A Systematic Comparison of Training Criteria for Statistical Machine Translation

Richard Zens and Saša Hasan and Hermann Ney

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

Abstract

We address the problem of training the free parameters of a statistical machine translation system. We show significant improvements over a state-of-the-art minimum error rate training baseline on a large Chinese-English translation task. We present novel training criteria based on maximum likelihood estimation and expected loss computation. Additionally, we compare the maximum a-posteriori decision rule and the minimum Bayes risk decision rule. We show that, not only from a theoretical point of view but also in terms of translation quality, the minimum Bayes risk decision rule is preferable.

1 Introduction

Once we specified the Bayes decision rule for statistical machine translation, we have to address three problems (Ney, 2001):

- the search problem, i.e. how to find the best translation candidate among all possible target language sentences;
- the modeling problem, i.e. how to structure the dependencies of source and target language sentences;
- the training problem, i.e. how to estimate the free parameters of the models from the training data.

Here, the main focus is on the training problem. We will compare a variety of training criteria for statistical machine translation. In particular, we are considering criteria for the log-linear parameters or model scaling factors. We will introduce new training criteria based on maximum likelihood estimation and expected loss computation. We will show that some achieve significantly better results than the standard minimum error rate training of (Och, 2003).

Additionally, we will compare two decision rules, the common maximum a-posteriori (MAP) decision rule and the minimum Bayes risk (MBR) decision rule (Kumar and Byrne, 2004). We will show that the minimum Bayes risk decision rule results in better translation quality than the maximum a-posteriori decision rule for several training criteria.

The remaining part of this paper is structured as follows: first, we will describe related work in Sec. 2. Then, we will briefly review the baseline system, Bayes decision rule for statistical machine translation and automatic evaluation metrics for machine translation in Sec. 3 and Sec. 4, respectively. The novel training criteria are described in Sec. 5 and Sec. 6. Experimental results are reported in Sec. 7 and conclusions are given in Sec. 8.

2 Related Work

The most common modeling approach in statistical machine translation is to use a log-linear combination of several sub-models (Och and Ney, 2002). In (Och and Ney, 2002), the log-linear weights were tuned to maximize the mutual information criterion (MMI). The current state-of-the-art is to optimize these parameters with respect to the final evaluation criterion; this is the so-called minimum error rate training (Och, 2003).

Minimum Bayes risk decoding for machine trans-

Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 524–532, Prague, June 2007.

lation was introduced in (Kumar and Byrne, 2004). It was shown that MBR outperforms MAP decoding for different evaluation criteria. Further experiments using MBR for Bleu were performed in (Venugopal et al., 2005; Ehling et al., 2007). Here, we will present additional evidence that MBR decoding is preferable over MAP decoding.

Tillmann and Zhang (2006) describe a perceptron style *algorithm* for training millions of features. Here, we focus on the comparison of different training *criteria*.

Shen et al. (2004) compared different algorithms for tuning the log-linear weights in a reranking framework and achieved results comparable to the standard minimum error rate training.

An annealed minimum risk approach is presented in (Smith and Eisner, 2006) which outperforms both maximum likelihood and minimum error rate training. The parameters are estimated iteratively using an annealing technique that minimizes the risk of an expected-BLEU approximation, which is similar to the one presented in this paper.

3 Baseline System

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

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

Here, $L(e_1^I, e'^{I'}_1)$ denotes the loss function under consideration. It measures the loss (or errors) of a candidate translation e_1^I assuming the correct translation is $e'^{I'}_1$. In the following, we will call this decision rule the MBR rule (Kumar and Byrne, 2004). This decision rule is optimal in the sense that any other decision rule will result (on average) in at least as many errors as the MBR rule. Despite this, most SMT systems do *not* use the MBR decision rule. The most common approach is to use the maximum a-posteriori (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 is equivalent to the MBR decision rule under a 0-1 loss function:

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

Hence, the MAP decision rule is optimal for the sentence or string error rate. It is *not* necessarily optimal for other evaluation metrics such as 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.

