Maximum Expected BLEU Training of Phrase and Lexicon Translation Models

Xiaodong He
Microsoft Research
One Microsoft Way, Redmond, WA, USA
xiaohe@microsoft.com

Li Deng
Microsoft Research
One Microsoft Way, Redmond, WA, USA
deng@microsoft.com

Abstract

This paper proposes a new discriminative training method in constructing phrase and lexicon translation models. In order to reliably learn a myriad of parameters in these models, we propose an expected BLEU score-based utility function with KL regularization as the objective, and train the models on a large parallel dataset. For training, we derive growth transformations for phrase and lexicon translation probabilities to iteratively improve the objective. The proposed method, evaluated on the Europarl German-to-English dataset, leads to a 1.1 BLEU point improvement over a state-of-the-art baseline translation system. In IWSLT 2011 Benchmark, our system using the proposed method achieves the best Chinese-to-English translation result on the task of translating TED talks.

1. Introduction

Discriminative training is an active area in statistical machine translation (SMT) (e.g., Och et al., 2002, 2003, Liang et al., 2006, Blunsom et al., 2008, Chiang et al., 2009, Foster et al, 2010, Xiao et al. 2011). Och (2003) proposed using a log-linear model to incorporate multiple features for translation, and proposed a minimum error rate training (MERT) method to train the feature weights to optimize a desirable translation metric.

While the log-linear model itself is discriminative, the phrase and lexicon translation features, which are among the most important components of SMT, are derived from either generative models or heuristics (Koehn et al., 2003, Brown et al., 1993). Moreover, the parameters in the phrase and lexicon translation models are estimated by relative frequency or maximizing joint likelihood, which may not correspond closely to the translation measure, e.g., bilingual evaluation understudy (BLEU) (Papineni et al., 2002). Therefore, it is desirable to train all these parameters to directly maximize an objective that directly links to translation quality.

However, there are a large number of parameters in these models, making discriminative training for them non-trivial (e.g., Liang et al., 2006, Chiang et al., 2009). Liang et al. (2006) proposed a large set of lexical and Part-of-Speech features and trained the model weights associated with these features using perceptron. Since many of the reference translations are non-reachable, an empirical *local updating* strategy had to be devised to fix this problem by picking a *pseudo* reference. Many such non-desirable heuristics led to moderate gains reported in that work. Chiang et al. (2009) improved a syntactic SMT system by adding as many as ten thousand syntactic features, and used Margin Infused Relaxed Algorithm (MIRA) to train the feature weights. However, the number of parameters in common phrase and lexicon translation models is much larger.

In this work, we present a new, highly effective discriminative learning method for phrase and lexicon translation models. The training objective is an expected BLEU score, which is closely linked to translation quality. Further, we apply a Kullback–Leibler (KL) divergence regularization to prevent over-fitting.

For effective optimization, we derive updating formulas of growth transformation (GT) for phrase and lexicon translation probabilities. A GT is a transformation of the probabilities that guarantees strict non-decrease of the objective over each GT iteration unless a local maximum is reached. A

similar GT technique has been successfully used in speech recognition (Gopalakrishnan et al., 1991, Povey, 2004, He et al., 2008). Our work demonstrates that it works with large scale discriminative training of SMT model as well.

Our work is based on a phrase-based SMT system. Experiments on the Europarl German-to-English dataset show that the proposed method leads to a 1.1 BLEU point improvement over a strong baseline. The proposed method is also successfully evaluated on the IWSLT 2011 benchmark test set, where the task is to translate TED talks (www.ted.com). Our experimental results on this open-domain spoken language translation task show that the proposed method leads to significant translation performance improvement over a state-of-the-art baseline, and the system using the proposed method achieved the best single system translation result in the Chinese-to-English MT track.

2. Related Work

One best known approach in discriminative training for SMT is proposed by Och (2003). In that work, multiple features, most of them are derived from generative models, are incorporated into a log-linear model, and the relative weights of them are tuned discriminatively on a small tuning set. However, in practice, this approach only works with a handful of parameters.

More closely related to our work, Liang et al. (2006) proposed a large set of lexical and Part-of-Speech features in addition to the phrase translation model. Weights of these features are trained using perceptron on a training set of 67K sentences. In that paper, the authors pointed out that forcing the model to update towards the reference translation could be problematic. This is because the hidden structure such as phrase segmentation and alignment could be abused if the system is forced to produce a reference translation. Therefore, instead of pushing the parameter update towards the reference translation (a.k.a. *bold updating*), the author proposed a *local updating* strategy where the model parameters are updated towards a pseudo-reference (i.e., the hypothesis in the n-best list that gives the best BLEU score). Experimental results showed that their approach outperformed a baseline by 0.8 BLEU point when using monotonic decoding, but there was no significant gain over a stronger baseline with a full-distortion model. In our work, we use the expectation of BLEU scores as the objective. This avoids the heuristics of picking the updating reference and therefore gives a more principal way of setting the training objective.

