Training MRF-Based Phrase Translation Models using Gradient Ascent

Jianfeng Gao
Microsoft Research
Redmond, WA, USA
jfgao@microsoft.com

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

Abstract

This paper presents a general, statistical framework for modeling phrase translation via Markov random fields. The model allows for arbituary features extracted from a phrase pair to be incorporated as evidence. The parameters of the model are estimated using a large-scale discriminative training approach that is based on stochastic gradient ascent and an N-best list based expected BLEU as the objective function. The model is easy to be incoporated into a standard phrase-based statistical machine translation system, requiring no code change in the runtime engine. Evaluation is performed on two Europarl translation tasks, German-English and French-English. Results show that incoporating the Markov random field model significantly improves the performance of a state-of-the-art phrase-based machine translation system, leading to a gain of 0.8-1.3 BLEU points.

1 Introduction

The phrase translation model, also known as the *phrase table*, is one of the core components of a phrase-based statistical machine translation (SMT) system. The most common method of constructing the phrase table takes a two-phase approach. First, the bilingual phrase pairs are extracted heuristically from an automatically word-aligned training data. The second phase is parameter estimation, where each phrase pair is assigned with some scores that are estimated based on counting of words or phrases on the same word-aligned training data.

There has been a lot of research on improving the quality of the phrase table using more principled methods for phrase extraction (e.g., Lamber and Banchs 2005), parameter estimation (e.g., Wuebker et al. 2010; He and Deng 2012), or both (e.g., Marcu and Wong 2002; Denero et al. 2006). The focus of this paper is on the parameter estimation phase. We revisit the problem of scoring a phrase translation pair by developing a new phrase translation model based on Markov random fields (MRFs) and large-scale discriminative training. We strive to address the following three primary concerns.

First of all, instead of parameterizing a phrase translation pair using a set of scoring functions that are learned independently (e.g., phrase translation probabilities and lexical weights) we use a general, statistical framework in which arbitrary features extracted from a phrase pair can be incorporated to model the translation in a unified way. To this end, we propose the use of a MRF model.

Second, because the phrase model has to work with other component models in an SMT system in order to produce good translations and the quality of translation is measured via BLEU score, it is desirable to optimize the parameters of the phrase model jointly with other component models with respect to an objective function that is closely related to the evaluation metric under consideration, i.e., BLEU in this paper. To this end, we resort to a large-scale discriminative training approach, following the pioneering work of Liang et al. (2006). Although there are established methods of tuning a handful of features on small training sets, such as the MERT method (Och 2003), the development of discriminative training methods for millions of features on millions of sentence pairs is still an ongoing area of research. A recent survey is due to Koehn (2010). In this paper we show that by using stochastic gradient ascent and an N-best list based

Proceedings of NAACL-HLT 2013, pages 450–459,
Atlanta, Georgia, 9–14 June 2013.

expected BLEU as the objective function, large-scale discriminative training can lead to significant improvements.

The third primary concern is the ease of adoption of the proposed method. To this end, we use a simple and well-established learning method, ensuring that the results can be easily reproduced. We also develop the features for the MRF model in such a way that the resulting model is of the same format as that of a traditional phrase table. Thus, the model can be easily incorporated into a standard phrase-based SMT system, requiring no code change in the runtime engine.

In the rest of the paper, Section 2 presents the MRF model for phrase translation. Section 3 describes the way the model parameters are estimated. Section 4 presents the experimental results on two Europarl translation tasks. Section 5 reviews previous work that lays the foundation of this study. Section 6 concludes the paper.

2 Model

The traditional translation models are directional models that are based on conditional probabilities. As suggested by the noisy-channel model for SMT (Brown et al. 1993):

$$E^* = \underset{E}{\text{argmax}}\, P(E|F) = \underset{E}{\text{argmax}}\, P(E)P(F|E \qquad (1)$$

The Bayes rule leads us to invert the conditioning of translation probability from a foreign (source) sentence F to an English (target) translation E.

However, in practice, the implementation of state-of-the-art phrase-based SMT systems uses a weighted log-linear combination of several models $h(F, E, A)$ including the logarithm of the phrase probability (and the lexical weight) in source-to-target and target-to-source directions (Och and Ney 2004)

$$\begin{aligned} E^* &= \text{argmax}_E \textstyle\sum_{m=1}^{M} \lambda_m h_m(F, E, A) \qquad (2) \\ &= \underset{E}{\text{argmax}}\, Score_{\boldsymbol{\lambda}}(F, E) \end{aligned}$$

where A in $h(F, E, A)$ is a hidden structure that best derives E from F, called the *Viterbi derivation* afterwards. In phrase-based SMT, A consists of (1) the segmentation of the source sentence into phrases, (2) the segmentation of the target sentence into phrases, and (3) an alignment between the source and target phrases.

In this paper we use Markov random fields (MRFs) to model the joint distribution $P_{\mathbf{w}}(\mathbf{f}, \mathbf{e})$ over a source-target translation phrase pair $(\mathbf{f}, \mathbf{e})$, parameterized by $\mathbf{w}$. Different from the directional translation models, as in Equation (1), the MRF model is undirected, which we believe upholds the spirit of the use of bi-directional translation probabilities under the log-linear framework. That is, the agreement or the *compatibility* of a phrase pair is more effective to score translation quality than a directional translation probability which is modeled based on an imagined generative story does.