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):

$$p_{\lambda_1^M}(e_1^I | f_1^J) = \frac{\exp\left(\sum_{m=1}^M \lambda_m h_m(e_1^I, f_1^J)\right)}{\sum_{I', e'^{I'}_1} \exp\left(\sum_{m=1}^M \lambda_m h_m(e'^{I'}_1, f_1^J)\right)} \tag{1}$$

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 and obtain:

$$\hat{e}_1^{\hat{I}} = \operatorname*{argmax}_{I, e_1^I} \left\{ \sum_{m=1}^M \lambda_m h_m(e_1^I, f_1^J) \right\}$$

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 (Koehn, 2004; Mauser 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.

In the following, we will discuss the so-called training problem (Ney, 2001): how do we train the free parameters λ_1^M of the model? The current state-of-the-art is to use minimum error rate training (MERT) as described in (Och, 2003). The free parameters are tuned to directly optimize the evaluation criterion.

Except for the MERT, the training criteria that we will consider are additive at the sentence-level. Thus, the training problem for a development set with S sentences can be formalized as:

$$\hat{\lambda}_1^M = \underset{\lambda_1^M}{\operatorname{argmax}} \sum_{s=1}^{S} F(\lambda_1^M, (e_1^I, f_1^J)_s) \qquad (2)$$

Here, $F(\cdot,\cdot)$ denotes the training criterion that we would like to maximize and $(e_1^I, f_1^J)_s$ denotes a sentence pair in the development set. The optimization is done using the Downhill Simplex algorithm from the Numerical Recipes book (Press et al., 2002). This is a general purpose optimization procedure with the advantage that it does not require the derivative information. Before we will describe the details of the different training criteria in Sec. 5 and 6, we will discuss evaluation metrics in the following section.

4 Evaluation Metrics

The automatic evaluation of machine translation is currently an active research area. There exists a variety of different metrics, e.g., word error rate, position-independent word error rate, BLEU score (Papineni et al., 2002), NIST score (Doddington, 2002), METEOR (Banerjee and Lavie, 2005), GTM (Turian et al., 2003). Each of them has advantages and shortcomings.

A popular metric for evaluating machine translation quality is the Bleu score (Papineni et al., 2002). It has certain shortcomings for comparing different machine translation systems, especially if comparing conceptually different systems, e.g. phrase-based versus rule-based systems, as shown in (Callison-Burch et al., 2006). On the other hand, Callison-Burch concluded that the Bleu score is reliable for comparing variants of the same machine translation system. As this is exactly what we will need in our experiments and as Bleu is currently the most popular metric, we have chosen it as our primary evaluation metric. Nevertheless, most of the methods we will present can be easily adapted to other automatic evaluation metrics.

In the following, we will briefly review the computation of the Bleu score as some of the training criteria are motivated by this. The Bleu score is a combination of the geometric mean of n-gram precisions and a brevity penalty for too short translation hypotheses. The Bleu score for a translation hypothesis e_1^I and a reference translation $\hat{e}_1^{\hat{I}}$ is computed as:

$$\text{Bleu}(e_1^I, \hat{e}_1^{\hat{I}}) = \text{BP}(I, \hat{I}) \cdot \prod_{n=1}^{4} \text{Prec}_n(e_1^I, \hat{e}_1^{\hat{I}})^{1/4}$$

with

$$\text{BP}(I, \hat{I}) = \begin{cases} 1 & \text{if } I \geq \hat{I} \\ \exp(1 - I/\hat{I}) & \text{if } I < \hat{I} \end{cases}$$

$$\text{Prec}_n(e_1^I, \hat{e}_1^{\hat{I}}) = \frac{\sum\limits_{w_1^n} \min\{C(w_1^n|e_1^I), C(w_1^n|\hat{e}_1^{\hat{I}})\}}{\sum\limits_{w_1^n} C(w_1^n|e_1^I)} \qquad (3)$$

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 denominators of the n-gram precisions evaluate to the number of n-grams in the hypothesis, i.e. $I - n + 1$.