As another closely related study, Chiang et al. (2009) incorporated about ten thousand syntactic features in addition to the baseline features. The feature weights are trained on a tuning set with 2010 sentences using MIRA. In our work, we have many more parameters to train, and the training is conducted on the entire training corpora. Our GT based optimization algorithm is highly parallelizable and efficient, which is the key for large scale discriminative training.

As a further related work, Rosti et al. (2011) have proposed using differentiable expected BLEU score as the objective to train system combination parameters. Other work related to the computation of expected BLEU in common with ours includes minimum Bayes risk approaches (Smith and Eisner 2006, Tromble et al., 2008) and lattice-based MERT (Macherey et al., 2008). In these earlier work, however, the phrase and lexicon translation models used remained unchanged.

Another line of research that is closely related to our work is phrase table refinement and pruning. Wuebker et al. (2010) proposed a method to train the phrase translation model using Expectation-Maximization algorithm with a *leave-one-out* strategy. The parallel sentences were forced to be aligned at the phrase level using the phrase table and other features as in a decoding process. Then the phrase translation probabilities were estimated based on the phrase alignments. To prevent overfitting, the statistics of phrase pairs from a particular sentence was excluded from the phrase table when aligning that sentence. However, as pointed out by Liang et al (2006), the same problem as in the *bold updating* existed, i.e., forced alignment between a source sentence and its reference translation was tricky, and the proposed alignment was likely to be unreliable. The method presented in this paper is free from this problem.

3. Phrase-based Translation System

The translation process of phrase-based SMT can be briefly described in three steps: segment source sentence into a sequence of phrases, translate each

source phrase to a target phrase, re-order target phrases into target sentence (Koehn et al., 2003).

In decoding, the optimal translation $\hat{E}$ given the source sentence F is obtained according to

$$\hat{E} = \underset{E}{\operatorname{argmax}} P(E|F) \tag{1}$$

where

$$P(E|F) = \frac{1}{Z} exp\left\{\sum_m \lambda_m \log h_m(E,F)\right\} \tag{2}$$

and $Z = \sum_E exp\{\sum_m \lambda_m \log h_m(E,F)\}$ is the normalization denominator to ensure that the probabilities sum to one. Note that we define the feature functions $\{h_m(E,F)\}$ in log domain to simplify the notation in later sections. Feature weights $\boldsymbol{\lambda} = \{\lambda_m\}$ are usually tuned by MERT.

Features used in a phrase-based system usually include LM, reordering model, word and phrase counts, and phrase and lexicon translation models. Given the focus of this paper, we review only the phrase and lexicon translation models below.

3.1. Phrase translation model

A set of phrase pairs are extracted from word-aligned parallel corpus according to phrase extraction rules (Koehn et al., 2003). Phrase translation probabilities are then computed as relative frequencies of phrases over the training dataset. i.e., the probability of translating a source phrase $\tilde{f}$ to a target phrase $\tilde{e}$ is computed by

$$p(\tilde{e}|\tilde{f}) = \frac{C(\tilde{e},\tilde{f})}{C(\tilde{f})} \tag{3}$$

where $C(\tilde{e},\tilde{f})$ is the joint counts of $\tilde{e}$ and $\tilde{f}$, and $C(\tilde{f})$ is the marginal counts of $\tilde{f}$.

In translation, the input sentence is segmented into K phrases, and the source-to-target forward phrase (FP) translation feature is scored as:

$$h_{FP}(E,F) = \prod_k p(\tilde{e}_k|\tilde{f}_k) \tag{4}$$

where $\tilde{e}_k$ and $\tilde{f}_k$ are the k-th phrase in E and F, respectively. The target-to-source (backward) phrase translation model is defined similarly.

3.2. Lexicon translation model

There are several variations in lexicon translation features (Ayan and Dorr 2006, Koehn et al., 2003, Quirk et al., 2005). We use the word translation table from IBM Model 1 (Brown et al., 1993) and compute the sum over all possible word alignments within a phrase pair without normalizing for length (Quirk et al., 2005). The source-to-target forward lexicon (FL) translation feature is:

$$h_{FL}(E,F) = \prod_k \prod_m \sum_r p(e_{k,m}|f_{k,r}) \tag{5}$$

where $e_{k,m}$ is the m-th word of the k-th target phrase $\tilde{e}_k$, $f_{k,r}$ is the r-th word in the k-th source phrase $\tilde{f}_k$, and $p(e_{k,m}|f_{k,r})$ is the probability of translating word $f_{k,r}$ to word $e_{k,m}$. In IBM model 1, these probabilities are learned via maximizing a joint likelihood between the source and target sentences. The target-to-source (backward) lexicon translation model is defined similarly.

4. Maximum Expected-BLEU Training

4.1. Objective function

We denote by $\boldsymbol{\theta}$ the set of all the parameters to be optimized, including forward phrase and lexicon translation probabilities and their backward counterparts. For simplification of notation, $\boldsymbol{\theta}$ is formed as a matrix, where its elements $\{\theta_{ij}\}$ are probabilities subject to $\sum_j \theta_{ij} = 1$. E.g., each row is a probability distribution.

