Minimum Bayes-risk System Combination

Jesús González-Rubio
Instituto Tecnológico de Informática
U. Politècnica de València
46022 Valencia, Spain
jegonzalez@iti.upv.es

Alfons Juan **Francisco Casacuberta**
D. de Sistemas Informáticos y Computación
U. Politècnica de València
46022 Valencia, Spain
{ajuan,fcn}@dsic.upv.es

Abstract

We present *minimum Bayes-risk system combination*, a method that integrates consensus decoding and system combination into a unified multi-system minimum Bayes-risk (MBR) technique. Unlike other MBR methods that re-rank translations of a single SMT system, MBR system combination uses the MBR decision rule and a linear combination of the component systems' probability distributions to search for the minimum risk translation among all the finite-length strings over the output vocabulary. We introduce expected BLEU, an approximation to the BLEU score that allows to efficiently apply MBR in these conditions. MBR system combination is a general method that is independent of specific SMT models, enabling us to combine systems with heterogeneous structure. Experiments show that our approach bring significant improvements to single-system-based MBR decoding and achieves comparable results to different state-of-the-art system combination methods.

1 Introduction

Once statistical models are trained, a decoding approach determines what translations are finally selected. Two parallel lines of research have shown consistent improvements over the max–derivation decoding objective, which selects the highest probability derivation. *Consensus decoding* procedures select translations for a single system with a minimum Bayes risk (MBR) (Kumar and Byrne, 2004). *System combination* procedures, on the other hand, generate translations from the output of multiple component systems by combining the best fragments of these outputs (Frederking and Nirenburg, 1994). In this paper, we present minimum Bayes risk system combination, a technique that unifies these two approaches by learning a consensus translation over multiple underlying component systems.

MBR system combination operates directly on the outputs of the component models. We perform an MBR decoding using a linear combination of the component models' probability distributions. Instead of re-ranking the translations provided by the component systems, we search for the hypothesis with the minimum expected translation error among all the possible finite-length strings in the target language. By using a loss function based on BLEU (Papineni et al., 2002), we avoid the hypothesis alignment problem that is central to standard system combination approaches (Rosti et al., 2007). MBR system combination assumes only that each translation model can produce expectations of n-gram counts; the latent derivation structures of the component systems can differ arbitrary. This flexibility allows us to combine a great variety of SMT systems.

The key contributions of this paper are three: the usage of a linear combination of distributions within the MBR decoding, which allows multiple SMT models to be involved in, and makes the computation of n-grams statistics to be more accurate; the decoding in an extended search space, which allows to find better hypotheses than the evidences provided by the component models; and the use of an expected BLEU score instead of the sentence-wise BLEU, which allows to efficiently apply MBR decoding in the huge search space under consideration.

We evaluate in a multi-source translation task obtaining improvements of up to $+2.0$ BLEU abs. over the best single system max-derivation, and state-of-the-art performance in the system combination task of the ACL 2010 workshop on SMT.

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pages 1268–1277,
Portland, Oregon, June 19-24, 2011. ©2011 Association for Computational Linguistics

2 Related Work

MBR system combination is a multi-system generalization of MBR decoding where the space of hypotheses is not constrained to the space of evidences. We expand the space of hypotheses following some underlying ideas of system combination techniques.

2.1 Minimum Bayes risk

In SMT, MBR decoding allows to minimize the loss of the output for a single translation system. MBR is generally implemented by re-ranking an N-best list of translations produced by a first pass decoder (Kumar and Byrne, 2004). Different techniques to widen the search space have been described (Tromble et al., 2008; DeNero et al., 2009; Kumar et al., 2009; Li et al., 2009). These works extend the traditional MBR algorithms based on N-best lists to work with lattices.

The use of MBR to combine the outputs of various MT systems has also been explored previously. Duan et al. (2010) present an MBR decoding that makes use of a mixture of different SMT systems to improve translation accuracy. Our technique differs in that we use a linear combination instead of a mixture, which avoids the problem of component systems not sharing the same search space; perform the decoding in a search space larger than the outputs of the component models; and optimize an expected BLEU score instead of the linear approximation to it described in (Tromble et al., 2008).

DeNero et al. (2010) present *model combination*, a multi-system lattice MBR decoding on the conjoined evidences spaces of the component systems. Our technique differs in that we perform the search in an extended search space not restricted to the provided evidences, have fewer parameters to learn, and optimizes an expected BLEU score instead of the linear BLEU approximation.

Another MBR-related technique to combine the outputs of various MT systems was presented by González-Rubio and Casacuberta (2010). They use different median string (Fu, 1982) algorithms to combine various machine translation systems. Our approach differs in that we take into account the posterior distribution over translations instead of considering each translation equally likely, optimize the expected BLEU score instead of a sentence-wise measure such as the edit distance or the sentence-level BLEU, and take into account the quality differences by associating a tunable scaling factor to each system.

