# Minimum Bayes-Risk Decoding Extended with Similar Examples: NAIST-NICT at IWSLT 2012

Hiroaki Shimizu[1,2], Masao Utiyama[2], Eiichiro Sumita[2], Satoshi Nakamura[1]

[1]Nara Institute of Science and Technology (NAIST), Nara, Japan
[2]National Institute of Information and Communication Technology (NICT)
Hikaridai 2-2-2, Keihanna Science City, 619-0288 Kyoto, Japan
hiroaki-sh@is.naist.jp

## Abstract

This paper describes our methods used in the NAIST-NICT submission to the International Workshop on Spoken Language Translation (IWSLT) 2012 evaluation campaign. In particular, we propose two extensions to minimum bayes-risk decoding which reduces a expected loss.

## 1. Introduction

Minimum Bayes-Risk (MBR) decoding has been proposed for statistical machine translation (SMT) to minimize expected loss of translation errors under loss functions that measure translation performance (Kumar and Byrne, 2004). Those loss functions are the inverse of evaluation metrics like BLEU (Papineni et al., 2001) and NIST (Doddington, 2002).

MBR outputs translations that are similar to the other translations in the n-best list, as this reduces the expected loss if one of these other translations is actually the correct answer (see Section. 2 for details).

We extend the MBR decoding with two methods: considering similarity of each translation to the Maximum A Posteriori (MAP) translation and using training sentences pairs that is similar to the input sentence for decoding.

The proposed methods are used in the NAIST-NICT system for the International Workshop on Spoken Language Translation (IWSLT) 2012 evaluation campaign. We participated in the OLYMPICS Task, which is from Chinese to English.

## 2. Minimum Bayes-Risk decoding

### 2.1. MAP decision rule

MAP decoding finds the most likely translation $\hat{E}$ from translation candidate $E_j$ given the input sentence $F$. The MAP translation of $F$ is defined by

$$\hat{E} = \arg\max_{E_j} P(E_j|F) \tag{1}$$

This is the traditional decision rule.

### 2.2. MBR decision rule

Let $F$ and $E$ be the source and target sentences, the MBR decoding is defined as follows.

$$\hat{E} = \arg\min_{E_i} \sum_{E_j} L(E_j, E_i) P(E_j|F) \tag{2}$$

$E_i$ is the $i$-th output sentence of the n-best translations. $P(E_j|F)$ is the probability of translation $E_j$ given $F$. $L(E_j, E_i)$ is the loss function. This loss function will be defined in Section 2.3.

Note that $P(E_j|F)$ can be scaled by

$$P(E_j|F) = \frac{\exp\left(\alpha H(E_j, F)\right)}{\sum_{E_k} \exp\left(\alpha H(E_k, F)\right)} \tag{3}$$

$H(\cdot,\cdot)$ is the weighted overall score. The scaling factor $\alpha$ lies in $[0, \infty)$. If $\alpha$ is smaller, $P$ becomes equal. If $\alpha$ is larger, $P$ becomes uneven.

### 2.3. BLEU

We use BLEU as the loss function. BLEU is defined by

$$BLEU(E_j, E_i) = \mathrm{BP} \times \exp\left(\frac{1}{4}\sum_{n=1}^{4} \log p_n(E_j, E_i)\right) \tag{4}$$

where $p_n$ is the n-gram precision of $E_j$ given $E_i$ as the reference. BP is the brevity penalty.

The loss function is defined by

$$L(E_j, E_i) = 1 - BLEU(E_j, E_i) \tag{5}$$

We use a sentence-level BLEU score (Papineni et al., 2001). To solve the problem that no matches make the sentence-level BLEU score zero, we add one count to the n-gram hit and total n-gram count for $n > 1$ (C. Lin et al., 2004). We use the sentence-level BLEU score only for MBR decoding and normal BLEU score for the translation results.