The utility function over the entire training set is defined as:

$$U(\boldsymbol{\theta}) = \sum_{E_1,\dots,E_N} P_{\boldsymbol{\theta}}(E_1,\dots,E_N|F_1,\dots,F_N)\left(\sum_{n=1}^{N} BLEU(E_n,E_n^*)\right) \tag{6}$$

where N is the number of sentences in the training set, E_n^* is the reference translation of the n-th source sentence F_n, and $E_n \in Hyp(F_n)$ that denotes the list of translation hypotheses of F_n. Since the sentences are independent with each other, the joint posterior can be decomposed:

$$P_{\boldsymbol{\theta}}(E_1,\dots,E_N|F_1,\dots,F_N) = \prod_{n=1}^{N} P_{\boldsymbol{\theta}}(E_n|F_n) \tag{7}$$

and $P_{\boldsymbol{\theta}}(E_n|F_n)$ is the posterior defined in (2), the subscript $\boldsymbol{\theta}$ indicates that it is computed based on the parameter set $\boldsymbol{\theta}$. $U(\boldsymbol{\theta})$ is proportional (with a factor of *N*) to the expected sentence BLEU score over the entire training set, i.e., after some algebra,

$$U(\boldsymbol{\theta}) = \sum_{n=1}^{N} \sum_{E_n} P_{\boldsymbol{\theta}}(E_n|F_n) BLEU(E_n, E_n^*)$$

In a phrase-based SMT system, the total number of parameters of phrase and lexicon translation models, which we aim to learn discriminatively, is very large (see Table 1). Therefore, regularization is critical to prevent over-fitting. In this work, we regularize the parameters with KL regularization.

KL divergence is commonly used to measure the distance between two probability distributions. For the whole parameter set $\boldsymbol{\theta}$, the KL regularization is defined in this work as the sum of KL divergence over the entire parameter space:

$$KL(\boldsymbol{\theta}^0||\boldsymbol{\theta}) = \sum_i \sum_j \theta_{ij}^0 \log \frac{\theta_{ij}^0}{\theta_{ij}} \quad (8)$$

where $\boldsymbol{\theta}^0$ is a constant prior parameter set. In training, we want to improve the utility function while keeping the changes of the parameters from $\boldsymbol{\theta}^0$ at minimum. Therefore, we design the objective function to be maximized as:

$$O(\boldsymbol{\theta}) = \log U(\boldsymbol{\theta}) - \tau \cdot KL(\boldsymbol{\theta}^0||\boldsymbol{\theta}) \quad (9)$$

where the prior model $\boldsymbol{\theta}^0$ in our approach is the relative-frequency-based phrase translation model and the maximum-likelihood-estimated IBM model 1 (word translation model). τ is a hyper-parameter controlling the degree of regularization.

4.2. Optimization

In this section, we derived GT formulas for iteratively updating the parameters so as to optimize objective (9). GT is based on extended Baum-Welch (EBW) algorithm first proposed by Gopalakrishnan et al. (1991) and commonly used in speech recognition (e.g., He et al. 2008).

4.2.1. Extended Baum-Welch Algorithm

Baum-Eagon inequality (Baum and Eagon, 1967) gives the GT formula to iteratively maximize positive-coefficient polynomials of random variables that are subject to sum-to-one constants. Baum-Welch algorithm is a model update algorithm for hidden Markov model which uses this GT. Gopalakrishnan et al. (1991) extended the algorithm to handle rational function, i.e., a ratio of two polynomials, which is more commonly encountered in discriminative training.

Here we briefly review EBW. Assuming a set of random variables $\mathbf{p} = \{p_{ij}\}$ that subject to the constraint that $\sum_j p_{ij} = 1$, and assume $g(\mathbf{p})$ and $h(\mathbf{p})$ are two positive polynomial functions of $\mathbf{p}$, a GT of $\mathbf{p}$ for the rational function $r(\mathbf{p}) = \frac{g(\mathbf{p})}{h(\mathbf{p})}$ can be obtained through the following two steps:

i) Construct the auxiliary function:

$$f(\mathbf{p}) = g(\mathbf{p}) - r(\mathbf{p}')h(\mathbf{p}) \quad (10)$$

where $\mathbf{p}'$ are the values from the previous iteration. Increasing f guarantees an increase of r, i.e., $h(\mathbf{p}) > 0$ and $r(\mathbf{p}) - r(\mathbf{p}') = \frac{1}{h(\mathbf{p})}\big(f(\mathbf{p}) - f(\mathbf{p}')\big)$.

ii) Derive GT formula for $f(\mathbf{p})$

$$p_{ij} = \frac{p'_{ij} \frac{\partial f(\mathbf{p})}{\partial p_{ij}}\Big|_{\mathbf{p}=\mathbf{p}'} + D \cdot p'_{ij}}{\sum_j p'_{ij} \frac{\partial f(\mathbf{p})}{\partial p_{ij}}\Big|_{\mathbf{p}=\mathbf{p}'} + D} \quad (11)$$

where *D* is a smoothing factor.