2.2 System Combination

System combination techniques in MT take as input the outputs $\{\mathbf{e}_1, \cdots, \mathbf{e}_N\}$ of N translation systems, where $\mathbf{e}_n$ is a structured translation object (or N-best lists thereof), typically viewed as a sequence of words. The dominant approach in the field chooses a primary translation $\mathbf{e}_p$ as a backbone, then finds an alignment $\mathbf{a}_n$ to the backbone for each $\mathbf{e}_n$. A new search space is constructed from these backbone-aligned outputs and then a voting procedure of feature-based model predicts a final consensus translation (Rosti et al., 2007). MBR system combination entirely avoids this alignment problem by considering hypotheses as n-gram occurrence vectors rather than word sequences. MBR system combination performs the decoding in a larger search space and includes statistics from the components' posteriors, whereas system combination techniques typically do not.

Despite these advantages, system combination may be more appropriate in some settings. In particular, MBR system combination is designed primarily for statistical systems that generate N-best or lattice outputs. MBR system combination can integrate non-statistical systems that generate either a single or an unweighted output. However, we would not expect the same strong performance from MBR system combination in these constrained settings.

3 Minimum Bayes risk Decoding

MBR decoding aims to find the candidate hypothesis that has the least expected loss under a probability model (Bickel and Doksum, 1977). We begin with a review of MBR for SMT.

SMT can be described as a mapping of a word sequence $\mathbf{f}$ in a source language to a word sequence $\mathbf{e}$ in a target language; this mapping is produced by the MT decoder $\mathcal{D}(\mathbf{f})$. If the reference translation $\mathbf{e}$ is known, the decoder performance can be measured by the loss function $\mathcal{L}(\mathbf{e}, \mathcal{D}(\mathbf{f}))$. Given such a loss function $\mathcal{L}(\mathbf{e}, \mathbf{e}')$ between an automatic translation $\mathbf{e}'$ and a reference $\mathbf{e}$, and an underlying proba-

bility model $P(\mathbf{e}|\mathbf{f})$, MBR decoding has the following form (Goel and Byrne, 2000; Kumar and Byrne, 2004):

$$\hat{\mathbf{e}} = \arg\min_{\mathbf{e}' \in E} \mathcal{R}(\mathbf{e}') \quad (1)$$

$$= \arg\min_{\mathbf{e}' \in E} \sum_{\mathbf{e} \in E} P(\mathbf{e}|\mathbf{f}) \cdot \mathcal{L}(\mathbf{e}, \mathbf{e}') \ , \quad (2)$$

where $\mathcal{R}(\mathbf{e}')$ denotes the Bayes risk of candidate translation $\mathbf{e}'$ under loss function $\mathcal{L}$, and E represents the space of translations.

If the loss function between any two hypotheses can be bounded: $\mathcal{L}(\mathbf{e}, \mathbf{e}') \leq \mathcal{L}_{max}$, the MBR decoder can be rewritten in term of a similarity function $\mathcal{S}(\mathbf{e}, \mathbf{e}') = \mathcal{L}_{max} - \mathcal{L}(\mathbf{e}, \mathbf{e}')$. In this case, instead of minimizing the Bayes risk, we maximize the Bayes gain $\mathcal{G}(\mathbf{e}')$:

$$\hat{\mathbf{e}} = \arg\max_{\mathbf{e}' \in E} \mathcal{G}(\mathbf{e}') \quad (3)$$

$$= \arg\max_{\mathbf{e}' \in E} \sum_{\mathbf{e} \in E} P(\mathbf{e}|\mathbf{f}) \cdot \mathcal{S}(\mathbf{e}, \mathbf{e}') \ . \quad (4)$$

MBR decoding can use different spaces for hypothesis selection and gain computation ($\arg\max$ and summatory in Eq. (4)). Therefore, the MBR decoder can be more generally written as follows:

$$\hat{\mathbf{e}} = \arg\max_{\mathbf{e}' \in E_h} \sum_{\mathbf{e} \in E_e} P(\mathbf{e}|\mathbf{f}) \cdot \mathcal{S}(\mathbf{e}, \mathbf{e}') \ , \quad (5)$$

where E_h refers to the hypotheses space form where the translations are chosen and E_e refers to the evidences space that is used to compute the Bayes gain. We will investigate the expansion of the hypotheses space while keeping the evidences space as provided by the decoder.