The n-gram counts for the Bleu score computation are usually collected over a whole document. For our purposes, a sentence-level computation is preferable. A problem with the sentence-level Bleu score is that the score is zero if not at least one four-gram matches. As we would like to avoid this problem, we use the smoothed sentence-level Bleu score as suggested in (Lin and Och, 2004). Thus, we increase the nominator and denominator of $\text{Prec}_n(\cdot,\cdot)$ by one for $n > 1$. Note that we will use the sentence-level Bleu score only during training. The evaluation on the development and test sets will be carried out using the standard Bleu score, i.e. at the corpus level. As the MERT baseline does not require the use of the sentence-level Bleu score, we use the standard Bleu score for training the baseline system.

In the following, we will describe several criteria for training the log-linear parameters λ_1^M of our model. For notational convenience, we assume that there is just one reference translation. Nevertheless, the methods can be easily adapted to the case of multiple references.

5 Maximum Likelihood

5.1 Sentence-Level Computation

A popular approach for training parameters is maximum likelihood estimation (MLE). Here, the goal is to maximize the joint likelihood of the parameters and the training data. For log-linear models, this results in a nice optimization criterion which is convex and has a single optimum. It is equivalent to the maximum mutual information (MMI) criterion. We obtain the following training criterion:

$$F_{ML-S}(\lambda_1^M, (e_1^I, f_1^J)) = \log p_{\lambda_1^M}(e_1^I | f_1^J)$$

A problem that we often face in practice is that the correct translation might not be among the candidates that our MT system produces. Therefore, (Och and Ney, 2002; Och, 2003) defined the translation candidate with the minimum word-error rate as pseudo reference translation. This has some bias towards minimizing the word-error rate. Here, we will use the translation candidate with the maximum Bleu score as pseudo reference to bias the system towards the Bleu score. However, as pointed out in (Och, 2003), there is no reason to believe that the resulting parameters are *optimal* with respect to translation quality measured with the Bleu score.

The goal of this sentence-level criterion is to discriminate the single correct translation against all the other "incorrect" translations. This is problematic as, even for human experts, it is very hard to define a single best translation of a sentence. Furthermore, the alternative target language sentences are not all equally bad translations. Some of them might be very close to the correct translation or even equivalent whereas other sentences may have a completely different meaning. The sentence-level MLE criterion does not distinguish these cases and is therefore a rather harsh training criterion.

5.2 N-gram Level Computation

As an alternative to the sentence-level MLE, we performed experiments with an n-gram level MLE. Here, we limit the order of the n-grams and assume conditional independence among the n-gram probabilities. We define the log-likelihood (LLH) of a target language sentence e_1^I given a source language sentence f_1^J as:

$$F_{ML-N}(\lambda_1^M, (e_1^I, f_1^J)) = \sum_{n=1}^{N} \sum_{w_1^n \in e_1^I} \log p_{\lambda_1^M}(w_1^n | f_1^J)$$

Here, we use the n-gram posterior probability $p_{\lambda_1^M}(w_1^n | f_1^J)$ as defined in (Zens and Ney, 2006). The n-gram posterior distribution is smoothed using a uniform distribution over all possible n-grams.

$$p_{\lambda_1^M}(w_1^n | f_1^J) = \alpha \cdot \frac{N_{\lambda_1^M}(w_1^n, f_1^J)}{\sum_{w'^n_1} N_{\lambda_1^M}(w'^n_1, f_1^J)} + (1-\alpha) \cdot \frac{1}{V^n}$$

Here, V denotes the vocabulary size of the target language; thus, V^n is the number of possible n-grams in the target language. We define $N_{\lambda_1^M}(w_1^n, f_1^J)$ as in (Zens and Ney, 2006):

$$N_{\lambda_1^M}(w_1^n, f_1^J) = \sum_{I, e_1^I} \sum_{i=1}^{I-n+1} p_{\lambda_1^M}(e_1^I | f_1^J) \cdot \delta(e_i^{i+n-1}, w_1^n) \quad (4)$$

The sum over the target language sentences is limited to an N-best list, i.e. the N best translation candidates according to the baseline model. In this equation, we use the Kronecker function $\delta(\cdot, \cdot)$, i.e. the term $\delta(e_i^{i+n-1}, w_1^n)$ evaluates to one if and only if the n-gram w_1^n occurs in the target sentence e_1^I starting at position i.