4.2.2. GT of Translation Models

Now we derive the GTs of translation models for our objective. Since maximizing $O(\boldsymbol{\theta})$ is equivalent to maximizing $e^{O(\boldsymbol{\theta})}$, we have the following auxiliary function:

$$R(\boldsymbol{\theta}) = U(\boldsymbol{\theta}) e^{-\tau \cdot KL(\boldsymbol{\theta}^0||\boldsymbol{\theta})} \quad (12)$$

After substituting (2) and (7) into (6), and drop optimization irrelevant terms in KL regularization, we have $R(\boldsymbol{\theta})$ in a rational function form:

$$R(\boldsymbol{\theta}) = \frac{G(\boldsymbol{\theta}) \cdot J(\boldsymbol{\theta})}{H(\boldsymbol{\theta})} \quad (13)$$

where $H(\boldsymbol{\theta}) = \sum_{E_1,\ldots,E_N} \prod_{n=1}^{N} \prod_m h_m^{\lambda_m}(E_n, F_n)$, $J(\boldsymbol{\theta}) = \prod_i \prod_j \theta_{ij}^{\tau\theta_{ij}^0}$, and $G(\boldsymbol{\theta}) =$

$\sum_{E_1,\ldots,E_N} \prod_{n=1}^{N} \prod_m h_m^{\lambda_m}(E_n, F_n) \left(\sum_{n=1}^{N} BLEU(E_n, E_n^*)\right)$ are all positive polynomials of $\boldsymbol{\theta}$. Therefore, we can follow the two steps of EBW to derive the GT formulas for $\boldsymbol{\theta}$.

If we denote by p_{ij} the probability of translating the source phrase *i* to the target phrase *j*. Then, the updating formula is (derivation omitted):

$$p_{ij} = \frac{\sum_n \sum_{E_n} \gamma_{FP}(E_n, n, i, j) + U(\boldsymbol{\theta}')\tau_{FP} p_{ij}^0 + D_i p'_{ij}}{\sum_n \sum_{E_n} \sum_j \gamma_{FP}(E_n, n, i, j) + U(\boldsymbol{\theta}')\tau_{FP} + D_i} \quad (14)$$

where $\tau_{FP} = \tau/\lambda_{FP}$ and $\gamma_{FP}(E_n, n, i, j) = P_{\boldsymbol{\theta}'}(E_n|F_n) \cdot [BLEU(E_n, E_n^*) - U_n(\boldsymbol{\theta}')] \cdot \sum_k \mathbf{1}(\tilde{f}_{n,k} = i, \tilde{e}_{n,k} = j)$. In which $U_n(\boldsymbol{\theta}')$ takes a form similar to (6), but is the expected BLEU score for sentence *n* using models from the previous iteration. $\tilde{f}_{n,k}$ and $\tilde{e}_{n,k}$ are the *k*-th phrases of F_n and E_n, respectively.

The smoothing factor set of D_i according to the Baum-Eagon inequality is usually far too large for practical use. In practice, one general guide of setting D_i is to make all updated value positive. Similar to (Povey 2004), we set D_i by

$$D_i = \sum_n \sum_{E_n} \sum_j \max(0, -\gamma_{FP}(E_n, n, i, j)) \quad (15)$$

to ensure the denominator of (15) is positive. Further, we set a low-bound of D_i as $\max_j\{\frac{-\sum_n \sum_{E_n} \gamma_{FP}(E_n, n, i, j)}{p'_{ij}}\}$ to guarantee the numerator to be positive.

We denote by l_{ij} the probability of translating the source word *i* to the target word *j*. Then following the same derivation, we get the updating formula for forward lexicon translation model:

$$l_{ij} = \frac{\sum_n \sum_{E_n} \gamma_{FL}(E_n, n, i, j) + U(\boldsymbol{\theta}')\tau_{FL} l_{ij}^0 + D_i l'_{ij}}{\sum_n \sum_{E_n} \sum_j \gamma_{FL}(E_n, F_n, i, j) + U(\boldsymbol{\theta}')\tau_{FL} + D_i} \quad (16)$$

where $\tau_{FL} = \tau/\lambda_{FL}$ and $\gamma_{FL}(E_n, n, i, j) = P_{\boldsymbol{\theta}'}(E_n|F_n) \cdot [BLEU(E_n, E_n^*) - U_n(\boldsymbol{\theta}')] \cdot \sum_k \sum_m \mathbf{1}(e_{n,k,m} = j)\gamma(n, k, m, i)$, and $\gamma(n, k, m, i) = \frac{\sum_r \mathbf{1}(f_{n,k,r}=i) p_{\boldsymbol{\theta}'}(e_{n,k,m}|f_{n,k,r})}{\sum_r p_{\boldsymbol{\theta}'}(e_{n,k,m}|f_{n,k,r})}$, in which $f_{n,k,r}$ and $e_{n,k,m}$ are the *r*-th and *m*-th word in the *k*-th phrase of the source sentence F_n and the target hypothesis E_n, respectively. Value of D_i is set in a way similar to (15).