2.1 MRF

MRFs, also known as undirected graphical models, are widely used in modeling joint distributions of spatial or contextual dependencies of physical phenomena (Bishop 2006). A Markov random field is constructed from a graph G. The nodes of the graph represent random variables, and edges define the independence semantics between the random variables. An MRF satisfies the Markov property, which states that a node is independent of all of its non-neighbors, defined by the clique configurations of G. In modeling a phrase translation pair, we define two types of nodes, (1) two phrase nodes and (2) a set of word nodes, each for a word in these phrases, such as the graph in Figure 1. Let us denote a clique by c and the set of variables in that clique by $(\mathbf{f}, \mathbf{e})_c$. Then, the joint distribution over the random variables in G is defined as

$$P_{\mathbf{w}}(\mathbf{f}, \mathbf{e}) = \frac{1}{Z} \textstyle\prod_{c \epsilon C(G)} \varphi_c((\mathbf{f}, \mathbf{e})_c; \mathbf{w}), \qquad (3)$$

where $\mathbf{e} = e_1, \dots, e_{|\mathbf{e}|}$, $\mathbf{f} = f_1, \dots, f_{|\mathbf{f}|}$ and $C(G)$ is the set of cliques in G, and each $\varphi_c((\mathbf{f}, \mathbf{e})_c; \mathbf{w})$ is a non-negative potential function defined over a clique c that measures the *compatibility* of the variables in c, $\mathbf{w}$ is a set of parameters that are used within the potential function. Z in Equation (3), sometimes called the *partition function*, is a normalization constant and is given by

$$\begin{aligned} Z &= \textstyle\sum_{\mathbf{f}} \sum_{\mathbf{e}} \prod_{c \epsilon C(G)} \varphi_c((\mathbf{f}, \mathbf{e})_c; \mathbf{w}) \qquad (4) \\ &= \textstyle\sum_{\mathbf{f}} \sum_{\mathbf{e}} Score(\mathbf{f}, \mathbf{e}), \end{aligned}$$

which ensures that the distribution $P_{\mathbf{w}}(\mathbf{f}, \mathbf{e})$ given by Equation (3) is correctly normalized. The pres-

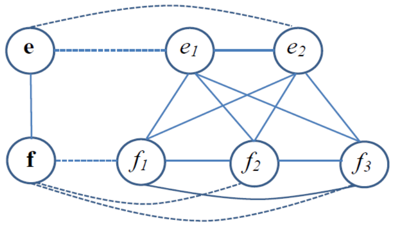

Figure 1: A Markov random field model for phrase translation of $\mathbf{e} = e_1, e_2$ and $\mathbf{f} = f_1, f_2, f_3$.

ence of Z is one of the major limitations of MRFs because it is generally not feasible to compute due to the exponential number of terms in the summation. However, we notice that Z is a global constant which is independent of $\mathbf{e}$ and $\mathbf{f}$. Therefore, in ranking phrase translation hypotheses, as performed by the decoder in SMT systems, we can drop Z and simply rank each hypothesis by its unnormalized joint probability. In our implementation, we only store in the phrase table for each translation pair $(\mathbf{f}, \mathbf{e})$ its unnormalized probability, i.e., $Score(\mathbf{f}, \mathbf{e})$ as defined in Equation (4).

It is common to define MRF potential functions of the exponential form as $\varphi_c((\mathbf{f}, \mathbf{e})_c; \mathbf{w}) = \exp(w_c \phi(c))$, where $\phi(c)$ is a real-valued feature function over clique c and w_c is the weight of the feature function. In phrase-based SMT systems, the sentence-level translation probability from F to E is decomposed as the product of a set of phrase translation probabilities. By dropping the phrase segmentation and distortion model components, we have

$$P(E|F) \approx \max_A P(E|A, F) \qquad (5)$$

$$P(E|A, F) = \prod_{(\mathbf{f},\mathbf{e}) \in A} P(\mathbf{e}|\mathbf{f}),$$

where A is the Viterbi derivation. Similarly, the joint probability $P(F, E)$ can be decomposed as

$$P(F, E) \approx \max_A P(F, A, E) \qquad (6)$$

$$\begin{aligned} P(F, A, E) &= \prod_{(\mathbf{f},\mathbf{e}) \in A} P_{\mathbf{w}}(\mathbf{f}, \mathbf{e}) \\ &\propto \sum_{(\mathbf{f},\mathbf{e}) \in A} \log P_{\mathbf{w}}(\mathbf{f}, \mathbf{e}) \\ &\propto \sum_{(\mathbf{f},\mathbf{e}) \in A} \sum_{c \in C(G_{(\mathbf{f},\mathbf{e})})} w_c \phi(c) \\ &= \sum_{(\mathbf{f},\mathbf{e}) \in A} \mathbf{w} \cdot \boldsymbol{\phi}(\mathbf{f}, \mathbf{e}) \end{aligned}$$

which is essentially proportional to a weighted linear combination of a set of features.

To instantiate an MRF model, one needs to define a graph structure representing the translation dependencies between source and target phrases, and a set of potential functions over the cliques of this graph.

