Improving Phrase Extraction via MBR Phrase Scoring and Pruning

Nan Duan, Tianjin University
Mu Li, Ming Zhou, Microsoft Research Asia
Lei Cui Harbin Institute of Technology
{v-naduan,muli,mingzhou,v-lecu}@microsoft.com

Abstract

One of the major reasons for translation errors in phrase-based SMT systems is the incorrect phrases induced from inaccuracy word-aligned parallel data. In this paper, we propose a novel approach that uses the minimum Bayes-risk (MBR) principle to improve the accuracy of phrase extraction. Our approach performs as a four-stage pipeline: first, bilingual phrases are extracted from parallel corpus using a standard phrase induction method; then, phrases are separated into groups under specific constraints and scored using an MBR model; next, word alignment links contained in phrases with their MBR scores lower than a certain threshold are pruned in the parallel data; last, a new phrase table is learned from the link-pruned parallel data and used in SMT decoding. We evaluate our approach on the NIST Chinese-to-English MT tasks, and show significant improvements on parallel data sets of different scales.

1 Introduction

Bilingual phrases are the fundamental building blocks for phrase-based SMT systems (Och and Ney, 2004; Koehn et al., 2004a; Chiang, 2005), and their abilities to handle local reorderings and translation ambiguity as well as many-to-many word translations are key factors to the success of phrasal SMT models.

The common practice of extracting bilingual phrases from the parallel data usually consists of three steps: first, words in bilingual sentence pairs are aligned using state-of-the-art automatic word alignment tools, such as GIZA++ (Och and Ney, 2003), in both directions; second, word alignment links are refined using heuristics, such as Grow-Diagonal-Final (GDF) method; third, bilingual phrases are extracted from the parallel data based on the refined word alignments with predefined constraints (Och and Ney, 2003).

Such phrase extraction methods, however, are not performed in a clean room. They are usually subject to various kinds of errors, such as noises in training corpus or mistakes caused by word alignment models. These errors could produce low-quality bilingual phrases in the final phrase table. For example, Figure 1 shows four English translations for the source Chinese phrase 氢弹 extracted from different training instances, and two of them contain errors: in (c), there are two irrelevant words included and one word missing, while phrase pair in (d) is totally wrong. Because of appearing in training corpus several times, all these bilingual phrases were maintained in the generated phrase table as valid entries.

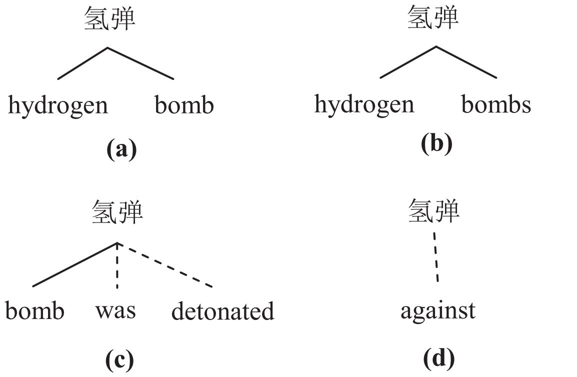

Figure 1: Phrase pairs extracted from different bilingual sentence pairs with the same source phrases, in which dashed lines denote wrong alignment links.

Incorrect phrase entries fed into SMT decoder are one of the major reasons for translation errors in phrase-based SMT systems. For example, even (d) doesn't occupy a large portion of probabilities in all translation alternatives of the source phrase 氢弹, it is still picked up by SMT decoder sometimes, because it is strongly preferred by the language model in certain circumstances.

Motivated by the success of consensus-based methods in SMT research (Kumar and Byrne, 2002; Kumar and Byrne, 2004; Ueffing and Ney, 2007; Kumar et al., 2009), this paper proposes a novel approach that makes use of MBR principle to improve the accuracy of phrase extraction. Our approach operates as a four-stage pipeline: first, bilingual phrases are extracted from parallel corpus using a standard phrase induction method; then, phrases are separated into groups under specific constraints and scored using an MBR model; next, word alignment links contained in phrases with their MBR scores lower than a certain threshold are pruned in the parallel data; last, a new phrase table is learned from the link-pruned parallel data and used in SMT decoding.

We evaluate on a state-of-the-art phrase-based SMT decoder on the NIST Chinese-to-English MT tasks, and experiments show that our MBR-based approach outperforms the standard phrase extraction method by up to 1.45 BLEU points.

2 A New MBR-based Phrase Extraction Pipeline

In order to learn a phrase table from the bilingual corpus, two major issues need to be addressed: i) which phrase pairs should be considered as valid entries; and ii) how to estimate feature values, e.g. translation probabilities and lexical weights, for these entries. The first issue is often referred as *phrase extraction*.

Let $\{(E_k, F_k)\}$ denote sentence pairs contained in a training corpus $\mathcal{C}$, $\mathcal{A}_k = \{(i,j)\}$ denote word alignment links of (E_k, F_k) that are generated by refining the bi-direction alignment results using heuristics rules (such as GDF). The objective of phrase extraction is to extracted all phrase pairs $\{(a, e, f)\}$ from word-aligned sentence pairs in $\mathcal{C}$ based on the *alignment consistency*[1] with length constraints, in which (a, e, f) denotes one phrase pair, f and e are the source and target phrases respectively, $a \in \mathcal{A}$ denotes a set of links that connect words contained in f and e.