GTs for updating backward phrase and lexicon translation models can be derived in a similar way, and is omitted here.

4.3. Implementation issues

4.3.1. Normalizing $\boldsymbol{\lambda}$

The posterior $p_{\boldsymbol{\theta}'}(E_n|F_n)$ in the model updating formula is computed according to (2). In decoding, only the relative values of $\boldsymbol{\lambda}$ matters. However, the absolute value will affect the posterior distribution, e.g., an overly large absolute value of $\boldsymbol{\lambda}$ would lead to a very sharp posterior distribution. In order to control the sharpness of the posterior distribution, we normalize $\boldsymbol{\lambda}$ by its L1 norm:

$$\hat{\lambda}_m = \frac{\lambda_m}{\sum_m |\lambda_m|} \quad (17)$$

4.3.2. Computing the sentence BLEU sore

The commonly used BLEU-4 score is computed by

$$BLEU\text{-}4 = \mathrm{BP} \cdot \exp\left(\frac{1}{4}\sum_{n=1}^{4} \log p_n\right) \quad (18)$$

In the updating formula, we need to compute the sentence-level $BLEU(E_n, E_n^*)$. Since the matching count may be sparse at the sentence level, we smooth raw precisions of high-order *n*-grams by:

$$p_n = \frac{\#(n\text{-}gram\ matched) + \eta \cdot p_n^0}{\#(n\text{-}gram) + \eta} \quad (19)$$

where p_n^0 is the prior value of p_n, η is a smoothing factor usually takes a value of 5 and p_n^0 can be set by $p_n^0 = p_{n-1} \cdot p_{n-1}/p_{n-2}$, for n = 3, 4. p_1 and p_2 are estimated empirically. Brevity penalty (BP) also plays a key role. Instead of clip it at 1, we use a non-clipped BP, $BP = e^{(1-\frac{r}{c})}$, for sentence-level BLEU[1]. We further scale the reference length, *r*, by a factor such that the total length of references on the training set equals that of the baseline output[2].

[1] This is to better approximate corpus-level BLEU, i.e., as discussed in (Chiang, et al., 2008), the per-sentence BP might effectively exceed unity in corpus-level BLEU computation.

[2] This is to focus the training on improving BLEU by improving n-gram match instead of by improving BP, e.g., this makes the BP of the baseline output already being perfect.

4.3.3. Training procedure

The parameter set $\boldsymbol{\theta}$ is optimized on the training set while the feature weights $\boldsymbol{\lambda}$ are tuned on a small tuning set[3]. Since $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ affect the training of each other, we train them in alternation. I.e., at each iteration, we first fix $\boldsymbol{\lambda}$ and update $\boldsymbol{\theta}$, then we re-tune $\boldsymbol{\lambda}$ given the new $\boldsymbol{\theta}$. Due to mismatch between training and tuning data, the training process might not always converge. Therefore, we need a validation set to determine the stop point of training. At the end, $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ that give the best score on the validation set are selected and applied to the test set. Fig. 1 gives a summary of the training procedure. Note that step 2 and 4 are parallelize-able across multiple processors.

1. Build the baseline system, estimate { $\boldsymbol{\theta}$, $\boldsymbol{\lambda}$ }.
2. Decode N-best list for training corpus using the baseline system, compute $BLEU(E_n, E_n^*)$.
3. set $\boldsymbol{\theta}' = \boldsymbol{\theta}$, $\boldsymbol{\lambda}' = \boldsymbol{\lambda}$.
4. Max expected BLEU training
   a. Go through the training set.
      i. Compute $P_{\boldsymbol{\theta}'}(E_n|F_n)$ and $U_n(\boldsymbol{\theta}')$.
      ii. Accumulate statistics $\{\gamma\}$.
   b. Update: $\boldsymbol{\theta}' \rightarrow \boldsymbol{\theta}$ by one iteration of GT.
5. MERT on the tuning set: $\boldsymbol{\lambda}' \rightarrow \boldsymbol{\lambda}$.
6. Test on the validation set using { $\boldsymbol{\theta}$, $\boldsymbol{\lambda}$ }.
7. Go to step 3 unless training converges or reaches a certain number of iterations.
8. Pick the best { $\boldsymbol{\theta}$, $\boldsymbol{\lambda}$ } on the validation set.

Figure 1. The max expected-BLEU training algorithm.

5. Evaluation

In evaluating the proposed method, we use two separate datasets. We first describe the experiments with the *Europarl* dataset (Koehn 2002), followed by the experiments with the more recent IWSLT-2011 task (Federico et al., 2011).