2.2 Cliques and Potential Functions

The MRF model studied in this paper is constructed from the graph G in Figure 1. It contains two types of nodes, including two phrase nodes for the source and target phrases respectively and word nodes, each for a word in these phrases. The cliques and their corresponding potential functions (or features) attempt to abstract the idea behind those translation models that have been proved effective for machine translation in previous work. In this study we focus on three types of cliques.

First, we consider cliques that contain two phrase nodes. A potential function over such a clique captures phrase-to-phrase translation dependencies similar to the use the bi-directional translation models in phrase-based SMT systems. The potential is defined as $\varphi_p(\mathbf{f}, \mathbf{e}) = w_p \phi_p(\mathbf{f}, \mathbf{e})$, where the feature $\phi_p(\mathbf{f}, \mathbf{e})$, called the *phrase-pair feature*, is an indicator function whose value is 1 if $\mathbf{e}$ is target phrase and $\mathbf{f}$ is source phrase, and 0 otherwise. While the conditional probabilities in a directional translation model are estimated using relative frequencies of phrase pairs extracted from word-aligned parallel sentences, the parameter of the phrase-pair function w_p is learned discriminatively, as we will describe in Section 3.

Second, we consider cliques that contain two word nodes, one in source phrase and the other in target phrase. A potential over such a clique captures word-to-word translation dependencies similar to the use the IBM Model 1 for lexical weighting in phrase-based SMT systems (Koehn et al. 2003). The potential function is defined as $\varphi_t(f, e) = w_t \phi_t(f, e)$, where the feature $\phi_t(f, e)$, called the *word-pair feature*, is an indicator function whose value is 1 if e is a word in target phrase $\mathbf{e}$ and f is a word in source phrase $\mathbf{f}$, and 0 otherwise.

The third type of cliques contains three word nodes. Two of them are in one language and the third in the other language. A potential over such a clique is intended to capture inter-word dependencies for selecting word translations. The potential function is inspired by the triplet lexicon model (Hasan et al. 2008) which is based on lexicalized triplets (e, f, f'). It can be understood as two source (or target) words triggering one target (or source) word. The potential function is defined as $\varphi_{tp}(f, f', e) = w_{tp}\phi_{tp}(f, f', e)$, where the feature $\phi_{tp}(f, f', e)$, called the *triplet feature*, is an indicator function whose value is 1 if e is a word in target phrase **e** and f and f' are two different words in source phrase **f**, and 0 otherwise.

For any clique c that contains nodes in only one language we assume that $\varphi(c) = 1$ for all setting of the clique, which has no impact on scoring a phrase pair. One may wish to define a potential over cliques containing a phrase node and word nodes in target language, which could act as a form of target language model. One may also add edges in the graph so as to define potentials that capture more sophisticated translation dependencies. The optimal potential set could vary among different language pairs and depend to a large degree upon the amount and quality of training data. We leave a comprehensive study of features to future work.

3 Training

This section describes the way the parameters of the MRF model are estimated. Although MRFs are by nature generative models, it is not always appropriate to train the parameters using conventional likelihood based approaches mainly for two reasons. The first is due to the difficulty in computing the partition function in Equation (4), especially in a task of our scale. The second is due to the metric divergence problem (Morgan et al. 2004). That is, the maximum likelihood estimation is unlikely to be optimal for the evaluation metric under consideration, as demonstrated on a variety of tasks including machine translation (Och 2003) and information retrieval (Metzler and Croft 2005; Gao et al. 2005). Therefore, we propose a large-scale discriminative training approach that uses stochastic gradient ascent and an N-best list based expected BLEU as the objective function.

We cast machine translation as a structured classification task (Liang et al. 2006). It maps an input source sentence F to an output pair (E, A) where E is the output target sentence and A the Viterbi derivation of E. A is assumed to be constructed during the translation process. In phrase-based SMT, A consists of a segmentation of the source and target sentences into phrases and an alignment between source and target phrases.

We also assume that translations are modeled using a linear model parameterized by a vector $\boldsymbol{\theta}$. Given a vector $\mathbf{h}(F, E, A)$ of feature functions on (F, E, A), and assuming $\boldsymbol{\theta}$ contains a component for each feature, the output pair (E, A) for a given input F are selected using the argmax decision rule

$$(E^*, A^*) = \underset{(E,A)}{\text{argmax}}\, \boldsymbol{\theta}^T \mathbf{h}(F, E, A) \qquad (7)$$

In phrase-based SMT, computing the argmax exactly is intractable, so it is performed approximately by beam decoding.

In a phrase-based SMT system equipped by a MRF-based phrase translation model, the parameters we need to learn are $\boldsymbol{\theta} = (\boldsymbol{\lambda}, \mathbf{w})$, where $\boldsymbol{\lambda}$ is a vector of a handful parameters used in the log-linear model of Equation (2), with one weight for each component model; and $\mathbf{w}$ is a vector containing millions of weights, each for one feature function in the MRF model of Equation (3). Our method takes three steps to learn $\boldsymbol{\theta}$:

1. Given a baseline phrase-based SMT system and a pre-set $\boldsymbol{\lambda}$, we generate for each source sentence in training data an N-best list of translation hypotheses.
2. We fix $\boldsymbol{\lambda}$, and optimize $\mathbf{w}$ with respect to an objective function on training data.
3. We fix $\mathbf{w}$, and optimize $\boldsymbol{\lambda}$ using MERT (Och 2003) to maximize the BLEU score on development data.

