The UPV-PRHLT combination system for WMT 2011

Jesús González-Rubio and **Francisco Casacuberta**
Instituto Tecnológico de Informática
Departamento de Sistemas Informáticos y Computación
Universitat Politècnica de València
{jegonzalez|fcn}@dsic.upv.es

Abstract

This paper presents the submissions of the pattern recognition and human language technology (PRHLT) group to the system combination task of the sixth workshop on statistical machine translation (WMT 2011). Each submissions is generated by a multi-system minimum Bayes risk (MBR) technique. Our technique 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 sentences in the target language.

1 Introduction

The UPV-PHRLT approach to machine translation (MT) system combination is based on the minimum Bayes risk system combination (MBRSC) algorithm (Gonzlez-Rubio et al., 2011). A multi-system MBR technique that computes consensus translations over multiple component systems.

MBRSC 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 reranking 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). MBRSC 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 MT systems.

2 Minimum Bayes risk Decoding

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

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

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

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 $\mathcal{G}(\mathbf{e}')$:

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

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

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

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

Proceedings of the 6th Workshop on Statistical Machine Translation, pages 140–144,
Edinburgh, Scotland, UK, July 30–31, 2011. ©2011 Association for Computational Linguistics

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.

3 MBR System Combination

MBRSC 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, MBRSC 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, MBRSC is written as follows:

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

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

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

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).

3.1 Computing BLEU-based Gain

We are interested in performing MBRSC under BLEU. Therefore, we rewrite the gain function $\mathcal{G}(\cdot)$ using single evidence (or reference) BLEU (Papineni et al., 2002) as the similarity function:

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

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

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 on the lattices provided by the component systems. However, we want to explore a hypotheses space not restricted to the evidences provided by the systems.

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

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

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

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

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

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

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

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).

3.2 Model Training

The scaling factors in Eq. (8) denote the "quality" of each system with respect to the rest of them, i.e. the relative importance of each system in the Bayes gain computation. This scaling factors must be carefully tuned to obtain good translations.

We compute the scaling factor of each system as the number of times the hypothesis of the system is the best TER-scoring translation in the tuning corpora. Previous works show that this measure obtains the best translation results among other heuristic measures (González-Rubio et al., 2010) and even as good results as more complex methods such as MERT (Och, 2003). A normalization is performed to transform these counts into the range $[0.0, 1.0]$. After the normalization, a weight value of 0.0 is assigned to the lowest-scoring system, i.e. the lowest-scoring system is discarded and not taken into account in the computation of the Bayes gain.

3.3 Model Decoding

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

[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.

Algorithm 1 MBRSC decoding algorithm.

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

The AMS algorithm is shown in Algorithm 1. AMS starts with an initial hypothesis $\mathbf{e}$[3] 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 com-

[3]In the experimentation we use the evidence with minimum Bayes' risk as the initial hypothesis of the algorithm.

		cz→en	en→cz	de→en	en→de	es→en	en→es	fr→en	en→fr
#systems		12	14	25	34	15	22	23	21
dev	Worst	15.6	8.8	12.8	4.5	15.1	20.3	15.8	13.9
	Best	25.9	**16.9**	**22.2**	16.3	27.8	32.7	28.6	**35.5**
	MBRSC	**26.7**	15.9	**22.2**	**17.1**	**30.5**	**33.3**	**30.2**	34.7
test	Worst	13.3	9.1	12.9	5.1	14.7	20.7	16.1	13.0
	Best	27.2	**18.6**	21.9	**16.7**	27.4	32.5	28.1	**33.5**
	MBRSC	**27.9**	17.7	**22.1**	16.5	**30.4**	**32.9**	**29.6**	32.7

Table 1: BLEU scores (case-sensitive) on the shared translation task development and test corpora of the best and worst single systems and MBRSC. For each translation direction, we show the number of systems being combined. Best translation results are in bold.

puting the gain of a hypothesis, and usually only a moderate number of iterations (< 10) is needed to converge (Martínez et al., 2000).

4 Results

Experiments were conducted on all the 8 translation directions of the shared translation task Czech–English (cz↔en), German–English (de↔en), Spanish–English (es↔en) and French–English (fr↔en) and also on the raw and clean versions of the Haitian creole–English featured translation task (ht→en). All the experiments were carried out with the true-cased, detokenized version of the tuning and test corpora, following the WMT 2011 submission guidelines.

4.1 Shared translation task

Table 1 shows the BLEU scores of MBRSC on the development and test corpora in comparison with the score of the best and worst individual systems. In most of the translation directions, MBRSC improved the results of the best individual system, e.g. +2.7/+3.0 BLEU point in es→en. However, in en→cz and en→fr, MBRSC performs worse than the best individual system. One thing we noticed is that for these translation directions, the translations from one provided single system (online-B) were much better in terms of BLEU than those of all other systems (in the former case by more than 14% relative in development). In our experience, MBRSC requires "comparably good" systems to be able to achieve significant improvements (particularly if using heuristic scaling factors). On the other hand, we would have achieved improvements over all remaining systems leaving out online-B.

4.2 Featured translation task

Regarding the ht→en featured translation task, MBRSC is not able to improve the results of the best individual system in any case. As in the en→cz and en→fr translation directions, one of the systems (bm-i2r) perform much much better than all other systems. We can notice the surprisingly low score of one of the systems (umd-hu) in the clean task. The translations of this system are all equal ("N / A") so we suppose that some error occurred during the translation or submission processes.

	ht→en	
	raw	clean
#systems	8	16
worst	15.4	2.9
best	**29.6**	**33.1**
MBRSC	28.6	32.2

Table 2: BLEU scores (case-sensitive) on the featured translation task development corpora of the best and worst single systems and MBRSC. Best translation results are in bold.

5 summary

The UPV-PRHLT submissions for WMT 2011 system combination task were described in this paper. The combination was based on a multi-system MBR technique that 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 in the output vocabulary. We introduced expected BLEU,

an approximation to the BLEU score that allows to efficiently apply MBR in these conditions. In most of the translation directions we were able to obtain BLEU gains over the best individual systems.

Acknowledgements

This paper is based upon 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 and the UPV under grant 20091027. 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

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.

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

Jesús González-Rubio, Germán Sanchis-Trilles, Joan-Andreu Sánchez, Jesús Andrés-Ferrer, Guillem Gascó, Pascual Martínez-Gómez, Martha-Alicia Rocha, and Francisco Casacuberta. 2010. The upv-prhlt combination system for wmt 2010. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 296–300, Uppsala, Sweden, July. Association for Computational Linguistics.

Jess Gonzlez-Rubio, Alfons Juan, and Francisco Casacuberta. 2011. Minimum bayes-risk system combination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1268–1277, Portland, Oregon, USA, June. 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.

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.

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.

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.