By the way, if the loss function is defined as follows, (2) is as same as (1).

$$L(E_j, E_i) = \begin{cases} 1 & E_j = E_i \\ 0 & otherwise \end{cases} \tag{6}$$

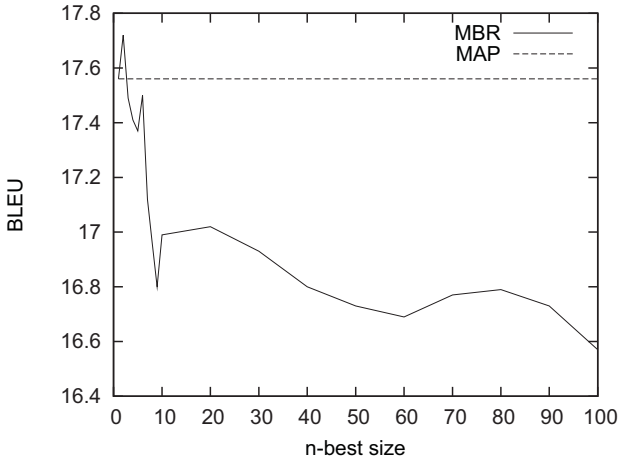


Figure 1: BLEU of MAP and MBR for development test set

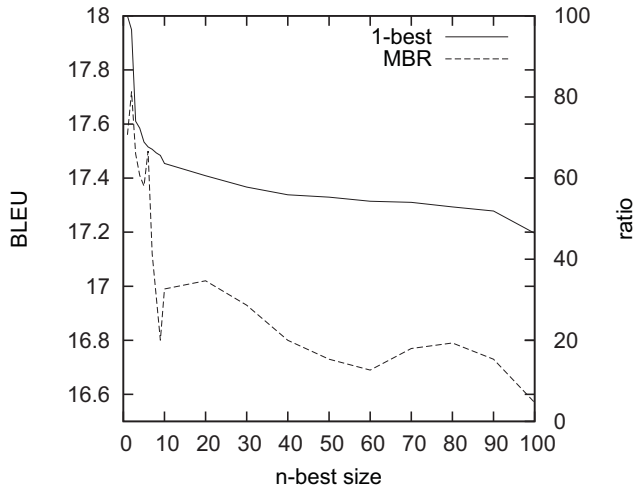


Figure 2: The relation between the percentage of the MAP in MBR and BLEU of MBR. The line "1-best" represents the percentages of MAP translations that used as MBR translations

MAP decoding is a special case of MBR translation where the loss function is defined as the 0-1 loss function formulation.

## 3. Considering similarity to MAP

In this Section, we introduce our first proposed method.

### 3.1. Observation

The motivation for the first proposed method is that we have noticed that the BLEU scores of the MBR translations were unstable with regards to different sizes of the n-best output.

Figure 1 shows to what extent the quality of the MBR translation depends on the n-best size. As shown in the figure, the BLEU of MBR translations for $n = 2$ was better than that of MAP translations on the development test set. However, $n \neq 2$ were inferior to those of MAP translations.

This observation made us conjecture that we need some modification to standard MBR decoding.

### 3.2. Proposed method 1

We conjecture that considering the similarity to the MAP translation is useful to obtain better translations, as the MAP translation is generally better than the other translations. The dotted line in Figure 2 indicates the BLEU scores of MBR translations. The line "1-best" represents the percentages of MAP translations that used as MBR translations. As shown in the figure, both lines gradually decreases as the n-best size increases. This means that when the BLEU scores of MBR translations are high, many of the MAP translations are adopted in MBR translations.

To consider the similarity to the MAP translations, we propose a method that limits the possible translations chosen by MBR to those above a certain similarity to the MAP translation. In other words, we only choose candidates that satisfy the following constraint.

$$BLEU(E_{MAP}, E_i) \geq B_1 \quad (7)$$

where $E_{MAP}$ is the MAP translation and $B_1$ is a threshold which indicates the similarity to the MAP translation. The sentence-level BLEU score (Papineni et al., 2001) also measures the similarity of MAP translations, because the loss function applied BLEU and the translation results are also measured by BLEU. Note that the n-best size, $\alpha$ and $B_1$ are decided by the development test set.

## 4. Using training data for MBR decoding

In this section, we introduce the second proposed method.

### 4.1. Motivation

Nearest neighbors choose the data which is nearest input data. Nearest neighbors of the source sentences have been used for tuning parameters (Utiyama et al. 2009, Liu et al. 2012). The second method uses nearest neighbors not for tuning but for reranking.