Now, we describe Steps 1 and 2 in detail.

3.1 N-Best Generation

Given a set of source-target sentence pairs as training data $(F_n, E_n^r), n = 1 \dots N$, we use the baseline phrase-based SMT system to generate for each source sentence F a list of 100-best candidate translations, each translation E coupled with its Viterbi derivation A, according to Equation (7). We denote the 100-best set by $\text{GEN}(F)$. Then, each output pair (E, A) is labeled by a sentence-level BLEU score, denoted by sBLEU, which is computed according to Equation (8) (He and Deng 2012),

$$\text{sBLEU}(E, E^r) = BP \times \tfrac{1}{4}\textstyle\sum_{n=1}^{4} \log p_n, \qquad (8)$$

where E^r is the reference translation, and $p_n, n = 1 \ldots 4$, are precisions of *n*-grams. While precisions of lower order *n*-grams, i.e., p_1 and p_2, are computed directly without any smoothing, matching counts for higher order *n*-grams could be sparse at the sentence level and need to be smoothed as

$$p_n = \frac{\#(matched\ ngram) + \alpha p_n^0}{\#(ngram) + \alpha}, \text{for } n = 3,4$$

where α is a smoothing parameter and is set to 5, and p_n^0 is the prior value of p_n, whose value is computed as $p_n^0 = (p_{n-1})^2 / p_{n-2}$ for $n = 3$ and 4. BP in Equation (8) is the sentence-level *brevity penalty*, computed as $BP = \exp\left(1 - \beta \frac{r}{c}\right)$, which differs from its corpus-level counterpart (Papineni et al. 2002) in two ways. First, we use a non-clipped BP, which leads to a better approximation to the corpus-level BLEU computation because the per-sentence BP might effectively exceed unity in corpus-level BLEU computation, as discussed in Chiang et al. (2008). Second, the ratio between the length of reference sentence r and the length of translation hypothesis c is scaled by a factor β such that the total length of the references on training data equals that of the 1-best translation hypotheses produced by the baseline SMT system. In our experiments, the value of β is computed, on the N-best training data, as the ratio between the total length of the references and that of the 1-best translation hypotheses

In our experiments we find that using sBLEU defined above leads to a small but consistent improvement over other variations of sentence-level BLEU proposed previously (e.g., Liang et al. 2006). In particular, the use of the scaling factor β in computing BP makes BP of the baseline's 1-best output close to perfect on training data, and has an effect of forcing the discriminative training to improve BLEU by improving *n*-gram precisions rather than by improving brevity penalty.

3.2 Parameter Estimation

We use an N-best list based expected BLEU, a variant of that in Rosti et al. (2011), as the objective function for parameter optimization. Given the current model $\boldsymbol{\theta}$, the expected BLEU, denoted by $\text{xBLEU}(\boldsymbol{\theta})$, over one training sample i.e., a labeled N-best list $\text{GEN}(F)$ generated from a pair of source and target sentences (F, E^r), is defined as

$$\begin{aligned} &\text{xBLEU}(\boldsymbol{\theta}) \\ &= \textstyle\sum_{E \in \text{GEN}(F)} P_{\boldsymbol{\theta}}(E|F) \text{sBLEU}(E, E^r), \end{aligned} \tag{9}$$

where sBLEU is the sentence-level BLEU, defined in Equation (8), and $P_{\boldsymbol{\theta}}(E|F)$ is a normalized translation probability from F to E computed using *softmax* as

$$P_{\boldsymbol{\theta}}(E|F) = \frac{\exp(Score_{\boldsymbol{\theta}}(F,E))}{\sum_{E'} \exp(Score_{\boldsymbol{\theta}}(F,E'))}, \tag{10}$$

where $Score(.)$ is the translation score according to the current model $\boldsymbol{\theta}$

$$\begin{aligned} Score_{\boldsymbol{\theta}}(F,E) &= \boldsymbol{\lambda} \cdot \mathbf{h}(F,E,A) \\ &+ \textstyle\sum_{(\mathbf{f},\mathbf{e}) \in A} \mathbf{w} \cdot \boldsymbol{\phi}(\mathbf{f},\mathbf{e}). \end{aligned} \tag{11}$$

The right hand side of (11) contains two terms. The first term is the score produced by the baseline system, which is fixed during phrase model training. The second term is the translation score produced by the MRF model, which is updated after each training sample during training. Comparing Equations (2) and (11), we can view the MRF model yet another component model under the log linear model framework with its λ being set to 1.