This paper presents an MBR–based approach to enhance the accuracy of phrase extraction, which contains four steps including: (1) *baseline phrase extraction*, (2) *MBR phrase scoring*, (3) *alignment pruning*, and (4) *phrase re-extraction*. In following subsections, we will present details of these four steps one by one.

2.1 Step 1: Baseline Phrase Extraction

In this step, all potential phrase pairs that are consistent with word alignments are extracted from a given training corpus $\mathcal{C}$ using the standard phrase extraction method. Furthermore, inspired by several studies (Mi et al., 2008; Dyer et al., 2008; Venugopal et al., 2008; Liu et al., 2009), in which *n*-best alternatives of annotations to SMT systems are leveraged to improve translation quality, we allow our proposed phrase extraction method to operate on *n*-best word alignments as well: given a sentence pair with *n*-best alignment candidates, we use alignments in the *n*-best list one at a time with the same sentence pair to form a new word-aligned sentence pair, and annotate it with the posterior probability of the alignment it used. These posterior probabilities will be used in the next step to compute MBR model scores.

2.2 Step 2: MBR Phrase Scoring

The objective of this step is to score phrase pairs extracted in Step 1 based on an MBR model. Similar to those bi-direction translation features, we compute two MBR scores for each phrase pair from two different language sides as well.

We first consider scoring phrase pairs based on their source phrases. Given all phrase pairs $\{(a', e', f)\}$ with the same source side f, we define a score $S_f(\rho)$ that is assigned to each phrase pair $\rho = (a', e', f)$ based on an MBR model as:

$$S_f(\rho) = \sum_{\rho' \in \mathcal{H}(f)} G(\rho, \rho') P(\rho' | f) \qquad (1)$$

- $\mathcal{H}(f)$ is the *hypothesis space* that contains all phrase pairs $\{(a', e', f)\}$ extracted in Step 1 sharing the same source phrase f. Each ρ' denotes one hypothesis.
- $G(\rho, \rho')$ is the *gain function* that is defined as $G(\rho, \rho') = M - L(\rho, \rho')$, the supplement of the loss function in a standard MBR definition, where M is a constant large enough. We define $G(\rho, \rho')$ as the similarity measure between two hypotheses ρ and ρ'. In this sense, $S_f(\rho)$ can be viewed as the expected similarity between ρ and all hypotheses in $\mathcal{H}(f)$. W formulate $G(\rho, \rho')$ as a weighted combination of a set of similarity features:

$$G(\rho, \rho') = \sum_i \lambda_i \theta_i(\rho, \rho') \qquad (2)$$

  where θ_i is the i^{th} feature with its weight λ_i.
- $P(\rho|f)$ is the *hypothesis distribution* over all hypotheses contained in $\mathcal{H}(f)$:

$$P(\rho|f) = \frac{\sum_{(E,F)\in\mathcal{C}} \sum_{\mathcal{A}} \delta_{(\mathcal{A},E,F)}(\rho) P(\mathcal{A}|E,F)}{\sum_{\rho' \in H(f)} \sum_{(E,F)\in\mathcal{C}} \sum_{\mathcal{A}} \delta_{(\mathcal{A},E,F)}(\rho') P(\mathcal{A}|E,F)}$$

[1] Given a source phrase f and a target phrase e, the phrase pair (e, f) is said to be *consistent with word alignment* if and only if: (1) at least one word in one phrase is aligned to one word in the other phrase; (2) no words in one phrase can be aligned to a word outside the other phrase.

where $\delta_{(\mathcal{A},E,F)}(\rho)$ equals to 1 when ρ can be extracted from $(\mathcal{A},E,F)$, and 0 otherwise, $P(\mathcal{A}|E,F)$ is the posterior probability[2] of $\mathcal{A}$:

$$P(\mathcal{A}|E,F) = \frac{exp\,\{\alpha \cdot \varphi(\mathcal{A},E,F)\}}{\sum_{\mathcal{A}'} exp\,\{\alpha \cdot \varphi(\mathcal{A}',E,F)\}}$$

where $\varphi(\mathcal{A},E,F)$ is the score predicted by the alignment model for an alignment $\mathcal{A}$, α controls the entropy of resulting distribution. $\alpha > 1$ makes the distribution more peak, while $0 \leq \alpha \leq 1$ makes the distribution more uniform. Due to the fact that varying α to modify the entropy of the alignment distribution doesn't have consistent impacts on translation quality (Venugopal et al., 2008), in this paper we just fix this value to be 1.0.

We rewrite Equation (1) by replacing $G(\rho,\rho')$ using Equation (2) as:

$$\begin{aligned} S_f(\rho) &= \sum_i \lambda_i \{ \sum_{\rho' \in \mathcal{H}(f)} \theta_i(\rho,\rho')P(\rho'|f)\} \\ &= \sum_i \lambda_i Sim_i(\rho,\mathcal{H}(f)) \end{aligned} \qquad (3)$$

$Sim_i\big(\rho,\mathcal{H}(f)\big) = \sum_{\rho' \in \mathcal{H}(f)} \theta_i(\rho,\rho')P(\rho'|f)$ is defined as the expected value of the i^{th} similarity feature θ_i for ρ based on the entire $\mathcal{H}(f)$.