### 4.2. Proposed method 2

We extended MBR decoding by referencing nearest neighbors. This method is that when we use MBR decoding, we make use of training sentence pairs that are similar to the input sentence. In particular, we use nearest neighbor to improve the probability estimate $P(E_j|F)$. This can be done by

Table 1: Corpus statistics

| | | Chinese | English |
|---|---|---|---|
| Training | Sentences | 75,552 | |
| | words | 675,602 | 739,246 |
| Tuning | Sentences | 1,007 | |
| | words | 5,973 | 10,413 |
| Development test | Sentences | 1,050 | |
| | words | 5,840 | 10,364 |

defining the probability of $P(E_j|F)$ in the following form.

$$\begin{aligned} P(E_j|F) &= \sum_{F_k \in \xi} P(E_j, F_k|F) \\ &= \sum_{F_k \in \xi} P(E_j|F_k, F)\ P(F_k|F) \\ &= \sum_{F_k \in \xi} P(E_j|F_k)\ P(F_k|F) \end{aligned} \quad (8)$$

$\xi$ is defined by

$$\xi = \{G : BLEU(G, F) \geq B_2\} \cup (G \in \mathcal{F}) \quad (9)$$

where $\xi$ is a collection of input training sentences above a certain similarity to the input sentence. $\mathcal{F}$ is a collection of input training sentences and input sentence $F$. $B_2$ is a threshold which indicates the similarity to the input sentence $F$. We use sentence-level BLEU.

We interpret the probability of $P(F_k|F)$ as the probability that the input sentence $F$ can be changed to $F_k$. To use sentence-level BLEU, we define $P(F_k|F)$ as

$$P(F_k|F) = \frac{BLEU(F_k, F)}{\sum_{F_l \in \xi} BLEU(F_l, F)} \quad (10)$$

$P(E_j|F_k)$ is defined by

$$P(E_j|F_k) = \begin{cases} 1 & F_k \neq F\ and\ E_j = F_k \\ 0 & F_k \neq F\ and\ E_j \neq F_k \\ (3) & F_k = F \end{cases} \quad (11)$$

where the probability is as same as normal MBR decoding as (3) in $F_k = F$.

The advantage of this method is making use of more information. As this method uses not only n-best list of input sentence but also training sentences pairs which is similar to input sentence.

This method can be combined with Equation (7) to constrain the candidates. We call it "proposed method 1+2".

## 5. Results

### 5.1. Experiment conditions

For building the translation system, we used the phrase-based Moses (Koehn et al., 2007) decoder. We used GIZA++ (Och and Ney, 2003) for word alignment and SRILM (Stolcke, 2002) for 5-gram language model. Minimum Error Rate Training (Och, 2007) was used for tuning. We used the Stanford word segmenter (Tseng et al., 2005) for Chinese segmentation. We used the Peking University (PKU) Stanford models.

We use OLYMPIC Task data: HIT (HIT Olympic Trilingual Corpus) and BTEC (Basic Travel Expression Corpus). For the training data, we used IWSLT_BTEC.train.*, IWSLT12_BTEC.devset* and IWSLT12_HIT.train.*. For the tuning data, we used IWSLT12_HIT.devset2_IWSLT12.*, and we used IWSLT12_HIT.devset1_IWSLT12.* for the development test set data. Statistics computed over these data sets are reported in Table 1.

Table 2: The development and official BLEU score. Moses default n-best size is 200 when we use Moses MBR decoding. n=2 is the development set optimal value.

| | Dev | Official |
|---|---|---|
| MAP (baseline) | 17.56 | 17.29 |
| MBR (n=200) | 16.53 | 17.72 |
| MBR (n=2) | 17.72 | 17.30 |
| Proposed 1 (non-scale) | 17.81 | 16.96 |
| Proposed 1 | 17.96 | 16.79 |
| Proposed 2 | 17.82 | 17.45 |
| Proposed 1+2 | 17.67 | 17.39 |

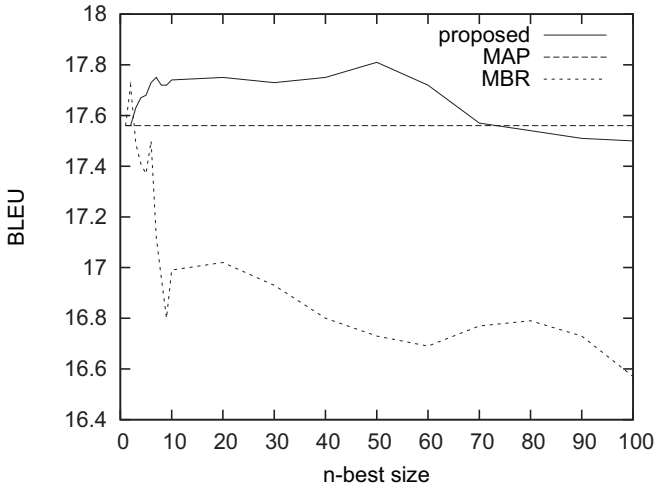


Figure 3: MAP, MBR and proposed BLEU score of development test set

### 5.2. Development test set

Table 2 shows the development BLEU score. Moses default n-best size is 200 when we use Moses MBR decoding. n=2 is the development set optimal value. When using the MAP similarity threshold over the development test set (proposed

1), we use the following parameters: the MBR scaling factor is 5, the MAP similarity threshold $B_1$is 0.75, and the n-best size is 50.