Given the objective function, the parameters of the MRF model are optimized using stochastic gradient ascent. As shown in Figure 2, we go through the training set T times, each time is considered an *epoch*. For each training sample, we update the model parameters as

$$\mathbf{w}^{new} = \mathbf{w}^{old} + \eta \cdot \mathbf{g}(\mathbf{w}^{old}) \tag{12}$$

where η is the learning rate, and the gradient $\mathbf{g}$ is computed as

$$\mathbf{g}(\mathbf{w}) = \frac{\partial \text{xBLEU}(\mathbf{w})}{\partial \mathbf{w}} \tag{13}$$

```
Initialize $\mathbf{w}$, assuming $\boldsymbol{\lambda}$ is fixed during training
For $t = 1 \ldots T$ ($T$ = the total number of iterations)
  For each training sample (labeled 100-best list)
    Compute $P_{\boldsymbol{\theta}}(E|F)$ for each translation hypothesis $E$ based on the current model $\boldsymbol{\theta} = (\boldsymbol{\lambda}, \mathbf{w})$
    Update the model via $\mathbf{w} = \mathbf{w} + \eta \cdot \mathbf{g}(\mathbf{w})$,
    where $\eta$ is the learning rate and $\mathbf{g}$ the gradient computed according to Equations (12) and (13)
```

Figure 2: The algorithm of training a MRF-based phrase translation model.

$$= \sum_{(E,A)} \mathrm{U}(\boldsymbol{\theta}, E) P_{\boldsymbol{\theta}}(E|F) \phi(F, E, A),$$

where $\mathrm{U}(\boldsymbol{\theta}, E) = \mathrm{sBLEU}(E, E^r) - \mathrm{xBLEU}(\boldsymbol{\theta})$.

Two considerations regarding the development of the training method in Figure 2 are worth mentioning. They significantly simplify the training procedure without sacrificing much the quality of the trained model. First, we do not include a regularization term in the objective function because we find early stopping and cross valuation more effective and simpler to implement. In experiments we produce a MRF model after each epoch, and test its quality on a development set by first combining the MRF model with other baseline component models via MERT and then examining BLEU score on the development set. We performed training for T epochs ($T = 100$ in our experiments) and then pick the model with the best BLEU score on the development set. Second, we do not use the leave-one-out method to generate the N-best lists (Wuebker et al. 2010). Instead, the models used in the baseline SMT system are trained on the same parallel data on which the N-best lists are generated. One may argue that this could lead to over-fitting. For example, comparing to the translations on unseen test data, the generated translation hypotheses on the training set are of artificially high quality with the derivations containing artificially long phrase pairs. The discrepancy between the translations on training and test sets could hurt the training performance. However, we found in our experiments that the impact of over-fitting on the quality of the trained MRF models is negligible[1].

4 Experiments

We conducted our experiments on two Europarl translation tasks, German-to-English (DE-EN) and French-to-English (FR-EN). The data sets are published for the shared task in NAACL 2006 Workshop on Statistical Machine Translation (WMT06) (Koehn and Monz 2006).

For DE-EN, the training set contains 751K sentence pairs, with 21 words per sentence on average. The official development set used for the shared task contains 2000 sentences. In our experiments, we used the first 1000 sentences as a development set for MERT training and optimizing parameters for discriminative training, such as learning rate and the number of iterations. We used the rest 1000 sentences as the first test set (TEST1). We used the WMT06 test data as the second test set (TEST2), which contains 2000 sentences.

Systems	DE-EN (TEST2)	FR-EN (TEST2)
Rank-1 system	27.3	30.8
Rank-2 system	26.0	30.7
Rank-3 system	25.6	30.5
Our baseline	26.0	31.4

Table 1: Baseline results in BLEU. The results of top ranked systems are reported in Koehn and Monz (2006)[2].

For FR-EN, the training set contains 688K sentence pairs, with 21 words per sentence on average. The development set contains 2000 sentences. We used 2000 sentences from the WMT05 shared task as TEST1, and the 2000 sentences from the WMT06 shared task as TEST2.

Two baseline phrase-based SMT systems, each for one language pair, are developed as follows. These baseline systems are used in our experiments both for comparison purpose and for generating N-best lists for discriminative training. First, we performed word alignment on the training set using a hidden Markov model with lexicalized distortion (He 2007), then extracted 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 a baseline system include a lexicalized reordering model, word count and phrase count, and a trigram language model trained on the English training data provided by the WMT06 shared task. A fast beam-search phrase-based decoder (Moore and Quirk 2007) is used and the distortion limit is set to four. The decoder is modified so as to output the Viterbi derivation for each translation hypothesis.

The metric used for evaluation is case insensitive BLEU score (Papineni et al. 2002). We also performed a significance test using the paired *t*-test. Differences are considered statistically significant when the *p*-value is less than 0.05. Table 1

[1] As pointed out by one of the reviewers, the fact that our training works fine without leave-one-out is probably due to the small phrase length limit (i.e., 4) we used. If a longer phrase limit (e.g., 7) is used the result might be different. We leave it to future work.

[2] The official results are accessible at http://www.statmt.org/wmt06/shared-task/results.html

#	Systems	DE-EN		FR-EN	
		TEST1	TEST2	TEST1	TEST2
1	**Baseline**	26.0	26.0	31.3	31.4
2	**MRF**$_{p+t+tp}$	**27.3**$^{\alpha}$	**27.1**$^{\alpha}$	**32.4**$^{\alpha}$	**32.2**$^{\alpha}$
3	**MRF**$_{p+t}$	27.2$^{\alpha}$	26.9$^{\alpha}$	32.3$^{\alpha}$	32.0$^{\alpha}$
4	**MRF**$_{p}$	26.8$^{\alpha\beta}$	26.7$^{\alpha\beta}$	32.2$^{\alpha}$	31.8$^{\alpha\beta}$
5	**MRF**$_{t}$	26.8$^{\alpha\beta}$	26.8$^{\alpha}$	32.1$^{\alpha}$	31.9$^{\alpha\beta}$

Table 2: Main results (BLEU scores) of MRF-based phrase translation models with different feature classes. The superscripts α and β indicate statistically significant difference ($p < 0.05$) from **Baseline** and **MRF**$_{p+t+tp}$, respectively.

