Large-scale Expected BLEU Training of Phrase-based Reordering Models

Michael Auli, Michel Galley, Jianfeng Gao
Microsoft Research
Redmond, WA, USA
{michael.auli,mgalley,jfgao}@microsoft.com

Abstract

Recent work by Cherry (2013) has shown that directly optimizing phrase-based reordering models towards BLEU can lead to significant gains. Their approach is limited to small training sets of a few thousand sentences and a similar number of sparse features. We show how the expected BLEU objective allows us to train a simple linear discriminative reordering model with millions of sparse features on hundreds of thousands of sentences resulting in significant improvements. A comparison to likelihood training demonstrates that expected BLEU is vastly more effective. Our best results improve a hierarchical lexicalized reordering baseline by up to 2.0 BLEU in a single-reference setting on a French-English WMT 2012 setup.

1 Introduction

Modeling reordering for phrase-based machine translation has been a long standing problem. Contrary to synchronous context free grammar-based translation models (Wu, 1997; Galley et al., 2004; Galley et al., 2006; Chiang, 2007), phrase-based models (Koehn et al., 2003; Och and Ney, 2004) have no in-built notion of reordering beyond what is captured in a single phrase pair, and the first phrase-based decoders simply scored inter-phrase reorderings using a restricted linear distortion feature, which scores a phrase reordering proportionally to the length of its displacement. While phrase-based models allow in theory completely unrestricted reordering patterns, movements are generally limited to a finite distance for complexity reasons. To address this limitation, extensive prior work focused on richer feature sets, in particular on lexicalized reordering models trained with maximum likelihood-based approaches (Tillmann, 2003; Xiong et al., 2006; Galley and Manning, 2008; Nguyen et al.,2009;§2).

More recently, Cherry (2013) proposed a very effective sparse ordering model relying on a set of only a few thousand indicator features which are trained towards a task-specific metric such as BLEU (Papineni et al., 2002). These features are simply added to the log-linear framework of translation that is trained with the Margin Infused Relaxed Algorithm (MIRA; Chiang et al., 2009) on a small development set of a few thousand sentences. While simple, the approach outperforms the state-of-the-art hierarchical reordering model of Galley and Manning (2008), a maximum likelihood-based model trained on millions of sentences to fit millions of parameters.

Ideally, we would like to scale sparse reordering models to similar dimensions but recent attempts to increase the amount of training data for MIRA was met with little success (Eidelman et al., 2013). In this paper we propose much larger sparse ordering models that combine the scalability of likelihood-based approaches with the higher accuracy of maximum BLEU training (§3). We train on the output of a hierarchical reordering model-based system and scale to millions of features learned on hundreds of thousands of sentences (§4). Specifically, we use the expected BLEU objective function (Rosti et al., 2010; Rosti et al., 2011; He and Deng, 2012; Gao and He, 2013; Gao et al., 2014; Green et al., 2014) which allows us to train models that use training data and feature sets that are two to three orders of magnitudes larger than in previous work (§5).

Our models significantly outperform the state-of-the-art hierarchical lexicalized reordering model on two language pairs and we demonstrate that richer feature sets result in significantly higher accuracy than with a feature set similar to Cherry (2013). We also demonstrate that our

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1250–1260,
October 25-29, 2014, Doha, Qatar.

approach greatly benefits from more training data than is typically used for maximum BLEU training. Previous work concluded that sparse reordering models perform better than maximum entropy models, however, the two approaches do not only differ in the objective function but also the type of training data (Cherry, 2013). Our analysis isolates the objective function and shows that expected BLEU optimization is the most important factor to train accurate ordering models. Finally, we compare expected BLEU training to pair-wise ranked optimization (PRO) on a feature set similar to Cherry (2013; §7).

2 Reordering Models

Reordering models for phrase-based translation are typically part of the log-linear framework which forms the basis of many statistical machine translation systems (Och and Ney, 2004).