An advantage of the n-gram level computation of the likelihood is that we do not have to define pseudo-references as for the sentence-level MLE. We can easily compute the likelihood for the human reference translation. Furthermore, this criterion has the desirable property that it takes partial correctness into account, i.e. it is not as harsh as the sentence-level criterion.

6 Expected Bleu Score

According to statistical decision theory, one should maximize the expected gain (or equivalently minimize the expected loss). For machine translation, this means that we should optimize the expected Bleu score, or any other preferred evaluation metric.

6.1 Sentence-Level Computation

The expected Bleu score for a given source sentence f_1^J and a reference translation $\hat{e}_1^{\hat{I}}$ is defined as:

$$\mathbb{E}[\text{Bleu}|\hat{e}_1^{\hat{I}}, f_1^J] = \sum_{e_1^I} Pr(e_1^I|f_1^J) \cdot \text{Bleu}(e_1^I, \hat{e}_1^{\hat{I}})$$

Here, $Pr(e_1^I|f_1^J)$ denotes the true probability distribution over the possible translations e_1^I of the given source sentence f_1^J. As this probability distribution is unknown, we approximate it using the log-linear translation model $p_{\lambda_1^M}(e_1^I|f_1^J)$ from Eq. 1. Furthermore, the computation of the expected Bleu score involves a sum over all possible translations e_1^I. This sum is approximated using an N-best list, i.e. the N best translation hypotheses of the MT system. Thus, the training criterion for the sentence-level expected Bleu computation is:

$$F_{EB-S}(\lambda_1^M, (\hat{e}_1^{\hat{I}}, f_1^J)) = \sum_{e_1^I} p_{\lambda_1^M}(e_1^I|f_1^J) \cdot \text{Bleu}(e_1^I, \hat{e}_1^{\hat{I}})$$

An advantage of the sentence-level computation is that it is straightforward to plug in alternative evaluation metrics instead of the Bleu score. Note that the minimum error rate training (Och, 2003) uses only the target sentence with the *maximum* posterior probability whereas, here, the whole probability *distribution* is taken into account.

6.2 N-gram Level Computation

In this section, we describe a more fine grained computation of the expected Bleu score by exploiting its particular structure. Hence, this derivation is specific for the Bleu score but should be easily adaptable to other n-gram based metrics. We can rewrite the expected Bleu score as:

$$\begin{aligned}\mathbb{E}[\text{Bleu}|\hat{e}_1^{\hat{I}}, f_1^J] &= \mathbb{E}[\text{BP}|\hat{I}, f_1^J] \\ &\quad \cdot \prod_{n=1}^{4} \mathbb{E}[\text{Prec}_n|\hat{e}_1^{\hat{I}}, f_1^J]^{1/4}\end{aligned}$$

We assumed conditional independence between the brevity penalty BP and the n-gram precisions Prec_n. Note that although these independence assumptions do not hold, the resulting parameters might work well for translation. In fact, we will show that this criterion is among the best performing ones in Sec. 7. This type of independence assumption is typical within the naive Bayes classifier framework. The resulting training criterion that we will use in Eq. 2 is then:

$$\begin{aligned}F_{EB-N}(\lambda_1^M, (\hat{e}_1^{\hat{I}}, f_1^J)) &= \mathbb{E}_{\lambda_1^M}[\text{BP}|\hat{I}, f_1^J] \\ &\quad \cdot \prod_{n=1}^{4} \mathbb{E}_{\lambda_1^M}[\text{Prec}_n|\hat{e}_1^{\hat{I}}, f_1^J]^{1/4}\end{aligned}$$

We still have to define the estimators for the expected brevity penalty as well as the expected n-gram precision:

$$\mathbb{E}_{\lambda_1^M}[\text{BP}|\hat{I}, f_1^J] = \sum_{I} \text{BP}(I, \hat{I}) \cdot p_{\lambda_1^M}(I|f_1^J)$$

$$\mathbb{E}_{\lambda_1^M}[\text{Prec}_n|\hat{e}_1^{\hat{I}}, f_1^J] = \tag{5}$$

$$\frac{\sum_{w_1^n} p_{\lambda_1^M}(w_1^n|f_1^J) \sum_{c} \min\{c, C(w_1^n|\hat{e}_1^{\hat{I}})\} \cdot p_{\lambda_1^M}(c|w_1^n, f_1^J)}{\sum_{w_1^n} p_{\lambda_1^M}(w_1^n|f_1^J) \sum_{c} c \cdot p_{\lambda_1^M}(c|w_1^n, f_1^J)}$$