We then consider scoring phrase pairs based on their target phrases. Given all phrase pairs $\{(a',e,f')\}$ with the same target side e, we score each of them in a similar way as in Equation (3):

$$\begin{aligned} S_e(\rho) &= \sum_j \lambda_j \sum_{\rho' \in \mathcal{H}(e)} \theta_j(\rho,\rho')P(\rho'|e) \\ &= \sum_j \lambda_j Sim_j(\rho,\mathcal{H}(e)) \end{aligned} \qquad (4)$$

$Sim_j\big(\rho,\mathcal{H}(e)\big) = \sum_{\rho' \in \mathcal{H}(e)} \theta_j(\rho,\rho')P(\rho'|e)$ is defined as the expected value of the j^{th} similarity feature θ_j for ρ based on the entire $\mathcal{H}(e)$.

Algorithm 1 shows how we obtain the feature weights $\{\lambda_i\}$ and $\{\lambda_j\}$ in Equation (3) and (4) that are necessary for computing $S_f(\rho)$ and $S_e(\rho)$.

First, we generate an temporary phrase table by using all phrase pairs extracted in Step 1 with two sets of similarity scores $\{Sim_i(\rho,\mathcal{H}(f)\}$ and $\{Sim_j(\rho,\mathcal{H}(e)\}$ maintained for each phase pair as additional phrasal features; then, we use this phrase table in our log-linear SMT system and optimize the weights of these similarity scores together with the weights of original SMT model features to maximize BLEU on development data set using the MERT algorithm proposed by Och (2003)[3]; last, we collect the well-tuned feature weights $\{\lambda_i\}$ and $\{\lambda_j\}$ and compute $S_f(\rho)$ and $S_e(\rho)$ for each ρ based on Equation (3) and (4).

Algorithm 1: MBR Phrase Scoring

```
for each unique source phrase $f$ do
    initialize a hypothesis space $\mathcal{H}(f) = \{\emptyset\}$
    for each phrase pair $(a',e',f)$ do
        add $\rho = (a',e',f)$ to $\mathcal{H}(f)$
    end for
    for each hypothesis $\rho$ in $\mathcal{H}(f)$ do
        compute $\{Sim_i(\rho,\mathcal{H}(f)\}$ based on $\mathcal{H}(f)$
        annotate $\rho$ with a set of similarity feature
        scores $\{Sim_i(\rho,\mathcal{H}(f)\}$
    end for
end for
for each unique target phrase $e$ do
    initialize a hypothesis space $\mathcal{H}(e) = \{\emptyset\}$
    for each phrase pair $\{a',e,f'\}$ do
        add $\rho = \{a',e,f'\}$ to $\mathcal{H}(e)$
    end for
    for each hypothesis $\rho$ in $\mathcal{H}(e)$ do
        compute $\{Sim_j(\rho,\mathcal{H}(e)\}$ based on $\mathcal{H}(e)$
        annotate $\rho$ with a set of similarity feature
        scores $\{Sim_j(\rho,\mathcal{H}(e)\}$
    end for
end for
generate a phrase table where each phrase pair $\rho$ is
annotated with similarity scores $\{Sim_i(\rho,\mathcal{H}(f)\}$
and $\{Sim_j(\rho,\mathcal{H}(e)\}$ as additional features
run MERT to optimize the weights of similarity
features together with original SMT features
collect feature weights $\{\lambda_i\}$ and $\{\lambda_j\}$ and compute
MBR scores $S_f(\rho)$ and $S_e(\rho)$ for each $\rho$
```

2.3 Step 3: Alignment Pruning

In order to discard low-quality phrase pairs from the final phrase table to alleviate decoding errors, an alignment pruning strategy is proposed.

Given each phrase pair $\rho = \{a,e,f\}$, we first find out two maximum MBR scores, $\hat{S}_f(\rho')$ and

[2] Aiming to estimate hypothesis distribution $P(\rho|f)$ for our MBR model, we utilize alignment posterior probabilities of the whole sentence pairs that contain current source phrase f. Doing so is based on an assumption that the distribution of any context independent phrase is identical to its distribution when it occurs within a sentence.

[3] In previous work, MBR has been used to introduce decision criteria that are directly related to the evaluation metric for the tasks (e.g., BLEU, AER, etc.). In our model, MBR is used to define similarity scores for phrase pairs, which have no direct correlations to the final evaluation criteria, BLEU. However, as we optimize these MBR features together with original SMT features, we have already established relations between our MBR model and final translation quality.

$\hat{S}_e(\rho')$, from its corresponding hypothesis spaces $\mathcal{H}(f)$ and $\mathcal{H}(e)$ respectively and multiply them to obtain a reference value as $S_{max}(\rho)$; we then prune all alignment links contained in ρ from $\mathcal{C}$, if the product of $S_f(\rho)$ and $S_e(\rho)$ is lower than $S_{max}(\rho)$ by a certain threshold t (Algorithm 2).

Algorithm 2: Alignment Pruning

```
for each phrase pair $\rho = \{a, e, f\}$ do
    find $\hat{S}_f(\rho) = Max_{\rho' \in \mathcal{H}(f)}\{S_f(\rho')\}$ from $\mathcal{H}(f)$
    find $\hat{S}_e(\rho) = Max_{\rho' \in \mathcal{H}(e)}\{S_e(\rho')\}$ from $\mathcal{H}(e)$
    $S_{max}(\rho) = \hat{S}_f(\rho) * \hat{S}_e(\rho)$
    if $\{S_f(\rho) * S_e(\rho)\} < \{S_{max}(\rho) * t\}$ then
        prune all alignment links contained in $\rho$ from
        the positions they were extracted in $\mathcal{C}$
    end if
end for
return $\mathcal{C}$ with link-pruned word alignments
```

One question may be asked is the reason that we remove all alignment links from phrase pairs of relative low MBR scores. In fact, although all alignment links are not necessarily bad in those low-quality phrase pairs, we just remove all of them from training corpus for convenience. By varying different values of t, we can empirically find an optimal setting, where alignment pruning can bring benefits for final translation quality.