4 MBR System Combination

MBR system combination is a multi-system generalization of MBR decoding. It uses the MBR decision rule on a linear combination of the probability distributions of the component systems. Unlike existing MBR decoding methods that re-rank translation outputs, MBR system combination search for the minimum risk hypotheses on the complete set of finite-length hypotheses over the output vocabulary. We assume the component systems to be statistically independent and define the Bayes gain as a linear combination of the Bayes gains of the components. Each system provides its own space of evidences $\mathcal{D}_n(\mathbf{f})$ and its posterior distribution over translations $P_n(\mathbf{e}|\mathbf{f})$. Given a sentence $\mathbf{f}$ in the source language, MBR system combination is written as follows:

$$\hat{\mathbf{e}} = \arg\max_{\mathbf{e}' \in E_h} \mathcal{G}(\mathbf{e}') \quad (6)$$

$$\approx \arg\max_{\mathbf{e}' \in E_h} \sum_{n=1}^{N} \alpha_n \cdot \mathcal{G}_n(\mathbf{e}') \quad (7)$$

$$= \arg\max_{\mathbf{e}' \in E_h} \sum_{n=1}^{N} \alpha_n \cdot \sum_{\mathbf{e} \in \mathcal{D}_n(\mathbf{f})} P_n(\mathbf{e}|\mathbf{f}) \cdot \mathcal{S}(\mathbf{e}, \mathbf{e}') \ , \quad (8)$$

where N is the total number of component systems, E_h represents the hypotheses space where the search is performed, $\mathcal{G}_n(\mathbf{e}')$ is the Bayes gain of hypothesis $\mathbf{e}'$ given by the n^{th} component system and α_n is a scaling factor introduced to take into account the differences in quality of the component models. It is worth mentioning that by using a linear combination instead of a mixture model, we avoid the problem of component systems not sharing the same search space (Duan et al., 2010).

MBR system combination parameters training and decoding in the extended hypotheses space are described below.

4.1 Model Training

We learn the scaling factors in Eq. (8) using minimum error rate training (MERT) (Och, 2003). MERT maximizes the translation quality of $\hat{\mathbf{e}}$ on a held-out set, according to an evaluation metric that compares to a reference set. We used BLEU, choosing the scaling factors to maximize BLEU score of the set of translations predicted by MBR system combination. We perform the maximization by means of the down-hill simplex algorithm (Nelder and Mead, 1965).

4.2 Model Decoding

In most MBR algorithms, the hypotheses space is equal to the evidences space. Following the underlying idea of system combination, we are interested in extend the hypotheses space by including new sentences created using fragments of the hypotheses in the evidences spaces of the component models. We perform the search ($argmax$ operation in Eq. (8))

Algorithm 1 MBR system combination decoding.

Require: Initial hypothesis $\mathbf{e}$
Require: Vocabulary the evidences Σ

1: $\hat{\mathbf{e}} \leftarrow \mathbf{e}$
2: **repeat**
3: $\quad \mathbf{e}_{cur} \leftarrow \hat{\mathbf{e}}$
4: $\quad$ **for** $j = 1$ **to** $|\mathbf{e}_{cur}|$ **do**
5: $\quad\quad \hat{\mathbf{e}}_s \leftarrow \mathbf{e}_{cur}$
6: $\quad\quad$ **for** $a \in \Sigma$ **do**
7: $\quad\quad\quad \mathbf{e}'_s \leftarrow Substitute(\mathbf{e}_{cur}, a, j)$
8: $\quad\quad\quad$ **if** $\mathcal{G}(\mathbf{e}'_s) > \mathcal{G}(\hat{\mathbf{e}}_s)$ **then**
9: $\quad\quad\quad\quad \hat{\mathbf{e}}_s \leftarrow \mathbf{e}'_s$
10: $\quad\quad \hat{\mathbf{e}}_d \leftarrow Delete(\mathbf{e}_{cur}, j)$
11: $\quad\quad \hat{\mathbf{e}}_i \leftarrow \mathbf{e}_{cur}$
12: $\quad\quad$ **for** $a \in \Sigma$ **do**
13: $\quad\quad\quad \mathbf{e}'_i \leftarrow Insert(\mathbf{e}_{cur}, a, j)$
14: $\quad\quad\quad$ **if** $\mathcal{G}(\mathbf{e}'_i) > \mathcal{G}(\hat{\mathbf{e}}_i)$ **then**
15: $\quad\quad\quad\quad \hat{\mathbf{e}}_i \leftarrow \mathbf{e}'_i$
16: $\quad\quad \hat{\mathbf{e}} \leftarrow \arg\max_{\mathbf{e}' \in \{\mathbf{e}_{cur}, \hat{\mathbf{e}}_s, \hat{\mathbf{e}}_d, \hat{\mathbf{e}}_i\}} \mathcal{G}(\mathbf{e}')$
17: **until** $\mathcal{G}(\hat{\mathbf{e}}) \not> \mathcal{G}(\mathbf{e}_{cur})$
18: **return** $\mathbf{e}_{cur}$

Ensure: $\mathcal{G}(\mathbf{e}_{cur}) \geq \mathcal{G}(\mathbf{e})$

using the approximate median string (AMS) algorithm (Martínez et al., 2000). AMS algorithm perform a search on a hypotheses space equal to the free monoid Σ^* of the vocabulary of the evidences $\Sigma = Voc(E_e)$.