Feature classes	# of features (weights)	
	DE-EN	FR-EN
phrase-pair features (*p*)	2.5M	2.3M
word-pair features (*t*)	12.2M	9.7M
triplet features (*tp*)	13.4M	13.8M

Table 3: Statistics of the features used in building MRF-based phrase translation models.

presents the baseline results. The performance of our phrase-based SMT systems compares favorably to the top-ranked systems, thus providing a fair baseline for our research.

4.1 Results

Table 2 shows the main results measured in BLEU evaluated on TEST1 and TEST2.

Row 1 is the baseline system. Rows 2 to 5 are the systems enhanced by integrating different versions of the MRF-based phrase translation model. These versions, labeled as **MRF**$_f$, are trained using the method described in Section 3, and differ in the feature classes (which are specified by the subscript *f*) incorporated in the MRF-based model. In this study we focused on three classes of features, as described in Section 2, phrase-pair features (*p*), word-pair features (*t*) and triplet features (*tp*). The statistics for these features are given in Table 3.

Table 2 shows that all the MRF models lead to a substantial improvement over the baseline system across all test sets, with a statistically significant margin from 0.8 to 1.3 BLEU points. As expected, the best phrase model incorporates all of the three classes of features (**MRF**$_{p+t+tp}$ in Row 2). We also find that both **MRF**$_p$ and **MRF**$_t$, although using only one class of features, perform quite well. In TEST2 of DE-EN and TEST1 of FR-EN, they are in a near statistical tie with **MRF**$_{p+t}$ and **MRF**$_{p+t+tp}$.

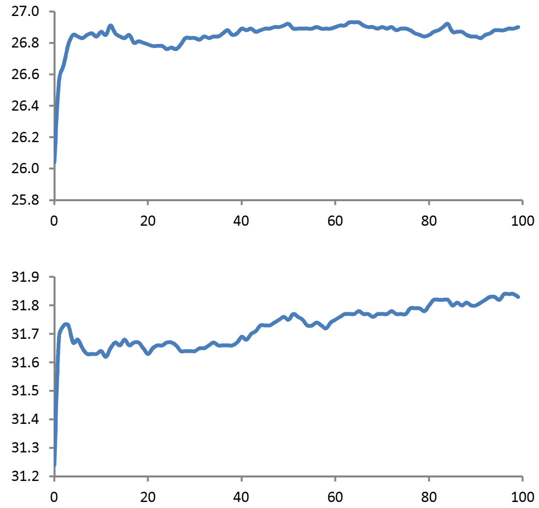

Figure 3: BLEU score on development data (*y* axis) for DE-EN (top) and FR-EN (bottom) as a function of the number of epochs (*x* axis).

The result suggests that while the MRF models are very effective in modeling phrase translations, the features we used in this study may not fully realize the potential of the modeling technology.

We also measured the sensitivity of the discriminative training method to different initializations and training parameters. Results show that our method is very robust. All the MRF models in Table 2 are trained by setting the initial feature vector to zero, and the learning rate η=0.01. Figure 3 plots the BLEU score on development sets as a function of the number of epochs *t*. The BLEU score improves quickly in the first 5 epochs, and then either remains flat, as on the DE-EN data, or keeps increasing but in a much slower pace, as on the FR-EN data.

4.2 Comparing Objective Functions

This section compares different objective functions for discriminative training. As shown in Table 4, xBLEU is compared to three widely used convex loss functions, i.e., hinge loss, logistic loss, and log loss. The hinge loss and logistic loss take into account only two hypotheses among an N-best list GEN: the one with the best sentence-level BLEU score with respect to its reference translation, denoted by (E^*, A^*), called the *oracle* candidate henceforth, and the highest scored *incorrect* candidate according to the current model, denoted by (E', A'), defined as

#	Objective functions	DE-EN		FR-EN	
		TEST 1	TEST2	TEST1	TEST2
1	**xBLEU**	27.2	26.9	32.3	32.0
2	**hinge loss**	26.4$^{\alpha}$	26.2$^{\alpha}$	31.8$^{\alpha}$	31.5$^{\alpha}$
3	**logistic loss**	26.3$^{\alpha}$	26.2$^{\alpha}$	31.7$^{\alpha}$	31.5$^{\alpha}$
4	**log loss**	26.5$^{\alpha}$	26.2$^{\alpha}$	32.1	31.7$^{\alpha}$

Table 4: BLEU scores of MRF-based phrase translation models trained using different objective functions. The MRF models use phrase-pair and word-pair features. The superscript α indicates statistically significant difference (p < 0.05) from **xBLUE**.