2.4 Step 4: Phrase Re-Extraction

Last, we re-extract bilingual phrases based on the link-pruned training corpus to learn a new phrase table[4]. For each phrase pair in this phrase table, we also compute two sets of expected similarity scores based on the MBR model, and use them as extra phrasal features. In our experimental part, we will show that besides alignment pruning, using similarity scores as additional features can provide further improvements as well.

When using *n*-best alignment results instead of 1-best ones, translation probabilities and lexical weights are estimated based on fractional counts instead of absolute frequencies of phrases.

3 Similarity Features

This section presents all similarity features that are used in computing $G(\rho, \rho')$. We summarize them into two categories as follows.

- ***alignment-based features***

1) $\theta_{W2W}(\rho, \rho')$. A feature that counts how many (source word)-to-(target word) link pairs in ρ co-occur in ρ'.

2) $\theta_{W2P}(\rho, \rho')$. A feature that counts how many (source word)-to-(target word's POS) link pairs in ρ co-occur in ρ'. Two MaxEnt-based POS taggers (Ratnaparkhi, 1996) are used to tag Chinese and English words contained in the bilingual corpus respectively.

3) $\theta_{W2C}(\rho, \rho')$. A feature that counts how many (source word)-to-(target word's class) link pairs in ρ co-occur in ρ'. Word clusters are obtained by using *mkcls* toolkit (Och, 1999) that trains word classes based on the maximum-likelihood criterion. The total numbers of word classes are set to be 80 for both Chinese and English.

4) $\theta_{W2S}(\rho, \rho')$. A feature that counts how many (source word)-to-(target word's stem) link pairs in ρ co-occur in ρ'. A stem dictionary that contains 22,660 entries is used to convert English words into their stem forms. We consider the stem for each Chinese word as the Chinese word itself.

5) $\theta_{Fert}(\rho, \rho')$. A feature that reflects the agreement on word fertilities for ρ and ρ':

$$\theta_{Fert}(\rho, \rho') = \#(a)\delta_{\#(a)}(\#(a'))$$

$\#(a)$ is the number of word link pairs in a, and $\delta_{\#(a)}(\#(a'))$ equals to 1 when $\#(a) = \#(a')$, and 0 otherwise.

6) $\theta_{Ratio}(\rho, \rho')$. A feature that reflects the agreement on link ratio defined as $r(\rho) = \#(\rho)/\{|\rho_s| + |\rho_t|\}$ between ρ and ρ', where $\#(\rho)$ is the total number of linked words contained in both sides of ρ, ρ_s and ρ_t are source and target phrases of ρ respectively:

$$\theta_{Ratio}(\rho, \rho') = r(\rho)\delta_{r(\rho)}(r(\rho'))$$

We use alignment-based features due to the fact that alignment quality usually determines the quality of phrase pairs. Besides words, we also incorporate the knowledge of POS tags, word classes and word stems to alleviate data sparseness and to make our model to be more general.

- ***n-gram-based features***

7) $\theta_n(\rho, \rho')$. A feature that counts how many n-grams in ρ_t co-occur in $\rho_t{}'$:

[4] By alignment pruning, low-quality bilingual phrases will be absent in final generated phrase table. Furthermore, some phrase pairs that cannot be extracted from training corpus based on original alignment results could become available as well, as the fact that those alignment errors are removed.

$$\theta_n(\rho, \rho') = \sum_{\omega \in \rho_t} \#_\omega(\rho_t)\delta_{\rho'_t}(\omega)$$

$\#_\omega(\rho_t)$ is the number of times that ω occurs in ρ_t, $\delta_{\rho'_t}(\omega)$ equals to 1 when ω occurs in ρ'_t, and 0 otherwise . In this paper, the order of n-gram considered varies from 1 to 4.

8) $\theta_{Len}(\rho, \rho')$. A feature that reflects the agreement on word lengths for ρ_t and ρ'_t:

$$\theta_{Len}(\rho, \rho') = |\rho_t|\delta_{|\rho_t|}(|\rho'_t|)$$

$\delta_{|\rho_t|}(|\rho'_t|)$ equals to 1 when $|\rho_t| = |\rho'_t|$, and 0 otherwise.

These features are motivated by the success of consensus-based techniques (Kumar and Byrne, 2004; Tromble et al., 2008; Kumar et al., 2009).

To summarize, 6 features are contained in the first category and 5 features are contained in the second category. Because that source and target phrases are exchangeable for each phrase pair, there will be (2*11=22) similarity features in total for each bilingual phrase[5].

4 Experiments

4.1 Data and Metric

We evaluate on the NIST Chinese-to-English MT tasks. The NIST 2003 (MT03) data set is used as the development set to tune model parameters, and evaluation results are reported on the NIST 2005 (MT05) and 2008 (MT08) data sets.

Two parallel data sets with different scales are used as training corpus: the first data set includes FBIS only, which contains 128K sentence pairs after pre-processing. We investigate the impacts of different parameter settings on this corpus; the second data set includes the following data sets, LDC2003E07, LDC2003E14, LDC2005T06 and LDC2005T10 with 354K sentence pairs contained after pre-processing. We confirm the effectiveness of our method on it using the optimal parameter setting determined on the first data set. We train a 5-gram language model on the Xinhua portion of LDC English Gigaword Version 3.0.