Formally, we are given K training pairs $\mathcal{D} = (f^{(1)}, e^{(1)})...(f^{(K)}, e^{(K)})$, where each $f^{(i)} \in \mathcal{F}$ is drawn from a set of possible foreign sentences, and each English sentence $e^{(i)} \in \mathcal{E}(f^{(i)})$ is drawn from a set of possible English translations of $f^{(i)}$. The log-linear model is parameterized by m parameters θ where each $\theta_k \in \theta$ is the weight of an associated feature $h_k(f, e)$ such as a language model or a reordering model. Function $h(f, e)$ maps foreign and English sentences to the vector $h_1(f, e)...h_m(f, e)$, and we usually choose translations $\hat{e}$ according to the following decision rule:

$$\hat{e} = \arg\max_{e \in \mathcal{E}(f)} \theta^{\mathsf{T}} h(f, e) \quad (1)$$

In practice, computing $\hat{e}$ exactly is intractable and we resort to an approximate but more efficient beam search (Och and Ney, 2004).

Early phrase-based models simply relied on a linear distortion feature, which measures the distance between the first word of the current source phrase and the last word of the previous source phrase (Koehn et al., 2003; Och and Ney, 2004). Unfortunately, this approach is agnostic to the actual phrases being reordered, and does not take into account that certain phrases are more likely to be reordered than others. This shortcoming led to a range of *lexicalized* reordering models that capture exactly those preferences for individual phrases (Tillmann, 2003; Koehn et al., 2007).

Reordering models generally assume a sequence of English phrases $e = \{\bar{e}_1, \ldots, \bar{e}_n\}$ currently hypothesized by the decoder, a phrase alignment $a = \{a_1, \ldots, a_n\}$ that defines a foreign phrase $\bar{f}_{a_i}$ for each English phrase $\bar{e}_i$, and an *orientation* o_i which describes how a phrase pair should be reordered with respect to the previous phrases. There are typically three orientation types and the exact definition depends on the specific models which we describe below. Orientations can be determined during decoding and from word-aligned training corpora. Most models estimate a probability distribution $p(o_i|pp_i, a_1, \ldots, a_i)$ for the i-th phrase pair $pp_i = \langle \bar{e}_i, \bar{f}_{a_i} \rangle$ and the alignments $a_1, \ldots, a_i$ of the previous target phrases.

Lexicalized Reordering. This model defines the three orientation types based only on the position of the current and previously translated source phrase a_i and a_{i-1}, respectively (Tillmann, 2003; Koehn et al., 2007). The orientation types generally are: monotone (M), indicating that a_{i-1} is directly followed by a_i. swap (S) assumes that a_i precedes a_{i-1}, i.e., the two phrases swap places. Finally, discontinuous (D) indicates that a_i is not adjacent to a_{i-1}. The probability distribution over these reordering events is based on a maximum likelihood estimate:

$$p(o|pp, a_{i-1}, a_i) = \frac{cnt(o, pp)}{cnt(pp)} \quad (2)$$

where $o \in \{M, S, D\}$ and cnt returns smoothed frequency counts over a word-aligned corpus.

Hierarchical Reordering. An extension of the lexicalized reordering model better handles long-distance reordering by conditioning the orientation of the current phrase on a context larger than just the previous phrase (Galley and Manning, 2008). In particular, the hierarchical reordering model does so by building a compact representations of the preceding context using an efficient shift-reduce parser. During translation new phrases get moved on a stack and are then combined with any previous phrase if they are adjacent. Figure 1 shows an illustrative example: when the decoder shifts phrase pp_8 onto the stack, this phrase is then merged with pp_7 (reduce operation), which then can be merged with previous phrases to finally form a hierarchical block h_1. These merge operations stop once we reach a phrase (here, pp_3) that is not contiguous with the current block. Then, as another phrase (pp_9) is hypothesized, the decoder uses the hierarchical block at the top of the stack (h_1) to determine the orientation of the current