Here, we use the sentence length posterior probability $p_{\lambda_1^M}(I|f_1^J)$ as defined in (Zens and Ney, 2006) and the n-gram posterior probability $p_{\lambda_1^M}(w_1^n|f_1^J)$ as described in Sec. 5.2. Additionally, we predict the number of occurrences c of an n-gram. This information is necessary for the so-called clipping in the Bleu score computation, i.e. the $\min$ operator in the nominator of formulae Eq. 3 and Eq. 5. The denominator of Eq. 5 is the expected number of n-grams in the target sentence, whereas the nominator denotes the expected number of correct n-grams.

To predict the number of occurrences within a translation hypothesis, we use relative frequencies smoothed with a Poisson distribution. The mean of the Poisson distribution $\mu(w_1^n, f_1^J, \lambda_1^M)$ is chosen to be the mean of the unsmoothed distribution.

$$\begin{aligned}p_{\lambda_1^M}(c|w_1^n, f_1^J) &= \beta \cdot \frac{N_{\lambda_1^M}(c, w_1^n, f_1^J)}{N_{\lambda_1^M}(w_1^n, f_1^J)} \\ &\quad + (1-\beta) \cdot \frac{\mu(w_1^n, f_1^J, \lambda_1^M)^c \cdot e^{-c}}{c!}\end{aligned}$$

Table 1: Chinese-English TC-Star task: corpus statistics.

			Chinese	English
Train	Sentence pairs		8.3 M	
	Running words		197 M	238 M
	Vocabulary size		224 K	389 K
Dev		Sentences	1 019	2 038
		Running words	26 K	51 K
Eval	2006	Sentences	1 232	2 464
		Running words	30 K	62 K
	2007	Sentences	917	1 834
		Running words	21 K	45 K

with

$$\mu(w_1^n, f_1^J, \lambda_1^M) = \sum_c c \cdot \frac{N_{\lambda_1^M}(c, w_1^n, f_1^J)}{N_{\lambda_1^M}(w_1^n, f_1^J)}$$

Note that in case the mean $\mu(w_1^n, f_1^J, \lambda_1^M)$ is zero, we do not need the distribution $p_{\lambda_1^M}(c|w_1^n, f_1^J)$. The smoothing parameters α and β are both set to 0.9.

7 Experimental Results

7.1 Task Description

We perform translation experiments on the Chinese-English TC-Star task. This is a broadcast news speech translation task used within the European Union project TC-Star[1]. The bilingual training data consists of virtually all publicly available LDC Chinese-English corpora. The 6-gram language model was trained on the English part of the bilingual training data and additional monolingual English parts from the GigaWord corpus. We use the modified Kneser-Ney discounting as implemented in the SRILM toolkit (Stolcke, 2002).

Annual public evaluations are carried out for this task within the TC-Star project. We will report results on manual transcriptions, i.e. the so-called verbatim condition, of the official evaluation test sets of the years 2006 and 2007. There are two reference translations available for the development and test sets. The corpus statistics are shown in Table 1.

7.2 Translation Results

In Table 2, we present the translation results for different training criteria for the development set and the two blind test sets. The reported case-sensitive Bleu scores are computed using the `mteval-v11b.pl`[2] tool using two reference translations, i.e. BLEUr2n4c. Note that already the baseline system (MERT-Bleu) would have achieved the first rank in the official TC-Star evaluation 2006; the best Bleu score in that evaluation was 16.1%.

The MBR hypotheses were generated using the algorithm described in (Ehling et al., 2007) on a 10 000-best list.

On the development data, the MERT-Bleu achieves the highest Bleu score. This seems reasonable as it is the objective of this training criterion.