The AMS algorithm is shown in Algorithm 1. AMS starts with an initial hypothesis $\mathbf{e}$ that is modified using edit operations until there is no improvement in the Bayes gain (Lines 3–16). On each position j of the current solution $\mathbf{e}_{cur}$, we apply all the possible single edit operations: substitution of the j^{th} word of $\mathbf{e}_{cur}$ by each word a in the vocabulary (Lines 5–9), deletion of the j^{th} word of $\mathbf{e}_{cur}$ (Line 10) and insertion of each word a in the vocabulary in the j^{th} position of $\mathbf{e}_{cur}$ (Lines 11–15). If the Bayes gain of any of the new edited hypotheses is higher than the Bayes gain of the current hypothesis (Line 17), we repeat the loop with this new hypotheses $\hat{\mathbf{e}}$, in other case, we return the current hypothesis.

AMS algorithm takes as input an initial hypothesis $\mathbf{e}$ and the combined vocabulary of the evidences spaces Σ. Its output is a possibly new hypothesis whose Bayes gain is assured to be higher or equal than the Bayes gain of the initial hypothesis.

The complexity of the main loop (lines 2-17) is $O(|\mathbf{e}_{cur}| \cdot |\Sigma| \cdot C_{\mathcal{G}})$, where $C_{\mathcal{G}}$ is the cost of computing the gain of a hypothesis, and usually only a moderate number of iterations (< 10) is needed to converge (Martínez et al., 2000).

5 Computing BLEU-based Gain

We are interested in performing MBR system combination under BLEU. BLEU behaves as a score function: its value ranges between 0 and 1 and a larger value reflects a higher similarity. Therefore, we rewrite the gain function $\mathcal{G}(\cdot)$ using single evidence (or reference) BLEU (Papineni et al., 2002) as the similarity function:

$$\mathcal{G}_n(\mathbf{e}') = \sum_{\mathbf{e} \in \mathcal{D}_n(\mathbf{f})} P_n(\mathbf{e}|\mathbf{f}) \cdot \text{BLEU}(\mathbf{e}, \mathbf{e}') \quad (9)$$

$$\text{BLEU} = \prod_{k=1}^{4} \left(\frac{m_k}{c_k} \right)^{\frac{1}{4}} \cdot \min\left(e^{1-\frac{r}{c}}, 1.0 \right) , \quad (10)$$

where r is the length of the evidence, c the length of the hypothesis, m_k the number of n-gram matches of size k, and c_k the count of n-grams of size k in the hypothesis.

The evidences space $\mathcal{D}_n(\mathbf{f})$ may contain a huge number of hypotheses[1] which often make impractical to compute Eq. (9) directly. To avoid this problem, Tromble et al. (2008) propose *linear BLEU*, an approximation to the BLEU score to efficiently perform MBR decoding when the search space is represented with lattices. However, our hypotheses space is the full set of finite-length strings in the target vocabulary and can not be represented in a lattice.

In Eq. (9), we have one hypothesis $\mathbf{e}'$ that is to be compared to a set of evidences $\mathbf{e} \in \mathcal{D}_n(\mathbf{f})$ which follow a probability distribution $P_n(\mathbf{e}|\mathbf{f})$. Instead of computing the expected BLEU score by calculating the BLEU score with respect to each of the evidences, our approach will be to use the expected n-gram counts and sentence length of the evidences to compute a single-reference BLEU score. We replace the reference statistics (r and m_n in Eq. (10)) by the expected statistics (r' and m'_n) given the pos-

[1]For example, in a lattice the number of hypotheses may be exponential in the size of its state set.

terior distribution $P_n(\mathbf{e}|\mathbf{f})$ over the evidences:

$$\mathcal{G}_n(\mathbf{e}') = \prod_{k=1}^{4} \left(\frac{m'_k}{c_k} \right)^{\frac{1}{4}} \cdot \min\left(e^{1-\frac{r'}{c}}, 1.0 \right) \quad (11)$$

$$r' = \sum_{\mathbf{e} \in \mathcal{D}_n(\mathbf{f})} |\mathbf{e}| \cdot P_n(\mathbf{e}|\mathbf{f}) \quad (12)$$

$$m'_k = \sum_{ng \in \mathcal{N}_k(\mathbf{e}')} \min(C_{\mathbf{e}'}(ng), C'(ng)) \quad (13)$$

$$C'(ng) = \sum_{\mathbf{e} \in \mathcal{D}_n(\mathbf{f})} C_{\mathbf{e}}(ng) \cdot P_n(\mathbf{e}|\mathbf{f}) \,, \quad (14)$$

where $\mathcal{N}_k(\mathbf{e}')$ is the set of n-grams of size k in the hypothesis, $C_{\mathbf{e}'}(ng)$ is the count of the n-gram ng in the hypothesis and $C'(ng)$ is the expected count of ng in the evidences. To compute the n-gram matchings m'_k, the count of each n-gram is truncated, if necessary, to not exceed the expected count for that n-gram in the evidences.