Translation quality is measured in terms of the case-insensitive *IBM-BLEU* scores that compute the brevity penalty using the closest reference translation (Papineni et al., 2002). Statistical significance is computed by using the bootstrap re-sampling method proposed by Koehn (2004b).

4.2 SMT Decoder

A re-implemented phrase-based SMT decoder (Xiong et al., 2006) with a lexicalized reordering component based on maximum entropy is used to generate translation outputs. Both the phrase table and the reordering model are trained on the same bilingual corpus used. The default beam size is set to be 100, and MERT algorithm (Och, 2003) is utilized to optimize model parameters.

Because of using MBR-inspired techniques, we also investigate the impacts of our method on MBR decoding over translation hypergraphs of the baseline system. We re-implement an MBR decoder (Kumar et al., 2009) that uses the linear BLEU score as its loss function.

4.3 Word Aligner

Although discriminative methods have already shown comparable word alignment accuracy in benchmarks, generative methods are still widely used to produce word alignments for large scale corpus. As a result, we evaluate our approach based on two different word aligners.

- *Disc-Aligner*. A discriminative word aligner (Fraser and Marcu, 2006) is re-implemented to predict alignments for the training corpus. A data set of 491 sentence pairs with human annotated word alignments is used to tune model parameters. Disc-Aligner can produce *n*-best alignment alternatives.
- *GIZA-Aligner*: An unsupervised word aligner GIZA++ (Och and Ney, 2003) is used with the default parameters. In this paper, we only use its Viterbi (1-best) alignment outputs.

4.4 Baseline Phrase Extraction Method

The standard phrase extraction method (*Base-PE*) proposed by Och and Ney (2004) is utilized to generate the baseline phrase table. The length limitations are set to be 5 and 10 for source and target phrases respectively. All phrase pairs that occur only once in the training corpus are pruned.

4.5 Results on the First Data Set

We denote our MBR-based phrase extraction method as *MBR-PE* and compare it to Base-PE on the first data set (Table 1). Disc-Aligner is used to predict word alignments.

We first use Base-PE to generate two baseline phrase tables, *Base-PE*$^{1\text{-}best}$ and *Base-PE*$^{15\text{-}best}$, using 1-best and 15-best alignments respectively; we then use MBR-PE to generate two improved phrase tables, *MBR-PE*$^{1\text{-}best}$ and *MBR-PE*$^{15\text{-}best}$, using link-pruned 1-best and 15-best alignments.

[5] The stability of MERT should be concerned when using so many features. Here, we alleviate this issue by enlarging the beam size and increasing the rounds of MERT iterations.

The pruning threshold is set to be 10^{-5}. Similarity features computed in our MBR model for phrase pairs are also used as features in decoding.

	IBM-BLEU%		
	MT03	MT05	MT08
Base-PE$^{1\text{-best}}$	32.87	31.40	21.09
Base-PE$^{15\text{-best}}$	33.38*	31.83*	21.57*
MBR-PE$^{1\text{-best}}$	**33.50***	**32.18***	**21.77***
MBR-PE$^{15\text{-best}}$	**34.47***	**32.85***	**22.41***

Table 1: MBR-PE vs. Base-PE on the first data set (*: significantly better than the results of Base-PE$^{1\text{-best}}$ with $\boldsymbol{p} < 0.05$)

We can draw several conclusions from Table 1: (i) by using n-best instead of 1-best alignments in phrase extraction, significant improvements are obtained (+0.43/+0.48 BLEU on MT05/MT08), which confirm that SMT systems can benefit from making the annotation pipeline wider; (ii) both MBR-PE$^{1\text{-best}}$ and MBR-PE$^{15\text{-best}}$ outperform two baseline phrase tables. These improvements, we think, mainly come from using all similarity features as additional features in decoding, which brings more discriminative power to the SMT model, and using alignment pruning to remove those low-quality phrase pairs, which reduces the possibility of making decoding errors; (iii) MBR-PE$^{15\text{-best}}$ performs significantly better than MBR-PE$^{1\text{-best}}$. We think that the reasons are two-fold: 1) more phrase pairs included in hypothesis spaces makes the computation of consensus statistics to be more accurate; 2) more alignments involved makes the hypothesis distributions to be more accurate. Compared to Base-PE$^{1\text{-best}}$, MBR-PE$^{15\text{-best}}$ obtains +1.45/+1.32 BLEU on MT05/MT08.

We arrange an experiment to test the effects of our method on MBR decoding technique. Given translation hypergraphs generated by the four SMT systems used in Table 1, we perform MBR decoding and present results in Table 2.

+ MBR Decoding	IBM-BLEU%		
	MT03	MT05	MT08
Base-PE$^{1\text{-best}}$	33.76	32.05	21.82
Base-PE$^{15\text{-best}}$	34.09	32.40	22.07
MBR-PE$^{1\text{-best}}$	**33.82**	**32.49**	**22.11**
MBR-PE$^{15\text{-best}}$	**34.69**	**33.14**	**22.60**

Table 2: MBR-PE vs. Base-PE on MBR decoding over translation hypergraphs

From Table 2 we can see that the results of MBR-PE are still consistently better than the results of Base-PE on MBR decoding. However, we notice that improvements of MBR decoding on MBR-PE are smaller than the ones on Base-PE. This may be caused by the fact that we have already included consensus-based information in our MBR-based phrase scoring model and used them as additional features in SMT decoding.