The maximum likelihood (MLE) criteria perform somewhat worse under MAP decoding. Interestingly, the MBR decoding can compensate this to a large extent: all criteria achieve a Bleu score of about 18.9% on the development set. The benefits of MBR decoding become even more evident on the two test sets. Here, the MAP results for the sentence-level MLE criterion are rather poor compared to the MERT-Bleu. Nevertheless, using MBR decoding results in very similar Bleu scores for most of the criteria on these two test sets. We can therefore support the claim of (Smith and Eisner, 2006) that MBR tends to have better generalization capabilities.

The n-gram level MLE criterion seems to perform better than the sentence-level MLE criterion, especially on the test sets. The reasons might be that there is no need for the use of pseudo references as described in Sec. 5 and that partial correctness is taken into account.

The best results are achieved using the expected Bleu score criteria described in Sec. 6. Here, the sentence level and n-gram level variants achieve more or less the same results. The overall improvement on the Eval'06 set is about 1.0% Bleu absolute for MAP decoding and 0.9% for MBR decoding. On the Eval'07 set, the improvements are even larger, about 1.8% Bleu absolute for MAP and 1.1% Bleu for MBR. All these improvements are statistically significant at the 99% level using a pairwise significance test[3].

Given that currently the most popular approach is to use MERT-Bleu MAP decoding, the overall im-

[1]http://www.tc-star.org

[2]http://www.nist.gov/speech/tests/mt/resources/scoring.htm

[3]The tool for computing the significance test was kindly provided by the National Research Council Canada.

Table 2: Translation results: Bleu scores [%] for the Chinese-English TC-Star task for various training criteria (MERT: minimum error rate training; MLE: maximum likelihood estimation; $\mathbb{E}$[Bleu]: expected Bleu score) and the maximum a-posteriori (MAP) as well as the minimum Bayes risk (MBR) decision rule.

			Development		Eval'06		Eval'07	
Decision Rule			MAP	MBR	MAP	MBR	MAP	MBR
Training Criterion	MERT-Bleu (baseline)		**19.5**	**19.4**	16.7	17.2	22.2	23.0
	MLE	sentence-level	17.8	18.9	14.8	17.1	18.9	22.7
		n-gram level	18.6	18.8	17.0	17.8	22.8	23.5
	$\mathbb{E}$[Bleu]	sentence-level	19.1	18.9	17.5	**18.1**	23.5	**24.1**
		n-gram level	18.6	18.8	**17.7**	17.6	**24.0**	**24.0**

provement is about 1.4% absolute for the Eval'06 set and 1.9% absolute on the Eval'07 set.

Note that the MBR decision rule almost always outperforms the MAP decision rule. In the rare cases where the MAP decision rule yields better results, the difference in terms of Bleu score are small and *not* statistically significant.

We also investigated the effect of the maximum n-gram order for the n-gram level maximum likelihood estimation (MLE). The results are shown in Figure 1. We observe an increase of the Bleu score with increasing maximum n-gram order for the development corpus. On the evaluation sets, however, the maximum is achieved if the maximum n-gram order is limited to four. This seems intuitive as the Bleu score uses n-grams up to length four. However, one should be careful here: the differences are rather small, so it might be just statistical noise.

Some translation examples from the Eval'07 test set are shown in Table 3 for different training criteria under the maximum a-posteriori decision rule.

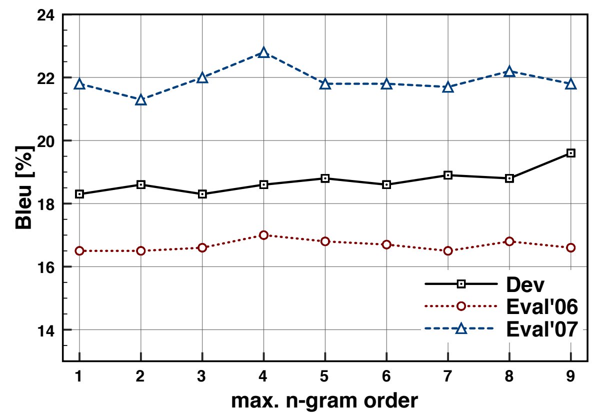

Figure 1: Effect of the maximum n-gram order on the Bleu score for the n-gram level maximum likelihood estimation under the maximum a-posteriori decision rule.