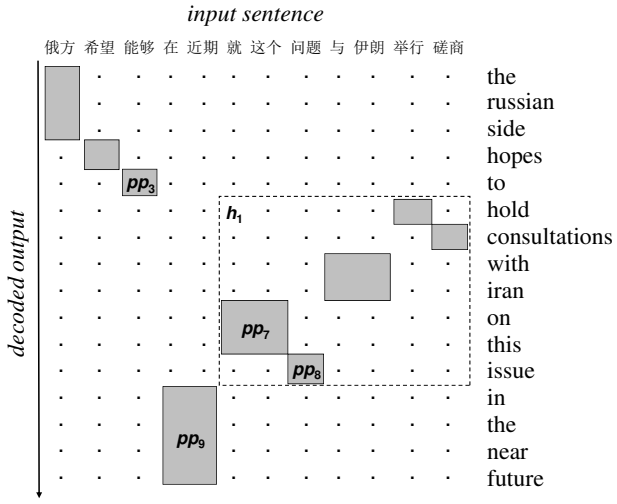

Figure 1: The hierarchical reordering model (HRM) analyzes a non-local context to determine the orientation of the current phrase. For example, the phrase pair pp_9 has a swap orientation ($o_9 = S$) with respect to a hierarchical block (h_1) that comprises the five preceding phrase pairs.

phrase pp_9, which in this case is a swap (S) orientation.[1] The model has the advantage that the orientations computed are more robust to derivational ambiguity of the underlying translation model. A given surface translation may be derived through different phrases but the shift-reduce parser combines them into a single representation which is more consistent with the orientations observed in the word-aligned training data.

Maximum Entropy-based models. The statistics used to estimate the lexicalized and the hierarchical reordering models are based on very sparse estimates, simply because certain phrases are not very frequent. Maximum entropy models address this problem by estimating Eq. 2 through sparse indicator features over phrase pairs instead, but prior work with such models still relies on word aligned corpora for estimation (Xiong et al., 2006; Nguyen et al., 2009). However, recent evaluations of the approach show little gain over the simpler frequency-based estimation method (Cherry, 2013).

Sparse Hierarchical Reordering model. All of the models so far are trained to maximize the likelihood of reordering decisions observed in word aligned corpora. Cherry (2013) argues that it is probably too difficult to learn human reordering patterns through noisy word alignments that were generated by unsupervised methods. Instead, he proposes to learn a discriminative reordering model based on the outputs of the actual machine translation system, adjusting the feature weights to maximize a task-specific objective, which is BLEU in their case. Their model is based on a set of sparse features derived from the hierarchical reordering model which we scale to millions of features (§6).

[1] Galley and Manning (2008) provide a more formal explanation.

3 A Simple Linear Reordering Model

Our reordering model is defined as a simple linear model over the basic orientation types, similar to Cherry (2013). In particular, our model defines score $s_\phi(o, e, f)$ over orientations $o = \{M, S, D\}$, and a sentence pair $\{e, f, a\}$ with alignment a as a linear combination of weighted indicator features:

$$\begin{aligned} s_\phi(o, e, f, a) &= \phi^\mathsf{T} u(o, e, f, a) \\ &= \sum_{i=1}^{I} \phi^\mathsf{T} u(o, pp_i, c_i) \\ &= \sum_{i=1}^{I} s_\phi(o, pp_i, c_i) \end{aligned} \quad (3)$$

where ϕ is a vector of weights, $\{pp_i\}_{i=1}^{I}$ is a set of phrases that decompose the sentence pair $\{e, f, a\}$, and $u(o, pp_i, c_i)$ is a function that maps orientation o, phrase pair pp_i and local context c_i to a sparse vector of indicator features. The local context c_i represents information used by the model that is in addition to the phrase pair. For example, the features of Cherry (2013) condition on the top-stack of the hierarchical shift reduce parser, information that is non-local with respect to the phrase pair. In our experiments, we use features that go beyond the top-stack, in order to condition on various parts of the source and target side contexts (§7).

4 Model Training

Optimization of our model is based on standard stochastic gradient descent (SGD; Bottou, 2004) with an expected BLEU loss $l(\phi)$ which we detail next (§5). The update is:

$$\phi_t = \phi_{t-1} - \mu \frac{\partial l(\phi_{t-1})}{\partial \phi_{t-1}} \quad (4)$$

where ϕ_t and ϕ_{t-1} are model weights at time t and $t-1$ respectively, and μ is a learning rate.