4.5.1 Effect of Using MBR-PE Iteratively

We are interested in a question that whether the iterative usage of MBR-PE could bring more improvements. In order to clarify this question, we perform MBR-PE on the same corpus used in Table 1 based on alignments pruned already, and denote the output phrase table as $MBR\text{-}PE_{iteration}$. Evaluation results using such phrase tables on all data sets are presented in Table 3.

	IBM-BLEU%		
	MT03	MT05	MT08
MBR-PE$_{\text{iteration}}$(1-best)	**33.01**	**31.78**	**21.49**
MBR-PE$_{\text{iteration}}$(15-best)	**33.85**	**32.37**	**21.91**

Table 3: Effects of using MBR-PE iteratively

The results of MBR-PE$_{\text{iteration}}$ are worse than those of MBR-PE on all test sets. We think it is caused by the fact that after the 1st round MBR-PE procedure finished, phrase pairs contained in the same hypothesis space have already been highly correlated. When more iterations of our MBR-PE proceed with more word links pruned, the recall of phrase pairs extracted became low, which degrade the translation quality. In fact, the effect of iterative usage of MBR-PE procedure equals to the effect of enlarging the pruning threshold somewhat. When the pruning threshold is well-tuned, few improvements can be further obtained by using our approach iteratively.

4.5.2 Effect of Alignment Pruning

We next investigate the effect of word alignment pruning. This time, we still use the similarity features to score each phrase pairs, but do not perform alignment pruning. The phrase tables generated are denoted as $MBR\text{-}PE_{nopruning}$. All similarity features are used as additional features in decoding procedures. Evaluation results are shown in Table 4.

	IBM-BLEU%		
	MT03	MT05	MT08
MBR-PE$_{\text{nopruning}}$(1-best)	**32.96**	**31.64**	**21.29**
MBR-PE$_{\text{nopruning}}$(15-best)	**33.74**	**32.19**	**21.77**

Table 4: MBR-PE without alignment pruning

The results of MBR-PE$_{\text{nopruning}}$ are better than the results of Base-PE$^{1\text{-best}}$ and Base-PE$^{15\text{-best}}$ but worse than the results of MBR-PE$^{1\text{-best}}$ and MBR-PE$^{15\text{-best}}$ in Table 1. We think this is caused by the fact that, although phrase pairs with low qualities have already penalized by MBR scores, they still take part in the competition in decoding, which could also bring translation errors.

4.5.3 Effect of Similarity Features

We investigate the effects of using similarity scores included in our MBR model as additional features in SMT decoding (Table 5). This time, we still use two phrase tables, MBR-PE$^{1\text{-best}}$ and MBR-PE$^{15\text{-best}}$, generated by MBR-PE in Table 1, but discard all similarity scores during decoding. We denote such systems as *MBR-PE$_{\text{-SimFeats}}$*.

	IBM-BLEU%		
	MT03	MT05	MT08
MBR-PE$_{\text{-SimFeats}}$(1-best)	**33.05**	**31.76**	**21.41**
MBR-PE$_{\text{-SimFeats}}$(15-best)	**33.97**	**32.23**	**21.89**

Table 5: Results of MBR-PE without using similarity scores as additional features in SMT decoding

Table 5 tells us that both of these two systems perform better than their baseline systems (Base-PE$^{1\text{-best}}$ and Base-PE$^{15\text{-best}}$). However, their results are still worse than the results of MBR-PE$^{1\text{-best}}$ and MBR-PE$^{15\text{-best}}$, which use similarity scores as additional features during decoding.

Table 4 and Table 5 show that both alignment pruning and using similarity scores as additional features are necessary for getting better results, and the best choice is to use them together.

4.5.4 Effect of Pruning Thresholds

The pruning threshold is an important parameter, because it is highly related to the percentage of alignment links to be pruned. In this experiment, we investigate the effects of different pruning thresholds, and show their impacts on translation results in Table 6. We also list the corresponding sizes of different phrase tables. The small scale corpus with 15-best alignments predicted by Disc-Aligner is used in MBR-PE.

Threshold	IBM-BLEU%			Table Size
	MT03	MT05	MT08	
10^{-8}	33.93	32.17	21.60	6.93M
10^{-7}	33.97	32.18	21.67	6.67M
10^{-6}	34.04	32.25	21.86	6.09M
$\mathbf{10^{-5}}$	**34.47**	**32.85**	**22.41**	**5.29M**
10^{-4}	33.31	31.92	21.34	4.23M
10^{-3}	32.34	30.66	19.76	2.84M

Table 6: Effects of different pruning thresholds on translation quality and phrase table sizes

When enlarging t to prune more word links as errors, sizes of phrase tables generated decrease monotonically. However, after reaching a peak point ($t = 10^{-5}$), translation performance begins to drop as t increases. When t is set to be 10^{-3}, the performance is even worse than baseline's. These results tell us that, in order to get the best translation quality, we must balance precision and recall for the phrase pairs extracted.

4.5.5 Effect of *N*-best Alignment Sizes

The size of *n*-best alignments used is another flexible parameter in our approach. The larger *n* we used, the more phrase pairs can be extracted and used in our MBR model. In this experiment, we vary this size from 1 to 20, and choose the best *n* based on the evaluation results on dev set (MT03) for use in blind tests. Based on different evaluation results, we draw a curve in Figure 2.