8 Conclusions

We have presented a systematic comparison of several criteria for training the log-linear parameters of a statistical machine translation system. Additionally, we have compared the maximum a-posteriori with the minimum Bayes risk decision rule.

We can conclude that the expected Bleu score is not only a theoretically sound training criterion, but also achieves the best results in terms of Bleu score. The improvement over a state-of-the-art MERT baseline is 1.3% Bleu absolute for the MAP decision rule and 1.1% Bleu absolute for the MBR decision rule for the large Chinese-English TC-Star speech translation task.

We presented two methods for computing the expected Bleu score: a sentence-level and an n-gram level approach. Both yield similar results. We think that the n-gram level computation has certain advantages: The n-gram posterior probabilities could be computed from a word graph which would result in more reliable estimates. Whether this pays off in terms of translation quality is left open for future work.

Another interesting result of our experiments is that the MBR decision rule seems to be less affected by sub-optimal parameter settings.

Although it is well-known that the MBR decision rule is more appropriate than the MAP decision rule, the latter is more popular in the SMT community (and many other areas of natural language processing). Our results show that it can be beneficial to

Table 3: Translation examples from the Eval'07 test set for different training criteria and the maximum a-posteriori decision rule. (MERT: minimum error rate training, MLE-S: sentence-level maximum likelihood estimation, $\mathbb{E}$[Bleu]: sentence-level expected Bleu)

Criterion	Translation
Reference 1	Saving Private Ryan ranks the third on the box office revenue list which is also a movie that is possible to win an 1999 Oscar award
2	Saving Private Ryan ranked third in the box office income is likely to compete in the nineteen ninety-nine Oscar Awards
MERT-Bleu	Saving private Ryan in box office income is possible ranked third in 1999 Oscar a film
MLE-S	Saving private Ryan box office revenue ranked third is possible in 1999 Oscar a film
$\mathbb{E}$[Bleu]-S	Saving private Ryan ranked third in the box office income is also likely to run for the 1999 Academy Awards a film
Reference 1	The following problem is whether people in countries like China and Japan and other countries will choose Euros rather than US dollars in international business activities in the future
2	The next question is whether China or Japan or other countries will choose to use Euros instead of US dollars when they conduct international business in the future
MERT-Bleu	The next question is in China or Japan international business activities in the future they will not use the Euro dollar
MLE-S	The next question was either in China or Japan international business activities in the future they will adopt the Euro instead of the dollar
$\mathbb{E}$[Bleu]-S	The next question was in China or Japan in the international business activities in the future they will adopt the Euro instead of the US dollar
Reference 1	The Chairman of the European Commission Jacques Santer pointed out in this September that the financial crisis that happened in Russia has not affected people's confidence in adopting the Euro
2	European Commission President Jacques Santer pointed out in September this year that Russia's financial crisis did not shake people's confidence for planning the use of the Euro
MERT-Bleu	President of the European Commission Jacques Santer on September this year that the Russian financial crisis has not shaken people 's confidence in the introduction of the Euro
MLE-S	President of the European Commission Jacques Santer September that the Russian financial crisis has not affected people 's confidence in the introduction of the Euro
$\mathbb{E}$[Bleu]-S	President of the European Commission Jacques Santer pointed out that Russia 's financial crisis last September has not shaken people 's confidence in the introduction of the Euro
Reference 1	After many years of friction between Dutch and French speaking Belgians all of them now hope to emphasize their European identities
2	After years of friction between Belgium's Dutch-speaking and French-speaking people they now all wish to emphasize their European identity
MERT-Bleu	Belgium's Dutch-speaking and French-speaking after many years of civil strife emphasized that they now hope that Europeans
MLE-S	Belgium's Dutch-speaking and francophone after years of civil strife that they now hope that Europeans
$\mathbb{E}$[Bleu]-S	Belgium's Dutch-speaking and French-speaking after many years of civil strife it is now want to emphasize their European identity

use the MBR decision rule. On the other hand, the computation of the MBR hypotheses is more time consuming. Therefore, it would be desirable to have a more efficient algorithm for computing the MBR hypotheses.