We add the model as a small number of dense features to the log-linear framework of translation

(Eq. 1). Specifically, we extend the m baseline features by a set of new features $h_{m+1}, \ldots, h_{m+j}$, where each represents a linear combination of sparse indicator features corresponding to one of the orientation types. Exposing each orientation as a separate dense feature within the log-linear model is common practice for lexicalized reordering models (Koehn et al., 2005):

$$h_{m+j} = s_\phi(o_j, e, f, a)$$

where $o_j \in \{M, S, D\}$.

The translation model is then parameterized by both θ, the log-linear weights of the baseline features, as well as ϕ, the weights of the reordering model. The reordering model is learned as follows (Gao and He, 2013; Gao et al., 2014):

1. We first train a baseline translation system to learn θ, without the discriminative reordering model, i.e., we set $\theta_{m+1} = 0, \ldots, \theta_{m+j} = 0$.

2. Using these weights, we generate n-best lists for the foreign sentences in the training data using the setup described in the experimental section (§7). The n-best lists serve as an approximation to $\mathcal{E}(f)$, the set of possible translations of f, used in the next step for expected BLEU training of the reordering model (§5).

3. Next, we fix θ, set $\theta_{m+1} = 1, \ldots \theta_{m+j} = 1$ and optimize ϕ with respect to the loss function on the training data using stochastic gradient descent.[2]

4. Finally, we fix ϕ and re-optimize θ in the presence of the discriminative reordering model using Minimum Error Rate Training (MERT; Och 2003; §7).

We found that re-optimizing θ after a few iterations of stochastic gradient descent in step 3 did not improve accuracy.

5 Expected BLEU Objective Function

The expected BLEU objective (Gao and He, 2013; Gao et al., 2014) allows us to efficiently optimize a large scale discriminative reordering model towards the desired task-specific metric, which in our setting is BLEU.

[2]We tuned $\theta_{m+1}, \ldots \theta_{m+j}$ on the development set but found that setting them uniformly to one resulted in faster training and equal accuracy.

Formally, we define our loss function $l(\phi)$ as the negative expected BLEU score, denoted as xBLEU(ϕ), for a given foreign sentence f and a log-linear parameter set θ:

$$\begin{aligned} l(\phi) &= -\text{ xBLEU}(\phi) \\ &= -\sum_{e \in \mathcal{E}(f)} p_{\theta,\phi}(e|f)\, \text{sBLEU}(e, e^{(i)}) \end{aligned} \quad (5)$$

where sBLEU$(e, e^{(i)})$ is a smoothed sentence-level BLEU score with respect to the reference translation $e^{(i)}$, and $\mathcal{E}(f)$ is the generation set approximated by an n-best list. In our experiments we use n-best lists with unique entries and therefore our definitions do not take into account multiple derivations of the same translation. Specifically, our n-best lists are generated by choosing the highest scoring derivation $\hat{e}$ amongst string identical translations e for f. We use a sentence-level BLEU approximation similar to Gao et al. (2014).[3] Finally, $p_{\theta,\phi}(e|f)$ is the normalized probability of translation e given f, defined as:

$$p_{\theta,\phi}(e|f) = \frac{\exp\{\gamma \theta^\mathsf{T} h(f, e)\}}{\sum_{e' \in \mathcal{E}(f)} \exp\{\gamma \theta^\mathsf{T} h(f, e')\}} \quad (6)$$

where $\theta^\mathsf{T} h(f, e)$ includes the discriminative reordering model $h_{m+1}(e, f), \ldots, h_{m+j}(e, f)$ parameterized by ϕ, and $\gamma \in [0, \inf)$ is a tuned scaling factor that flattens the distribution for $\gamma < 1$ and sharpens it for $\gamma > 1$ (Tromble et al., 2008).[4]

Next, we define the gradient of the expected BLEU loss function $l(\phi)$. To simplify our notation we omit the local context c in $s_\phi(o, pp, c)$ (Eq. 3) from now on and assume it to be part of pp. Using the observation that the loss does not explicitly depend on ϕ, we get:

$$\begin{aligned} \frac{\partial l(\phi)}{\partial \phi} &= \sum_{o,pp} \frac{\partial l(\phi)}{\partial s_\phi(o, pp)} \frac{\partial s_\phi(o, pp)}{\partial \phi} \\ &= \sum_{o,pp} -\delta_{o,pp} u(o, pp) \end{aligned}$$

where $\delta_{o,pp}$ is the *error term* for orientation o of phrase pair pp:

$$\delta_{o,pp} = -\frac{\partial l(\phi)}{\partial s_\phi(o, pp)}$$

[3]We found in early experiments that the BLEU+1 approximation used by Liang et al. (2006) and Nakov et. al (2012) worked equally well in our setting.