We have replaced a summation over a possibly exponential number of items ($\mathbf{e}' \in \mathcal{D}_n(\mathbf{f})$ in Eq. (9)) with a summation over a polynomial number of n-grams that occur in the evidences[2]. Both, the expected length of the evidences r' and their expected n-gram counts m'_k can be pre-computed efficiently from N-best lists and translation lattices (Kumar et al., 2009; DeNero et al., 2010).

6 Experiments

We report results on a multi-source translation task. From the Europarl corpus released for the ACL 2006 workshop on MT (WMT2006), we select those sentence pairs from the German–English (de–en), Spanish–English (es–en) and French–English (fr–en) sub-corpora that share the same English translation. We obtain a multi-source corpus with German, Spanish and French as source languages and English as target language. All the experiments were carried out with the lowercased and tokenized version of this corpus.

We report results using BLEU (Papineni et al., 2002) and translation edit rate (Snover et al., 2006) (TER). We measure statistical significance using 95% confidence intervals computed using paired bootstrap re-sampling (Zhang and Vogel, 2004). In all table cells (except for Table 3) systems without statistically significant differences are marked with the same superscript.

System		dev		test	
		BLEU	TER	BLEU	TER
de→en	MAX	25.3	60.5	25.6*	60.3
	MBR	25.1	60.7	25.4*	60.5
es→en	MAX	30.9*	53.3*	30.4*	53.9*
	MBR	31.0*	53.4*	30.4*	54.0*
fr→en	MAX	30.7*	53.9*	30.8*	53.4*
	MBR	30.7*	53.8*	30.9*	53.4*

Table 1: Performance of base systems.

Approach	dev		test	
	BLEU	TER	BLEU	TER
Best MAX	30.9*	53.3*	30.8*	53.4*
Best MBR	31.0*	53.4*	30.9*	53.4*
MBR-SC	**32.3**	**52.5**	**32.8**	**52.3**

Table 2: Performance from best single system max-derivation decoding (*Best MAX)*, the best single system minimum Bayes risk decoding (*Best MBR*) and minimum Bayes risk system combination (*MBR-SC*) combining three systems.

6.1 Base Systems

We combine outputs from three systems, each one translating from one source language (German, Spanish or French) into English. Each individual system is a phrase-based system trained using the Moses toolkit (Koehn et al., 2007). The parameters of the systems were tuned using MERT (Och, 2003) to optimize BLEU on the development set. Each base system yields state-of-the-art performance, summarized in Table 1. For each system, we report the performance of max-derivation decoding (MAX) and 1000-best[3] MBR decoding (Kumar and Byrne, 2004).

6.2 Experimental Results

Table 2 compares MBR system combination (MBR-SC) to the best MAX and MBR systems. Both Best

[2] If $\mathcal{D}_n(\mathbf{f})$ is represented by a lattice, the number of n-grams is polynomial in the number of edges in the lattice.

[3] Ehling et al. (2007) studied up to 10000-best and show that the use of 1000-best candidates is sufficient for MBR decoding.

Setup	BLEU	TER
Best MBR	30.9	53.4
MBR-SC Expected	30.9	53.5
MBR-SC E/Conjoin	32.4	52.1
MBR-SC E/C/evidences-best	30.9	53.5
MBR-SC E/C/hypotheses-best	31.8	52.5
MBR-SC E/C/Extended	32.7	52.3
MBR-SC E/C/Ex/MERT	**32.8**	**52.3**

Table 3: Results on the test set for different setups of minimum Bayes risk system combination.

MBR and MBR-SC were computed on 1000-best lists. MBR-SC uses expected BLEU as gain function using the conjoined evidences spaces of the three systems to compute expected BLEU statistics. It performs the search in the free monoid of the output vocabulary, and its model parameters were tuned using MERT on the development set. This is the standard setup for MBR system combination, and we refer to it as MBR-SC-E/C/Ex/MERT in Table 3.

MBR system combination improves single Best MAX system by $+2.0$ BLEU points in test, and always improves over MBR. This improvement could arise due to multiple reasons: the expected BLEU gain, the larger evidences space, the extended hypotheses space, or the MERT tuned scaling factor values. Table 3 teases apart these contributions.

We first apply MBR-SC to the best system (MBR-SC-Expected). Best MBR and MBR-SC-Expected differ only in the gain function: MBR uses sentence level BLEU while MBR-SC-Expected uses the expected BLEU gain described in Section 5. MBR-SC-Expected performance is comparable to MBR decoding on the 1000-best list from the single best system. The expected BLEU approximation performs as well as sentence-level BLEU and additionally requires less total computation.

We now extend the evidences space to the conjoined 1000-best lists (MBR-SC-E/Conjoin). MBR-SC-E/Conjoin is much better than the best MBR on a single system. This implies that either the expected BLEU statistics computed in the conjoined evidences space are stronger or the larger conjoined evidences spaces introduce better hypotheses.