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

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proc. *Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43th Annual Meeting of the Association of Computational Linguistics (ACL)*, pages 65–72, Ann Arbor, MI, June.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85, June.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of BLEU in machine translation research. In Proc. *11th Conf. of the Europ. Chapter of the Assoc. for Computational Linguistics (EACL)*, pages 249–256, Trento, Italy, April.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In Proc. *ARPA Workshop on Human Language Technology*.

Nicola Ehling, Richard Zens, and Hermann Ney. 2007. Minimum Bayes risk decoding for BLEU. In Proc. *45th Annual Meeting of the Assoc. for Computational Linguistics (ACL): Poster Session*, Prague, Czech Republic, June.

Philipp Koehn. 2004. Pharaoh: a beam search decoder for phrase-based statistical machine translation models. In Proc. *6th Conf. of the Assoc. for Machine Translation in the Americas (AMTA)*, pages 115–124, Washington DC, September/October.

Shankar Kumar and William 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.

Chin-Yew Lin and Franz Josef Och. 2004. Orange: a method for evaluating automatic evaluation metrics for machine translation. In Proc. *COLING '04: The 20th Int. Conf. on Computational Linguistics*, pages 501–507, Geneva, Switzerland, August.

Arne Mauser, Richard Zens, Evgeny Matusov, Saša Hasan, and Hermann Ney. 2006. The RWTH statistical machine translation system for the IWSLT 2006 evaluation. In Proc. *Int. Workshop on Spoken Language Translation (IWSLT)*, pages 103–110, Kyoto, Japan, November.

Hermann Ney. 2001. Stochastic modelling: from pattern classification to language translation. In Proc. *39th Annual Meeting of the Assoc. for Computational Linguistics (ACL): Workshop on Data-Driven Machine Translation*, pages 1–5, Morristown, NJ, July.

Franz Josef Och and Hermann 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.

Franz Josef 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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing 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.

William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. 2002. *Numerical Recipes in C++*. Cambridge University Press, Cambridge, UK.

Libin Shen, Anoop Sarkar, and Franz Josef Och. 2004. Discriminative reranking for machine translation. In Proc. *Human Language Technology Conf. / North American Chapter of the Assoc. for Computational Linguistics Annual Meeting (HLT-NAACL)*, pages 177–184, Boston, MA, May.

David A. Smith and Jason Eisner. 2006. Minimum risk annealing for training log-linear models. In Proc. *21st Int. Conf. on Computational Linguistics and 44th Annual Meeting of the Assoc. for Computational Linguistics (COLING/ACL): Poster Session*, pages 787–794, Sydney, Australia, July.

Andreas Stolcke. 2002. SRILM – an extensible language modeling toolkit. In Proc. *Int. Conf. on Speech and Language Processing (ICSLP)*, volume 2, pages 901–904, Denver, CO, September.

Christoph Tillmann and Tong Zhang. 2006. A discriminative global training algorithm for statistical MT. In Proc. *21st Int. Conf. on Computational Linguistics and 44th Annual Meeting of the Assoc. for Computational Linguistics (COLING/ACL)*, pages 721–728, Sydney, Australia, July.

Joseph P. Turian, Luke Shen, and I. Dan Melamed. 2003. Evaluation of machine translation and its evaluation. Technical Report Proteus technical report 03-005, Computer Science Department, New York University.

Ashish Venugopal, Andreas Zollmann, and Alex Waibel. 2005. Training and evaluating error minimization rules for statistical machine translation. In Proc. *43rd Annual Meeting of the Assoc. for Computational Linguistics (ACL): Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*, pages 208–215, Ann Arbor, MI, June.

Richard Zens and Hermann Ney. 2006. N-gram posterior probabilities for statistical machine translation. In Proc. *Human Language Technology Conf. / North American Chapter of the Assoc. for Computational Linguistics Annual Meeting (HLT-NAACL): Proc. Workshop on Statistical Machine Translation*, pages 72–77, New York City, NY, June.