$$(E', A') = \mathrm{argmax}_{(E,A)\in\mathrm{GEN}(F)\backslash\{(E^*,A^*)\}} Score_\theta(F, E, A),$$

where $Score_\theta(.)$ is defined in Equation (11). Let $\mathbf{x} = \mathbf{h}(F, E^*, A^*) - \mathbf{h}(F, E', A')$. The hinge loss under the N-best re-ranking framework is defined as $\max(0, 1 - \boldsymbol{\theta}^T\mathbf{x})$. It is easy to verify that to train a model using this version of hinge loss, the update rule of Equation (12) can be rewritten as

$$\mathbf{w}^{new} = \begin{cases} \mathbf{w}^{old}, & \text{if } \hat{E} = E^* \\ \mathbf{w}^{old} + \eta\mathbf{x}, & otherwise \end{cases} \quad (14)$$

where $\hat{E}$ is the highest scored candidate in GEN. Following Shalev-Shwartz (2012), by setting $\eta = 1$, we reach the Perceptron-based training algorithm that has been widely used in previous studies of discriminative training for SMT (e.g., Liang et al. 2006; Simianer et al. 2012).

The logistic loss $\log(1 + \exp(-\boldsymbol{\theta}^T\mathbf{x}))$ leads to an update rule similar to that of hinge loss

$$\mathbf{w}^{new} = \begin{cases} \mathbf{w}^{old}, & \text{if } \hat{E} = E^* \\ \mathbf{w}^{old} + \eta P_{\boldsymbol{\theta}}(\mathbf{x})\mathbf{x}, & otherwise \end{cases} \quad (15)$$

where $P_{\boldsymbol{\theta}}(\mathbf{x}) = 1/(1 + \exp(\boldsymbol{\theta}^T\mathbf{x}))$.

The log loss is widely used when a probabilistic interpretation of the trained model is desired, as in conditional random fields (CRFs) (Lafferty et al. 2001). Given a training sample, log loss is defined as $\log P_{\boldsymbol{\theta}}(E^*|F)$, where E^* is the oracle translation hypothesis with respect to its reference translation. $P_{\boldsymbol{\theta}}(E^*|F)$ is computed as Equation (10). So, unlike hinge loss and logistic loss, log loss takes into account the distribution over all hypotheses in an N-best list.

The results in Table 4 suggest that the objective functions that take into account the distribution over all hypotheses in an N-best list (i.e., xBLEU and log loss) are more effective than the ones that do not. xBLEU, although it is a non-concave function, significantly outperforms the others because it is more closely coupled with the evaluation metric under consideration (i.e., BLEU).

5 Related Work

Among the attempts to learning phrase translation probabilities that go beyond pure counting of phrases on word-aligned corpora, Wuebker et al. (2010) and He and Deng (2012) are most related to our work. The former find phrase alignment directly on training data and update the translation probabilities based on this alignment. The latter learn phrase translation probabilities discriminatively, which is similar to our approach. But He and Deng's method involves multiple stages, and is not straightforward to implement[3]. Our method differs from previous work in its use of a MRF model that is simple and easy to understand, and a stochastic gradient ascent based training method that is efficient and easy to implement.

A large portion of previous studies on discriminative training for SMT either use a handful of features or use small training sets of a few thousand sentences (e.g., Och 2003; Shen et al. 2004; Watanabe et al. 2007; Duh and Kirchhoff 2008; Chiang et al. 2008; Chiang et al. 2009). Although there is growing interest in large-scale discriminative training (e.g., Liang et al. 2006; Tillmann and Zhang 2006; Blunsom et al. 2008; Hopkins and May 2011; Zhang et al. 2011), only recently does some improvement start to be observed (e.g., Simianer et al. 2012; He and Deng 2012). It still remains uncertain if the improvement is attributed to new features, new training algorithms, objective functions, or simply large amounts of training data. We show empirically the importance of objective functions. Gimple and Smith (2012) also analyze objective functions, but more from a theoretical viewpoint.

The proposed MRF-based translation model is inspired by previous work of applying MRFs for information retrieval (Metzler and Croft 2005), query expansion (Metzler et al. 2007; Gao et al. 2012) and POS tagging (Haghighi and Klein 2006).

[3] For comparison, the method of He and Deng (2012) also achieved very similar results to ours using the same experimental setting, as described in Section 4.

Another undirected graphical model that has been more widely used for NLP is a CRF (Lafferty et al. 2001). An MRF differs from a CRF in that its partition function is no longer observation dependent. As a result, learning an MRF is harder than learning a CRF using maximum likelihood estimation (Haghighi and Klein 2006). Our work provides an alternative learning method that is based on discriminative training.

6 Conclusions

The contributions of this paper are two-fold. First, we present a general, statistical framework for modeling phrase translations via MRFs, where different features can be incorporated in a unified manner. Second, we demonstrate empirically that the parameters of the MRF model can be learned effectively using a large-scale discriminative training approach which is based on stochastic gradient ascent and an N-best list based expected BLEU as the objective function.

In future work we strive to fully realize the potential of the MRF model by developing features that can capture more sophisticated translation dependencies that those used in this study. We will also explore the use of MRF-based translation models for translation systems that go beyond simple phrases, such as hierarchical phrase based systems (Chiang 2005) and syntax-based systems (Galley et al. 2004).