[4]γ is only used during expected BLEU training.

The error term indicates how the expected BLEU loss changes with the reordering score which we derive in the next section.

Finally, the gradient of the reordering score $s_\phi(o, pp)$ with respect to ϕ is simply given by this:

$$\frac{\partial s_\phi(o, pp)}{\partial \phi} = \frac{\partial \phi^\mathsf{T} u(o, pp)}{\partial \phi} = u(o, pp)$$

5.1 Derivation of the Error Term $\delta_{o,pp}$

We rewrite the loss function (Eq. 5) using Eq. 6 and separate it into two terms $G(\phi)$ and $Z(\phi)$:

$$l(\phi) = -\text{xBLEU}(\phi) = -\frac{G(\phi)}{Z(\phi)} \quad (7)$$
$$= -\frac{\sum_{e \in \mathcal{E}(f)} \exp\{\gamma \theta^\mathsf{T} h(f, e)\} \, \text{sBLEU}(e, e^{(i)})}{\sum_{e' \in \mathcal{E}(f)} \exp\{\gamma \theta^\mathsf{T} h(f, e')\}}$$

Next, we apply the quotient rule of differentiation:

$$\delta_{o,pp} = \frac{\partial \text{xBLEU}(\phi)}{\partial s_\phi(o, pp)} = \frac{\partial (G(\phi)/Z(\phi))}{\partial s_\phi(o, pp)}$$
$$= \frac{1}{Z(\phi)} \left(\frac{\partial G(\phi)}{\partial s_\phi(o, pp)} - \frac{\partial Z(\phi)}{\partial s_\phi(o, pp)} \text{xBLEU}(\phi) \right)$$

The gradients for $G(\phi)$ and $Z(\phi)$ with respect to $s_\phi(o, pp)$ are:

$$\frac{\partial G(\phi)}{\partial s_\phi(o, pp)} = \sum_{e \in \mathcal{E}(f)} \text{sBLEU}(e, e^{(i)}) \frac{\partial \exp\{\gamma \theta^\mathsf{T} h(f, e)\}}{\partial s_\phi(o, pp)}$$
$$\frac{\partial Z(\phi)}{\partial s_\phi(o, pp)} = \sum_{e \in \mathcal{E}(f)} \frac{\partial \exp\{\gamma \theta^\mathsf{T} h(f, e)\}}{\partial s_\phi(o, pp)}$$

By using the following definition:

$$U(\phi, e) = \text{sBLEU}(e, e^{(i)}) - \text{xBLEU}(\phi)$$

together with the chain rule, Eq. 6 and Eq. 7, we can rewrite $\delta_{o,pp}$ as follows:

$$\delta_{o,pp} = \frac{1}{Z(\phi)} \sum_{e \in \mathcal{E}(f)} \left(\frac{\partial \exp\{\gamma \theta^\mathsf{T} h(f, e)\}}{\partial s_\phi(o, pp)} U(\phi, e) \right)$$
$$= \sum_{e \in \mathcal{E}(f)} \left(p_{\theta,\phi}(e|f) \frac{\partial \gamma \theta^\mathsf{T} h(f, e)}{\partial s_\phi(o, pp)} U(\phi, e) \right)$$

Because ϕ is only relevant to the reordering model, represented by $h_{m+1}, \ldots, h_{m+j}$, we have:

$$\frac{\partial \gamma \theta^\mathsf{T} h(f, e)}{\partial s_\phi(o, pp)} = \gamma \lambda_k \frac{\partial h_k(e, f)}{\partial s_\phi(o, pp)}$$
$$= \gamma \lambda_k \mathcal{N}(o, pp, e, f)$$