5.1 Experimental setup in the Europarl task

In evaluating the proposed method, we use two separate datasets. First, we conduct experiments on the Europarl German-to-English dataset. The training corpus contains 751K sentence pairs, 21 words per sentence on average. 2000 sentences are provided in the development set. We use the first 1000 sentences for $\boldsymbol{\lambda}$ tuning, and the rest for validation. The test set consists of 2000 sentences.

[3] Usually, the tuning set matches the test condition better, and therefore is preferable for $\boldsymbol{\lambda}$ tuning.

To build the baseline phrase-based SMT system, we first perform word alignment on the training set using a hidden Markov model with lexicalized distortion (He 2007), then extract the phrase table from the word aligned bilingual texts (Koehn et al., 2003). The maximum phrase length is set to four. Other models used in the baseline system include lexicalized ordering model, word count and phrase count, and a 3-gram LM trained on the English side of the parallel training corpus. Feature weights are tuned by MERT. A fast beam-search phrase-based decoder (Moore and Quirk 2007) is used and the distortion limit is set to four. Details of the phrase and lexicon translation models are given in Table 1. This baseline achieves a BLEU score of 26.22% on the test set. This baseline system is also used to generate a 100-best list of the training corpus during maximum expected BLEU training.

Translation model	# parameters
Phrase models (fore. & back.)	9.2 M
Lexicon model (IBM-1 src-to-tgt)	12.9 M
Lexicon model (IBM-1 tgt-to-src)	11.9 M

Table 1. Summary of phrase and lexicon translation models

5.2 Experimental results on the Europarl task

During training, we first tune the regularization factor τ based on the performance on the validation set. For simplicity reasons, the tuning of τ makes use of only the phrase translation models. Table 2 reports the BLEU scores and gains over the baseline given different values of τ. The results highlight the importance of regularization. While $\tau = 5 \times 10^{-5}$ gives the best score on the validation set, the gain is shown to be substantially reduced to merely 0.2 BLEU point when $\tau = 0$, i.e., no regularization. We set the optimal value of $\tau = 5 \times 10^{-5}$ in all remaining experiments.

Test on Validation Set	$BLEU\%$	$\Delta BLEU\%$
Baseline	26.70	--
$\tau = 0$ (no regularization)	26.91	+0.21
$\tau = 1 \times 10^{-5}$	27.31	+0.61
$\tau = 5 \times 10^{-5}$	27.44	+0.74
$\tau = 10 \times 10^{-5}$	27.27	+0.57

Table 2. Results on degrees of regularizations. BLEU scores are reported on the validation set. $\Delta BLEU$ denotes the gain over the baseline.

Fixing the optimal regularization factor τ, we then study the relationship between the expected

sentence-level BLEU (Exp. BLEU) score of N-best lists and the corpus-level BLEU score of 1-best translations. The conjectured close relationship between the two is important in justifying our use of the former as the training objective. Fig. 2 shows these two scores on the training set over training iterations. Since the expected BLEU is affected by λ strongly, we fix the value of λ in order to make the expected BLEU comparable across different iterations. From Fig. 2 it is clear that the expected BLEU score correlates strongly with the real BLEU score, justifying its use as our training objective.

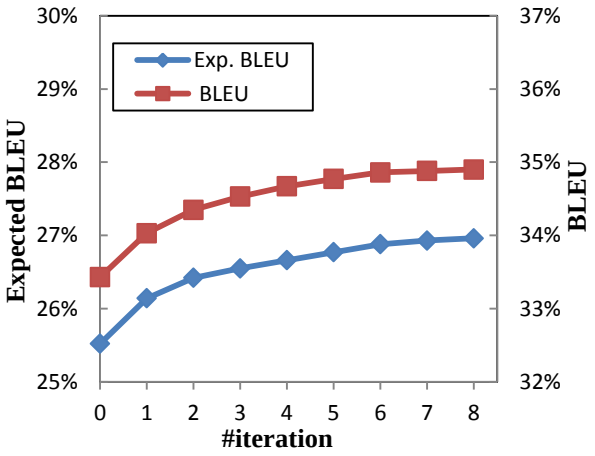

Figure 2. Expected sentence BLEU and 1-best corpus BLEU on the 751K sentence of training data.

Next, we study the effects of training the phrase translation probabilities and the lexicon translation probabilities according to the GT formulas presented in the preceding section. The break-down results are shown in Table 3. Compared with the baseline, training phrase or lexicon models alone gives a gain of 0.7 and 0.5 BLEU points, respectively, on the test set. For a full training of both phrase and lexicon models, we adopt two learning schedules: update both models together at each iteration (*simultaneously*), or update them in two stages (*two-stage*), where the phrase models are trained first until reaching the best score on the validation set and then the lexicon models are trained. Both learning schedules give significant improvements over the baseline and also over training phrase or lexicon models alone. The *two-stage* training of both models gives the best result of 27.33%, outperforming the baseline by 1.1 BLEU points.