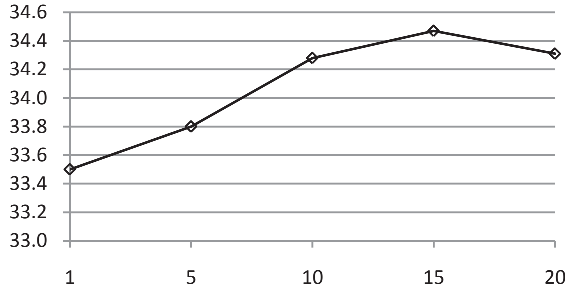

Figure 2: Effects of different *n*-best alignment sizes used in MBR-PE on MT03

The trend of the curve in Figure 2 shows that, when using more alignment candidates instead of using 1-best alignment only, translation qualities become consistently better. The potential reasons of these results are two-fold: (i) more involved alignments bring more potential phrase pairs, which enlarge the hypothesis spaces for MBR computation; (ii) more posterior probabilities of alignments make the hypothesis distributions to be more accurate. However, we didn't use word alignments more than 20-best for each sentence pair. This is due to the efficiency reason in MBR computation as well as the fact that performances change few after 15-best alignments are used.

4.5.6 Effect of Different Feature Categories

We want to know the impacts of different feature categories described in Section 3 on translation performances as well.

Given the phrase tables generated by MBR-PE in Table 1, we use *MBR-PE(align)* to denote the results using alignment-based similarity scores as additional features only during SMT decoding, and use *MBR-PE(consen)* to denote the results using consensus-based similarity scores as additional features only during SMT decoding. We compare them with the results of MBR-PE that use both of these two feature categories as additional features during SMT decoding, and list evaluation results in Table 7.

	IBM-BLEU%		
	MT03	MT05	MT08
MBR-PE$^{1\text{-best}}$(align)	33.09	31.87	21.58
MBR-PE$^{1\text{-best}}$(consen)	33.21	31.89	21.55
MBR-PE$^{1\text{-best}}$	**33.50**	**32.18**	**21.77**
MBR-PE$^{15\text{-best}}$(align)	34.02	32.43	21.99
MBR-PE$^{15\text{-best}}$(consen)	34.18	32.60	22.19
MBR-PE$^{15\text{-best}}$	**34.47**	**32.85**	**22.41**

Table 7: Effects of different feature categories

Both MBR-PE(align) and MBR-PE(consen) outperform baseline system (Base-PE) in Table 1 and MBR-PE$_{\text{-SimFeats}}$ in Table 5. From Table 7 we can see that consensus-based features contribute more than alignment-based features do. However, the best performances are achieved when using both of these feature categories at the same time.

4.6 Results on the Second Data Set

Last, we evaluate our MBR-based phrase extraction approach on the second data set.

For the alignments predicted by Disc-Aligner, we use the best parameter setting determined on the first data set where pruning threshold is set to be 10^{-5} and 15-best alignments are used to extract bilingual phrases. We also want to see whether simply cutting off phrase pairs with low frequencies could bring improvements, instead of using our more complicated MBR scoring and alignment pruning. We use *Base-PE(cutoff=2)* to denote phrase tables generated by using Base-PE with all phrase pairs that occur twice or less are pruned. Evaluation results are shown in Table 8.

	IBM-BLEU%		
	MT03	MT05	MT08
Base-PE$^{1\text{-best}}$	36.13	34.19	22.50
Base-PE$^{1\text{-best}}$ (cutoff=2)	35.81	33.96	22.24
Base-PE$^{15\text{-best}}$	36.54	34.50	22.94
Base-PE$^{15\text{-best}}$ (cutoff=2)	36.21	34.20	22.67
MBR-PE$^{1\text{-best}}$	**36.97***	**34.80***	**23.26***
MBR-PE$^{15\text{-best}}$	**37.21***	**35.07***	**23.85***

Table 8: MBR-PE vs. Base-PE on the second data set based on Disc-Aligner (*: significantly better than the results of Base-PE$^{1\text{-best}}$ with $p < 0.05$)

Compared to Base-PE$^{1\text{-best}}$ and Base-PE$^{15\text{-best}}$, MBR-PE$^{1\text{-best}}$ and MBR-PE$^{15\text{-best}}$ achieve significant improvements on both dev and test data sets, which solidly demonstrate the effectiveness of our approach again. When enlarging the cutoff threshold to be 2, results of Base-PE become worse than the ones using default cutoff size 1. It shows that the gains achieved in our approach cannot be obtained by simply discarding those low-frequency phrase pairs directly.

For the alignments predicted by GIZA++, evaluation results are listed in Table 9, where MBR-PE still obtains significant improvements by using 1-best word alignment only for phrase extraction. We also include the results of MBR decoding on two systems for comparison. Although the gains of MBR decoding obtained from MBR-PE are smaller than the ones obtained from Base-PE, they still perform best among all settings. These conclusions are consistent with the ones induced from Table 1 and 3.

	IBM-BLEU%		
	MT03	MT05	MT08
Base-PE$^{1\text{-best}}$	37.05	35.09	23.33
Base-PE$^{1\text{-best}}$ (+ MBR Decoding)	37.78	35.50	23.71
MBR-PE$^{1\text{-best}}$	**37.95***	**35.87***	**24.06***
MBR-PE$^{1\text{-best}}$ (+ MBR Decoding)	**38.31***	**36.02***	**24.23***

Table 9: MBR-PE vs. Base-PE on the second data set based on GIZA++ (*: significantly better than the results of Base-PE$^{1\text{-best}}$ with $p < 0.05$)

5 Conclusions

We have presented a novel MBR-based phrase extraction pipeline for SMT training. Under this pipeline, the quality of phrase pairs are measured by their internal similarities, and phrase pairs with low MBR model scores are pruned based on an alignment pruning strategy. One can put as many features as possible into this MBR framework to compute such similarity scores. To the best of our knowledge, it is the first attempt to improve the accuracy of phrase pairs by leveraging the MBR principle.

