior ministry *announced* in a statement to carry out the death sentence today in the area of medina (west) in saudi citizen *was killed* one of its citizens.
Reference	faruq al-shar'a takes the constitutional oath of office before the syrian president
MAP-Hyp	farouk al-shara *leads sworn in by* the syrian president
MBR-Hyp	farouk al-shara *lead the constitutional oath before* the syrian president

low-scoring translations, as can be seen in the GALE task. For an ASR task, similar results were reported by (Stolcke et al., 1997).

Some translation examples for the GALE Arabic-English newswire task are shown in Tab. 5. The differences between the MAP and the MBR hypotheses are set in *italics*.

5 Conclusions

We have shown that Minimum Bayes Risk decoding on N-best lists improves the BLEU score considerably. The achieved results are promising. The improvements were consistent among several evaluation sets. Even if the improvement is sometimes small, e.g. TC-STAR, it is statistically significant: the absolute improvement of the BLEU score is between 0.2% for the TC-STAR task and 1.1% for the GALE Chinese-English task. Note, that MBR decoding is never worse than MAP decoding, and is therefore promising for SMT. It is easy to integrate and can improve even well-trained systems by tuning them for a particular evaluation criterion.

Acknowledgments

This material is partly based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-06-C-0023, and was partly funded by the European Union under the integrated project TC-STAR (Technology and Corpora for Speech to Speech Translation, IST-2002-FP6-506738, http://www.tc-star.org).

References

P. F. Brown, J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85, June.

V. Goel and W. Byrne. 2003. Minimum bayes-risk automatic speech recognition. *Pattern Recognition in Speech and Language Processing*.

S. Kumar and W. Byrne. 2004. Minimum bayes-risk decoding for statistical machine translation. In Proc. *Human Language Technology Conf. / North American Chapter of the Assoc. for Computational Linguistics Annual Meeting (HLT-NAACL)*, pages 169–176, Boston, MA, May.

L. Mangu, E. Brill, and A. Stolcke. 2000. Finding consensus in speech recognition: Word error minimization and other applications of confusion networks. *Computer, Speech and Language*, 14(4):373–400, October.

E. Matusov, R. Zens, D. Vilar, A. Mauser, M. Popović, S. Hasan, and H. Ney. 2006. The RWTH machine translation system. In Proc. *TC-Star Workshop on Speech-to-Speech Translation*, pages 31–36, Barcelona, Spain, June.

F. J. Och and H. Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In Proc. *40th Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, pages 295–302, Philadelphia, PA, July.

F. J. Och. 2003. Minimum error rate training in statistical machine translation. In Proc. *41st Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, pages 160–167, Sapporo, Japan, July.

K. Papineni, S. Roukos, T. Ward, and W. J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Proc. *40th Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, PA, July.

A. Stolcke, Y. Konig, and M. Weintraub. 1997. Explicit word error minimization in N-best list rescoring. In Proc. *European Conf. on Speech Communication and Technology*, pages 163–166, Rhodes, Greece, September.