More detail of the two-stage training is provided in Fig. 3, where BLEU scores in each stage are shown as a function of the GT training iteration. The phrase translation probabilities (PT) are trained alone in the first stage, shown in blue color. After five iterations, the BLEU score on the validation set reaches the peak value, with further iteration giving BLEU score fluctuation. Hence, we perform lexicon model (LEX) training starting from the sixth iteration with the corresponding BLEU scores shown in red color in Fig. 3. The BLEU score is further improved by 0.4 points after additional three iterations of training the lexicon models. In total, nine iterations are performed to complete the two-stage GT training of all phrase and lexicon models.

BLEU (%)	validation	test
Baseline	26.70	26.22
Train phrase models alone	27.44	26.94*
Train lexicon models alone	27.36	26.71
Both models: *simultaneously*	27.65	27.13*
Both models: *two-stage*	27.82	27.33*

Table 3. Results on the Europarl German-to-English dataset. The BLEU measures from various settings of maximum expected BLEU training are compared with the baseline, where * denotes that the gain over the baseline is statistically significant with a significance level > 99%, measured by paired bootstrap resampling method proposed by Koehn (2004).

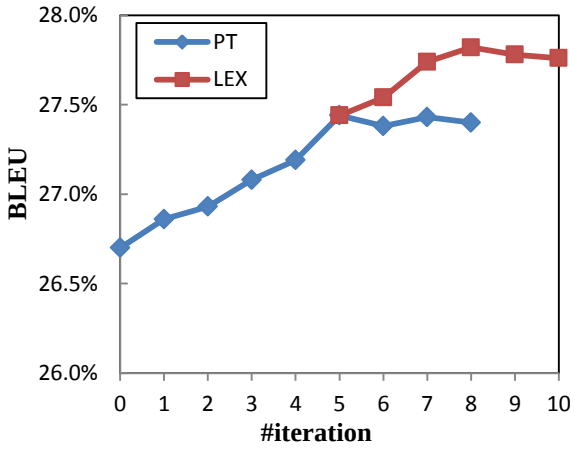

Figure 3. BLEU scores on the validation set as a function of the GT training iteration in two-stage training of both the phrase translation models (PT) and the lexicon models (LEX). The BLEU scores on training phrase models are shown in blue, and on training lexicon models in red.

5.3 Experiments on the IWSLT2011 benchmark

As the second evaluation task, we apply our new method described in this paper to the 2011 IWSLT Chinese-to-English machine translation benchmark (Federico et al., 2011). The main focus of the IWSLT2011 Evaluation is the translation of TED talks (www.ted.com). These talks are originally given in English. In the Chinese-to-English translation task, we are provided with human translated Chinese text with punctuations inserted. The goal is to match the human transcribed English speech with punctuations.

This is an open-domain spoken language translation task. The training data consist of 110K sentences in the transcripts of the TED talks and their translations, in English and Chinese, respectively. Each sentence consists of 20 words on average. Two development sets are provided, namely, dev2010 and tst2010. They consist of 934 sentences and 1664 sentences, respectively. We use dev2010 for $\boldsymbol{\lambda}$ tuning and tst2010 for validation. The test set tst2011 consists of 1450 sentences.

In our system, a primary phrase table is trained from the 110K TED parallel training data, and a 3-gram LM is trained on the English side of the parallel data. We are also provided additional out-of-domain data for potential usage. From them, we train a secondary 5-gram LM on 115M sentences of supplementary English data, and a secondary phrase table from 500K sentences selected from the supplementary UN corpus by the method proposed by Axelrod et al. (2011).

In carrying out the maximum expected BLEU training, we use 100-best list and tune the regularization factor to the optimal value of $\tau = 1 \times 10^{-5}$. We only train the parameters of the primary phrase table. The secondary phrase table and LM are excluded from the training process since the out-of-domain phrase table is less relevant to the TED translation task, and the large LM slows down the N-best generation process significantly.

At the end, we perform one final MERT to tune the relative weights with all features including the secondary phrase table and LM.

The translation results are presented in Table 4. The baseline is a phrase-based system with all features including the secondary phrase table and LM. The new system uses the same features except that the primary phrase table is discriminatively trained using maximum expected-BLEU and GT optimization as described earlier in this paper. The results are obtained using the two-stage training schedule, including six iterations for training phrase translation models and two iterations for training lexicon translation models. The results in Table 4 show that the proposed method leads to an improvement of 1.2 BLEU point over the baseline. This gives the best single system result on this task.

BLEU (%)	Validation	Test
Baseline	11.48	14.68
Max expected BLEU training	12.39	15.92

Table 4. The translation results on IWSLT 2011 MT_CE task.

6. Summary

The contributions of this work can be summarized as follows. First, we propose a new objective function (Eq. 9) for training of large-scale translation models, including phrase and lexicon models, with more parameters than all previous methods have attempted. The objective function consists of 1) the utility function of expected BLEU score, and 2) the regularization term taking the form of KL divergence in the parameter space. The expected BLEU score is closely linked to translation quality and the regularization is essential when many parameters are trained at scale. The importance of both is verified experimentally with the results presented in this paper.