When we restrict the BLEU statistics to be computed from only the best system's evidences space (MBR-SC-E/C/evidences-best), BLEU scores dramatically decrease relative to MBR-SC-E/Conjoin. This implies that the expected BLEU statistics computed over the conjoined 1000-best lists are stronger than the corresponding statistics from the single best system. On the other hand, if we restrict the search space to only the 1000-best list of the best system (MBR-SC-E/C/hypotheses-best), BLEU scores also decrease relative to MBR-SC-E/Conjoin. This implies that the conjoined search space also contains better hypotheses than the single best system's search space.

These results validate our approach. The linear combination of the probability distributions in the conjoined evidences spaces allows to compute much stronger statistics for the expected BLEU gain and also contains some better hypotheses than the single best system's search space does.

We next expand the conjoined evidences spaces using the decoding algorithm described in Section 4.2 (MBR-SC-E/C/Extended). In this case, the expected BLEU statistics are computed from the conjoined 1000-best lists of the three systems, but the hypotheses space where we perform the decoding is expanded to the set of all possible finite-length hypotheses over the vocabulary of the evidences. We take the output of MBR-SC-E/Conjoin as the initial hypotheses of the decoding (see Algorithm 1). MBR-SC-E/C/Extended improves BLEU score of MBR-SC-E/Conjoin but obtains a slightly worse TER score. Since these two systems are identical in their expected BLEU statistics, the improvements in BLEU imply that the extended search space has introduced better hypotheses. The degradation in TER performance can be explained by the use of a BLEU-based gain function in the decoding process.

We finally compute the optimum values for the scaling factors of the different system using MERT (MBR-SC-E/C/Ex/MERT). MBR-SC-E/C/Ex/MERT slightly improves BLEU score of MBR-SC-E/C/Extended. This implies that the optimal values of the scaling factors do not deviate much from 1.0; a similar result was reported in (Och and Ney, 2001). We hypothesize that this is because the three component systems share the same SMT model, pre-process and decoding. We expect to obtain larger improvements when combining systems implementing different MT paradigms.

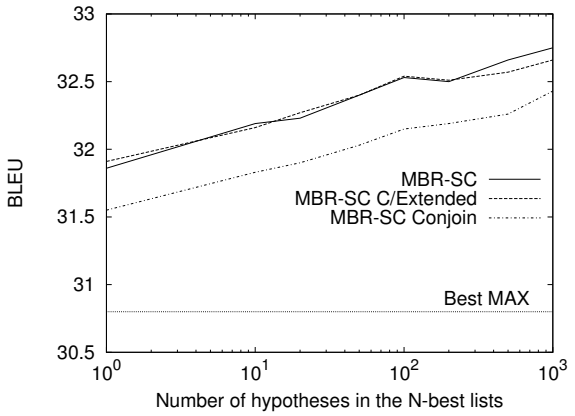

Figure 1: Performance of minimum Bayes risk system combination (MBR-SC) for different sizes of the evidences space in comparison to other MBR-SC setups.

MBR-SC-E/C/Ex/MERT is the standard setup for MBR system combination and, from now, on we will refer to it as MBR-SC.

We next evaluate performance of MBR system combination on N-best lists of increasing sizes, and compare it to MBR-SC-E/C/Extended and MBR-SC-E/Conjoin in the same N-best lists. We list the results of the Best MAX system for comparison.

Results in Figure 1 confirm the conclusions extracted from results displayed in Table 3. MBR-SC-Conjoin is consistently better than the Best MAX system, and differences in BLEU increase with the size of the evidences space. This implies that the linear combination of posterior probabilities allow to compute stronger statistics for the expected BLEU gain, and, in addition, the larger the evidences space is, the stronger the computed statistics are. MBR-SC-C/Extended is also consistently better than MBR-SC-Conjoin with an almost constant improvement of $+0.4$ BLEU points. This result show that the extended search space always contains better hypotheses than the conjoined evidences spaces; also confirms the soundness of Algorithm 1 that allows to reach them. Finally, MBR-SC also slightly improves MBR-SC-C/Extended. The optimization of the scaling factors allows only small improvements in BLEU.

Figure 2 display the MBR system combination translation and compare it to the max-derivation translations of the three component systems. Reference translation is also listed for comparison. MBR-SC adds word "*point*" to create a new translation equal to the reference. MBR-SC is able to detect that this is valuable word even though it does not appear in the max-derivation hypotheses.

MAX de→en	i will return later .
MAX es→en	i shall come back to that later .
MAX fr→en	i will return to this later .
MBR-SC	i will return to this point later .
Reference	i will return to this point later .

Figure 2: MBR system combination example.

6.3 Comparison to System Combination

Figure 3 compares MBR system combination (MBR-SC) with state-of-the-art system combination techniques presented to the system combination task of the ACL 2010 workshop on MT (WMT2010). All system combination techniques build a "word sausage" from the outputs of the different component systems and choose a path trough the sausage with the highest score under different models. A description of these systems can be found in (Callison-Burch et al., 2010).