Figure 3 shows the proposed method is stable and does not depend on the size of the n-best list. When $n = 50$, we can obtain the best BLEU score. If the similarity is less than 0.75, the graph looks like MBR and many translations are normal MBR translations, because similarity to MAP is not considered. If the similarity is more than 0.75, the graph looks like MAP and many of translations are MAP translation. As the limiting condition becomes very severe. The scaling factor 5 is the optimal value for the development test set.

Table 2 shows the BLEU of the nearest neighbor method also uses on the development test set (proposed 2). We use the following parameters: MBR scaling factor is 5, the similarity to the input sentence $B_2$ is 0.7 and n-best size is 50.

### 5.3. Official test set

Table 2 also shows the official test set result. The proposed method 1 is not better than the baseline by 0.5. However, the proposed method 2 is better than the baseline by 0.16. The proposed 1+2 is better than the baseline by 0.1.

## 6. Discussion

First, we discuss about the wrong result of the proposed 1. It can be seen that the results for the MBR decoding for development test set is not good. One of the reasons is that many of MAP translations are good so considering the similarity to MAP translation worked well. However, MBR decoding for official test set is good. We guess that most MAP translations did not have as good quality, and standard MBR translation is stable in the size of the n-best list. So, considering similarity to MAP translation made the result worse. We guess that the proposed method 1 is a valid method for data in which MBR decoding does not have a positive effect. In addition, we use Word Error Rate (WER) for the loss function (5). However the result of WER is worse than that of BLEU.

The proposed method 2 is effective for this official test set. However proposed method 1+2 is not better than method as method 1 is not effective.

## 7. Conclusion

We participated in the OLYMPICS Task. Our system extended MBR decoding with two methods. The method using training data for MBR decoding improved BLEU scores.

## 8. Acknowledgement

We thank Graham Neubig for his comments on this paper.

## 9. References

[1] G. Doddington. 2002. Automatic Evaluation of Machine Translation Quality using N-gram Co-occurrence Statistics. In Proceedings of Human Language Technology Conference.

[2] S. Kumar and W. Byrne. 2004. Minimum Bayes-Risk Decoding for Statistical Machine Translation. In Human Language Technologies: North American Chapter of the Association for Computational Linguistics, pages 169–176, Boston, MA, USA.

[3] K. Papineni, S. Roukos, T. Ward and W. Zhu. 2001. BLEU: a Method for Automatic Evaluation of Machine Translation. Technical Report RC22176 (W0109–022) , IBM Research Division.

[4] L. Chin-Yew and F. Och. 2004. ORANGE : a Method for Evaluating Automatic Evaluation Metric for Machine Translation. In Proceedings of the 20th International Conference on Computational Linguistics, Geneva, Switzerland.

[5] M. Utiyama, H. Yamamoto and E. Sumita. 2009. Two Methods for Stabilizing MERT: NICT at IWSLT 2009. In proceedings of IWSLT, page 79–82.

[6] L. Liu, H. Cao, T. Watanabe, T. Zhao, M. Yu and C. Zhu. 2012. Locally Training the Log-Linear Model for SMT. In Proceedings of Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, page 402–411.

[7] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C.Dyer, O. Bojar, A. Constantin and E. Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, page 177–180.

[8] F. J. Och and H. Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. In Proceedings of Computational Linguistics, vol. 29, No.1 , page 19–51.

[9] A. Stolcke. 2003. SRILM - An Extensible Language Modeling Toolkit. In Proceedings of International Conference on Spoken Language Processing.

[10] F. J. Och. 2007. Minimum Error Rate Training in Statistical Machine Translation. In Proceedings of Association for Computational Linguistics.

[11] H. Tseng, P. Chang, G. Andrew, D. Jurafsky and C. Manning. 2005. A Conditional Random Field Word Segmenter. In Proceedings of Fourth SIGHAN Workshop on Chinese Language Processing.