Second, through non-trivial derivation, we show that the novel objective function of Eq. (9) is amenable to iterative GT updates, where each update is equipped with a closed-form formula.

Third, the new objective function and new optimization technique are successfully applied to two important machine translation tasks, with implementation issues resolved (e.g., training schedule and hyper-parameter tuning, etc.). The superior results clearly demonstrate the effectiveness of the proposed algorithm.

Acknowledgments

The authors are grateful to Chris Quirk, Mei-Yuh Hwang, and Bowen Zhou for the assistance with the MT system and/or for the valuable discussions.

References

Amittai Axelrod, Xiaodong He, Jianfeng Gao. 2011, Domain adaptation via pseudo in-domain data selection. In Proc. of EMNLP, 2011.

Necip Fazil Ayan, and Bonnie J. Dorr. Going. 2006. Beyond AER: an extensive analysis of word alignments and their impact on MT. In Proc. of COLING-ACL, 2006.

Leonard Baum and J. A. Eagon. 1967. An inequality with applications to statistical prediction for functions of Markov processes and to a model of ecology, Bulletin of the American Mathematical Society, Jan. 1967.

Phil Blunsom, Trevor Cohn, and Miles Osborne. 2008. A discriminative latent variable model for statistical machine translation. In Proc. of ACL 2008.

Peter F. Brown, Stephen A. Della Pietra, Vincent Della J.Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics, 1993.

David Chiang, Steve DeNeefe, Yee Seng Chan, and Hwee Tou Ng, 2008. Decomposability of translation metrics for improved evaluation and efficient algorithms. In Proc. of EMNLP, 2008.

David Chiang, Kevin Knight and Weri Wang, 2009. 11,001 new features for statistical machine translation. In Proc. of NAACL-HLT, 2009.

Marcello Federico, L. Bentivogli, M. Paul, and S. Stueker. 2011. Overview of the IWSLT 2011 Evaluation Campaign. In Proc. of IWSLT, 2011.

George Foster, Cyril Goutte and Roland Kuhn. 2010. Discriminative Instance Weighting for Domain Adaptation in Statistical Machine Translation. In Proc. of EMNLP, 2010.

P. S. Gopalakrishnan, Dimitri Kanevsky, Arthur Nadas, and David Nahamoo. 1991. An inequality for rational functions with applications to some statistical estimation problems. IEEE Trans. Inform. Theory, 1991.

Xiaodong He. 2007. Using Word-Dependent Transition Models in HMM based Word Alignment for Statistical Machine Translation. In *Proc. of the Second ACL Workshop on Statistical Machine Translation*.

Xiaodong He, Li Deng, Wu Chou, 2008. Discriminative learning in sequential pattern recognition. IEEE Signal Processing Magazine, Sept. 2008.

Philipp Koehn. 2002. Europarl: A Multilingual Corpus for Evaluation of Machine Translation.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase based translation. In Proc. of NAACL. 2003.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In Proc. of EMNLP 2004.

Percy Liang, Alexandre Bouchard-Cote, Dan Klein and Ben. Taskar. 2006. An end-to-end discriminative approach to machine translation, In Proc. of COLING-ACL, 2006.

Wolfgang Macherey, Franz Josef Och, gnacio Thayer, and Jakob Uskoreit. 2008. Lattice-based minimum error rate training for statistical machine translation. In Proc. of EMNLP 2008.

Robert Moore and Chris Quirk. 2007. Faster Beam-Search Decoding for Phrasal Statistical Machine Translation. In Proc. of MT Summit XI.

Franz Josef Och and H. Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation, In Proc. of ACL 2002.

Franz Josef Och, 2003, Minimum error rate training in statistical machine translation. In Proc. of ACL 2003.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In Proc. of ACL 2002.

Daniel Povey. 2004. Discriminative Training for large Vocabulary Speech Recognition. Ph.D. dissertation, Cambridge University, Cambridge, UK, 2004.

Chris Quirk, Arul Menezes, and Colin Cherry. 2005. Dependency treelet translation: Syntactically informed phrasal SMT. In Proc. of ACL 2005.

Antti-Veikko Rosti, Bing hang, Spyros Matsoukas, and Richard Schard Schwartz. 2011. Expected BLEU training for graphs: bbn system description for WMT system combination task. In Proc. of workshop on statistical machine translation 2011.

David A Smith, Jason Eisner. 2006. Minimum risk annealing for training log-linear models, In Proc. of COLING-ACL 2006.

Joern Wuebker, Arne Mauser and Hermann Ney. 2010. Training phrase translation models with leaving-one-out, In Proc. of ACL 2010.

Roy Tromble, Shankar Kumar, Franz Och, and Wolfgang Macherey. 2008. Lattice Minimum Bayes-Risk decoding for statistical machine translation. In Proc. of EMNLP 2008.

Xinyan Xiao, Yang Liu, Qun Liu, and Shouxun Lin. 2011. Fast Generation of Translation Forest for Large-Scale SMT Discriminative Training. In Proc. Of EMNLP 2011.