In this task, the output of the component systems are single hypotheses or unweighted lists thereof. Therefore, we lack of the statistics of the components' posteriors which is one of the main advantages of MBR system combination over system combination techniques. However, we find that, even in these constrained setting, MBR system combination performance is similar to the best system combination techniques for all translation directions. These experiments validate our approach. MBR system combination yields state-of-the-art performance while avoiding the challenge of aligning translation hypotheses.

7 Conclusion

MBR system combination integrates consensus decoding and system combination into a unified multi-system MBR technique. MBR system combination uses the MBR decision rule on a linear combination of the component systems' probability distributions to search for the sentence with the minimum Bayes risk on the complete set of finite-length

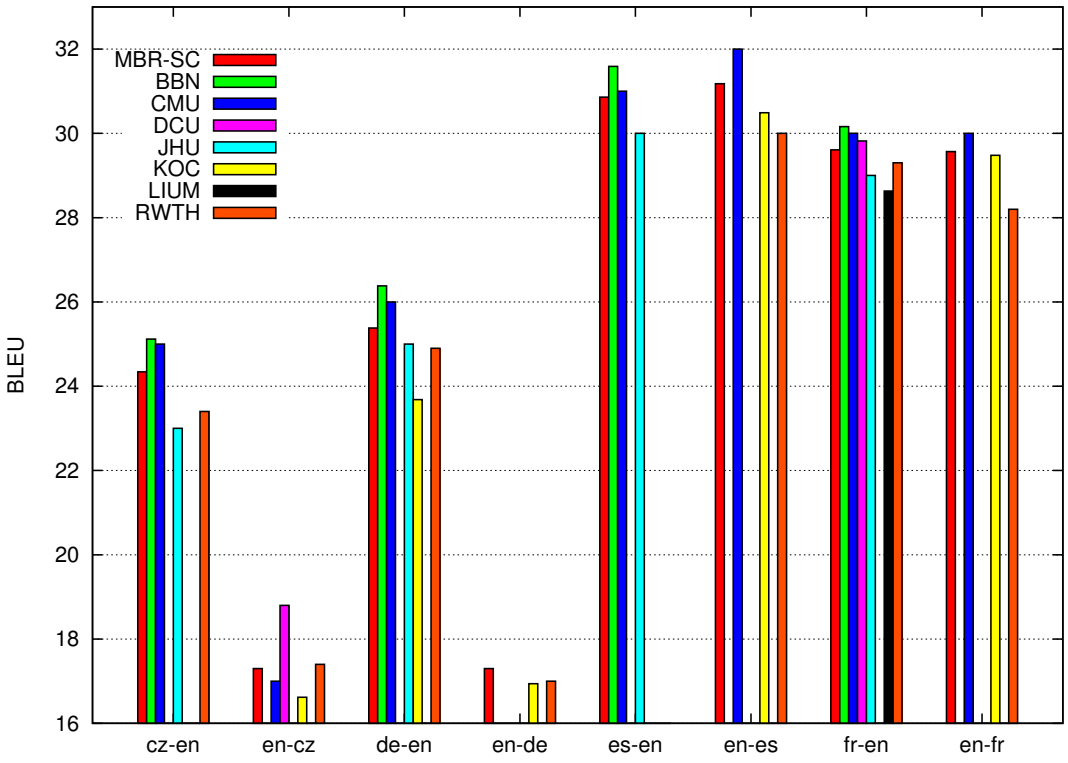

Figure 3: Performance of minimum Bayes risk system combination (MBR-SC) for different language directions in comparison to the rest of system combination techniques presented in the WMT2010 system combination task.

strings in the output vocabulary. Component systems can have varied decoding strategies; we only require that each system produce an N-best list (or a lattice) of translations. This flexibility allows the technique to be applied quite broadly. For instance, Leusch et al. (2010) generate intermediate translations in several pivot languages, translate them separately into the target language, and generate a consensus translation out of these using a system combination technique. Likewise, these pivot translations could be combined via MBR system combination.

MBR system combination has two significant advantages over current approaches to system combination. First, it does not rely on hypothesis alignment between outputs of individual systems. Aligning translation hypotheses can be challenging and has a substantial effect on combination performance (He et al., 2008). Instead of aligning the sentences, we view the sentences as vectors of n-gram counts and compute the expected statistics of the BLEU score to compute the Bayes gain. Second, we do not need to pick a backbone system for combination. Choosing a backbone system can also be challenging and also affects system combination performance (He and Toutanova, 2009). MBR system combination sidesteps this issue by working directly on the conjoined evidences space produced by the outputs of the component systems, and allows the consensus model to express system preferences via scaling factors.

Despite its simplicity, MBR system combination provides strong performance by leveraging different consensus, decoding and training techniques. It outperforms best MAX or MBR derivation on each of the component systems. In addition, it obtains state-of-the-art performance in a constrained setting better suited for dominant system combination techniques.