One future work to do is to investigate more features to measure similarities between phrase pairs. The other is to compare our approach with both discriminative learning-based methods and significance test-based methods for SMT phrase extraction. We expect to adapt our approach to other rule extraction procedures as well.

References

David Chiang. 2005. *A Hierarchical Phrase-based Model for Statistical Machine Translation*. In *Proc. ACL*, pages 263-270.

Yonggang Deng, Jia Xu, and Yuqing Gao. 2008. *Phrase Table Training for Precision and Recall: What Makes a Good Phrase and a Good Phrase Pair?* In *Proc. ACL*, pages 81-88.

Christopher Dyer, Smaranda Muresan, and Philip Resnik. 2008. *Generalizing Word Lattice Translation*. In *Proc. ACL*, pages 1012-1020.

Alexander Fraser and Daniel Marcu. 2006. *Semi-Supervised Training for Statistical Word Alignment.* In *Proc. ACL*, pages 769-776.

Alexander Fraser and Daniel Marcu. 2007. *Measuring word alignment quality for Statistical Machine Translation. Computational Linguistics*, 33(3): 293-303.

Fei Huang and Bing Xiang. 2010. *Feature-Rich Discriminative Phrase Rescoring for SMT*. In *Proc. COLING*, pages 492-500.

Aria Haghighi, John Blitzer, John DeNero, and Dan Klein. 2009. *Better Word Alignments with Supervised ITG Models*. In *Proc. ACL-IJCNLP*, pages 923-931.

Zhongjun He, Yao Meng, Yajuan Lv, Hao Yu and Qun Liu. 2009. *Reducing SMT Rule Table with Monolingual Key Phrase*. In *Proc. ACL-IJCNLP*, pages 121-124.

Howard Johnson, Joel Martin, George Foster, and Roland Kuhn. 2007. *Improving Translation Quality by Discarding Most of the Phrase Table*. In *Proc. EMNLP-CoNLL*, pages 967-975.

Philipp Koehn. 2004a. *Phrase-based Model for SMT. Computational Linguistics*, 28(1): 114-133.

Philipp Koehn. 2004b. *Statistical Significance Tests for Machine Translation Evaluation*. In *Proc. EMNLP*, pages 388-395.

Shankar Kumar and William Byrne. 2002. *Minimum Bayes-Risk Word Alignment of Bilingual Texts*. In *Proc. EMNLP*, pages 140-147.

Shankar Kumar and William Byrne. 2004. *Minimum Bayes-Risk Decoding for Statistical Machine Translation*. In *Proc. 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 *Proc. ACL*, pages 163-171.

Adam Lopez and Philip Resnik. 2006. *Word-based Alignment, Phrase-based Translation: What's the link*. In *Proc. AMTA*, pages 90-99.

Yang Liu, Tian Xia, Xinyan Xiao, and Qun Liu. 2009. *Weighted Alignment Matrices for Statistical Machine Translation*. In *Proc. EMNLP*, pages 1017-1026.

Haitao Mi, Liang Huang, and Qun Liu. 2008. *Forest-based Translation. In Proc. ACL*, pages 192-199.

Robert Moore. 2004. *On Log-Likelihood-Ratios and the Significance of Rare Events. In Proc. EMNLP*, pages 333-340.

Franz Och. 1999. *An Efficient Method for Determining Bilingual Word Classes*. In *Proc. EACL*, pages 160-167.

Franz Och. 2003. *Minimum Error Rate Training in Statistical Machine Translation*. In *Proc. ACL*, pages 160-167.

Franz Och and Hermann Ney. 2004. *The Alignment template approach to Statistical Machine Translation. Computational Linguistics*, 30(4): 417-449.

Kishore Papineni, Salim Roukos, Todd Ward, and Weijing Zhu. 2002. *BLEU: a method for automatic evaluation of machine translation*. In *Proc. ACL*, pages 311-318.

Adwait Ratnaparkhi. 1996. *A Maximum Entropy Model for Part-Of-Speech Tagging*, In *Proc. ACL,* pages 133-142.

Roy Tromble, Shankar Kumar, Franz Och, and Wolfgang Macherey. 2008. *Lattice Minimum Bayes-Risk Decoding for Statistical Machine Translation*. In *Proc. EMNLP*, pages 620-629.

Nicola Ueffing and Hermann Ney. 2007. *Word-level Confidence Estimation for Machine Translation. Computational Linguistics*, 33(1): 9-40.

Ashish Venugopal, Andreas Zollmann, Noah Smith, and Stephan Vogel. 2008. *Wider Pipelines: N-best Alignments and Parses in MT Training*. In *Proc. AMTA*, pages 192-201.

Deyi Xiong, Qun Liu, and Shouxun Lin. 2006. *Maximum Entropy based Phrase Reordering Model for Statistical Machine Translation*. In *Proc. ACL*, pages 521-528.

Mei Yang and Jing Zheng. 2009. *Toward Smaller, Faster, and Better Hierarchical Phrase-based SMT*. In *Proc. ACL-IJCNLP*, pages 237-240.

Bing Zhao, Stephan Vogel, and Alex Waibel. 2004. *Phrase Pair Rescoring with Term Weighting for Statistical Machine Translation*. In *Proc. EMNLP*, pages 206-213.