References

Bishop, C. M. 2006. *Patten recognition and machine learning.* Springer.

Blunsom, P., Cohn, T., and Osborne, M. 2008. A discriminative latent variable models for statistical machine translation. In *ACL-HLT*.

Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., and Mercer, R. L. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2): 263-311.

Chiang, D. 2005. A hierarchical phrase-based model for statistical machine translation. In *ACL*, pp. 263-270.

Chiang, D., Knight, K., and Wang, W. 2009. 11,001 new features for statistical machine translation. In *NAACL-HLT*.

Chiang, D., Marton, Y., and Resnik, P. 2008. Online large-margin training of syntactic and structural translation features. In *EMNLP*.

DeNero, J., Gillick, D., Zhang, J., and Klein, D. 2006. Why generative phrase models underperform surface heuristics. In *Workshop on Statistical Machine Translation*, pp. 31-38.

Duh, K., and Kirchhoff, K. 2008. Beyond log-linear models: boosted minimum error rate training for n-best ranking. In *ACL*.

Galley, M., Hopkins, M., Knight, K., Marcu, D. 2004. What's in a translation rule? In *HLT-NAACL*, pp. 273-280.

Gao, J., Xie, S., He, X., and Ali, A. 2012. Learning lexicon models from search logs for query expansion. In *EMNLP-CoNLL*, pp. 666-676.

Gao, J., Qi, H., Xia, X., and Nie, J-Y. 2005. Linear discriminant model for information retrieval. In *SIGIR*, pp. 290-297.

Gimpel, K., and Smith, N. A. 2012. Structured ramp loss minimization for machine translation. In *NAACL-HLT*.

Haghighi, A., and Klein, D. 2006. Prototype-driven learning for sequence models. In *NAACL*.

Hasan, S., Ganitkevitch, J., Ney, H., and Andres-Fnerre, J. 2008. Triplet lexicon models for statistical machine translation. In *EMNLP*, pp. 372-381.

He, X. 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*.

He, X., and Deng, L. 2012. Maximum expected bleu training of phrase and lexicon translation models. In *ACL*, pp. 292-301.

Hopkins, H., and May, J. 2011. Tuning as ranking. In *EMNLP*.

Koehn, P. 2010. *Statistical machine translation. Cambridge University Press*.

Koehn, P., and Monz, C. 2006. Manual and automatic evaluation of machine translation between European languages. In *Workshop on Statistical Machine Translation*, pp. 102-121.

Koehn, P., Och, F., and Marcu, D. 2003. Statistical phrase-based translation. In *HLT-NAACL*, pp. 127-133.

Lafferty, J., McCallum, A., and Pereira, F. 2001. Conditional random fields: probablistic models for segmenting and labeling sequence data. In *ICML*.

Lambert, P., and Banchs, R.E. 2005. Data inferred multi-word expressions for statistical machine translation. In *MT Summit X*, Phuket, Thailand.

Liang, P., Bouchard-Cote, A. Klein, D., and Taskar, B. 2006. An end-to-end discriminative approach to machine translation. In *COLING-ACL*.

Marcu, D., and Wong, W. 2002. A phrase-based, joint probability model for statistical machine translation. In *EMNLP*.

Metzler, D., and Croft, B. 2005. A markov random field model for term dependencies. In *SIGIR*, pp. 472-479.

Metzler, D., and Croft, B. 2007. Latent concept expansion using markov random fields. In *SIGIR*, pp. 311-318.

Morgan, W., Greiff, W., and Henderson, J. 2004. *Direct maximization of average precision by hill-climbing with a comparison to a maximum entropy approach*. Technical report. MITRE.

Moore, R., and Quirk, C. 2007. Faster beam-search decoding for phrasal statistical machine translation. In *MT Summit XI*.

Och, F., and Ney, H. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 29(1): 19-51.

Och, F. 2003. Minimum error rate training in statistical machine translation. In *ACL*, pp. 160-167.

Papinein, K., Roukos, S., Ward, T., and Zhu W-J. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL*.

Rosti, A-V., Hang, B., Matsoukas, S., and Schwartz, R. S. 2011. Expected BLEU training for graphs: bbn system description for WMT system combination task. In *Workshop on Statistical Machine Translation*.

Shalev-Shwartz, Shai. 2012. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107-194.

Shen, L., Sarkar, A., and Och, F. 2004. Discriminative reranking for machine translation. In *HLT/NAACL*.

Simianer, P., Riezler, S., and Dyer, C. 2012. Joint feature selection in distributed stochasic learning for large-scale discriminative training in SMT. In *ACL*, pp. 11-21.

Tillmann, C., and Zhang, T. 2006. A discriminative global training algorithm for statistical MT. In *COLING-ACL*.

Watanabe, T., Suzuki, J., Tsukada, H., and Isozaki, H. 2007. Online large-margin training for statistical machine translation. In *EMNLP*.

Wuebker, J., Mauser, A., and Ney, H. 2010. Training phrase translation models with leaving-one-out. In *ACL*, pp. 475-484.

Zhang, Y., Deng, L., He, X., and Acero, A., 2011. A Novel decision function and the associated decision-feedback learning for speech translation, in *ICASSP*.