Acknowledgements

Work supported by the EC (FEDER/FSE) and the Spanish MEC/MICINN under the MIPRCV "Consolider Ingenio 2010" program (CSD2007-00018), the iTrans2 (TIN2009-14511) project, the UPV

under grant 20091027 and the FPU scholarship AP2006-00691. Also supported by the Spanish MITyC under the erudito.com (TSI-020110-2009-439) project and by the Generalitat Valenciana under grant Prometeo/2009/014.

References

Peter J. Bickel and Kjell A Doksum. 1977. *Mathematical statistics : basic ideas and selected topics*. Holden-Day, San Francisco.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar F. Zaidan. 2010. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 17–53, Morristown, NJ, USA. Association for Computational Linguistics.

John DeNero, David Chiang, and Kevin Knight. 2009. Fast consensus decoding over translation forests. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, pages 567–575, Morristown, NJ, USA. Association for Computational Linguistics.

John DeNero, Shankar Kumar, Ciprian Chelba, and Franz Och. 2010. Model combination for machine translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 975–983, Morristown, NJ, USA. Association for Computational Linguistics.

Nan Duan, Mu Li, Dongdong Zhang, and Ming Zhou. 2010. Mixture model-based minimum bayes risk decoding using multiple machine translation systems. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 313–321, Beijing, China, August. Coling 2010 Organizing Committee.

Nicola Ehling, Richard Zens, and Hermann Ney. 2007. Minimum bayes risk decoding for bleu. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 101–104, Morristown, NJ, USA. Association for Computational Linguistics.

Robert Frederking and Sergei Nirenburg. 1994. Three heads are better than one. In *Proceedings of the fourth conference on Applied natural language processing*, pages 95–100, Morristown, NJ, USA. Association for Computational Linguistics.

K.S. Fu. 1982. *Syntactic Pattern Recognition and Applications*. Prentice Hall.

Vaibhava Goel and William J. Byrne. 2000. Minimum bayes-risk automatic speech recognition. *Computer Speech & Language*, 14(2):115–135.

Jesús González-Rubio and Francisco Casacuberta. 2010. On the use of median string for multi-source translation. In *In Proceedings of the International Conference on Pattern Recognition (ICPR2010)*, pages 4328–4331.

Xiaodong He and Kristina Toutanova. 2009. Joint optimization for machine translation system combination. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3*, pages 1202–1211, Morristown, NJ, USA. Association for Computational Linguistics.

Xiaodong He, Mei Yang, Jianfeng Gao, Patrick Nguyen, and Robert Moore. 2008. Indirect-hmm-based hypothesis alignment for combining outputs from machine translation systems. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 98–107, Morristown, NJ, USA. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180, Morristown, NJ, USA. Association for Computational Linguistics.

Shankar Kumar and William J. Byrne. 2004. Minimum bayes-risk decoding for statistical machine translation. In *HLT-NAACL*, pages 169–176.

Shankar Kumar, Wolfgang Macherey, Chris Dyer, and Franz Och. 2009. Efficient minimum error rate training and minimum bayes-risk decoding for translation hypergraphs and lattices. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, pages 163–171, Morristown, NJ, USA. Association for Computational Linguistics.

Gregor Leusch, Aurélien Max, Josep Maria Crego, and Hermann Ney. 2010. Multi-pivot translation by system combination. In *International Workshop on Spoken Language Translation*, Paris, France, December.

Zhifei Li, Jason Eisner, and Sanjeev Khudanpur. 2009. Variational decoding for statistical machine translation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Process-*

ing of the AFNLP: Volume 2 - Volume 2, pages 593–601, Morristown, NJ, USA. Association for Computational Linguistics.

C. D. Martínez, A. Juan, and F. Casacuberta. 2000. Use of Median String for Classification. In *Proceedings of the 15th International Conference on Pattern Recognition*, volume 2, pages 907–910, Barcelona (Spain), September.

John A. Nelder and Roger Mead. 1965. A Simplex Method for Function Minimization. *The Computer Journal*, 7(4):308–313, January.

Franz Josef Och and Hermann Ney. 2001. Statistical multi-source translation. In *In Machine Translation Summit*, pages 253–258.

Franz J. Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, pages 160–167, Morristown, NJ, USA. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Morristown, NJ, USA. Association for Computational Linguistics.

Antti-Veikko Rosti, Necip Fazil Ayan, Bing Xiang, Spyros Matsoukas, Richard Schwartz, and Bonnie Dorr. 2007. Combining outputs from multiple machine translation systems. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 228–235, Rochester, New York, April. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and Ralph Weischedel. 2006. A study of translation error rate with targeted human annotation. In *In Proceedings of the Association for Machine Transaltion in the Americas*.

Roy W. Tromble, Shankar Kumar, Franz Och, and Wolfgang Macherey. 2008. Lattice minimum bayes-risk decoding for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 620–629, Morristown, NJ, USA. Association for Computational Linguistics.

Ying Zhang and Stephan Vogel. 2004. Measuring confidence intervals for the machine translation evaluation metrics. In *In Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-2004*, pages 4–6.