1: **function** TRAINSGD($\mathcal{D}$, μ)
2: $t \leftarrow 0$
3: **for all** $(f^{(i)}, e^{(i)})$ in $\mathcal{D}$ **do**
4: xBLEU $= 0$ ▷ Compute xBLEU
5: **for all** e in $\mathcal{E}(f^{(i)})$ **do**
6: wBLEU $\leftarrow p_{\theta,\phi_t}(e|f)$ sBLEU$(e, e^{(i)})$
7: xBLEU $\leftarrow$ xBLEU + wBLEU
8: **end for**
9: **for all** e in $\mathcal{E}(f^{(i)})$ **do**
10: $D =$ sBLEU$(e, e^{(i)}) -$ xBLEU
11: **for all** o, pp in $\langle e, f^{(i)} \rangle$ **do**
12: $N = \mathcal{N}(o, pp, e, f)$
13: $\delta_{o,pp} = p_{\theta,\phi_t}(e|f^{(i)}) \gamma \lambda_k N D$
14: $\phi_{t+1} = \phi_t - \mu \delta_{o,pp} u(o, pp))$
15: **end for**
16: **end for**
17: $t \leftarrow t + 1$
18: **end for**
19: **end function**

Figure 2: Algorithm for computing the expected BLEU loss with SGD updates (Eq. 4) based on training data $\mathcal{D}$ and learning rate μ.

where $m + 1 \leq k \leq m + j$ and $\mathcal{N}(o, pp, e, f)$ is the number of times pp with orientation o occurs in the current sentence pair.

This simplifies the error term to:

$$\delta_{o,pp} = \sum_{e \in \mathcal{E}(f)} p_{\theta,\phi}(e|f) \gamma \lambda_k \mathcal{N}(o, pp, e, f) U(\phi, e) \quad (8)$$

where λ_k is the weight of the dense feature summarizing orientation o in the log-linear model. We use Eq. 8 in a simple algorithm to train our model (Figure 2). Our SGD trainer uses a mini-batch size of a single sentence (§7) which entails all hypothesis in the n-best list for this sentence and the parameters are updated after each mini-batch.

6 Feature Sets

Our features are inspired by Cherry (2013) who bases his features on the local phrase-pair $pp = \langle \bar{e}, \bar{f} \rangle$ as well as the top stack of the shift reduce parser of the baseline hierarchical ordering model. We experiment with these variants and extensions:

- SparseHRMLocal: This feature set is exclusively based on the local phrase-pair and

consists of features over the first and last word of both the source and target phrase.[5] We use four different word representations: The word identity itself, but only for the 80 most common source and target language words. The three other word representations are based on Brown clustering with either 20, 50 or 80 classes (Brown et al., 1992). There is one feature for every orientation type.

- SparseHRM: The main feature set of Cherry (2013). This is an extension of SparseHRM-Local adding features based on the first and last word of both the source and the target of the hierarchical block at the top of the stack. There are also features based on the source words *in-between* the current phrase and the hierarchical block at the top of the stack.

- SparseHRM+UncommonWords: This set is identical to SparseHRM, except that word-identity features are not restricted to the 80 most frequent words, but can be instantiated for all words, regardless of frequency.

- SparseHRM+BiPhrases: This augments SparseHRM by phrase-identity features resulting in millions of instances compared to only a few thousand for SparseHRM. We add three features for each possible phrase pair: the source phrase, the target phrase, and the whole phrase pair.

The baseline hierarchical lexicalized reordering model is most similar to SparseHRM+BiPhrases feature set since both have parameters for phrase, orientation pairs.[6] The feature set closest to Cherry (2013) is SparseHRM. However, while Cherry had to severely restrict his features for batch lattice MIRA-based training, our maximum expected BLEU approach can handle millions of features.

7 Experiments

Baseline. We experiment with a phrase-based system similar to Moses (Koehn et al., 2007), scoring translations by a set of common features including maximum likelihood estimates of source given target phrases $p_{MLE}(e|f)$ and vice versa, $p_{MLE}(f|e)$, lexically weighted estimates $p_{LW}(e|f)$ and $p_{LW}(f|e)$, word and phrase-penalties, as well as a linear distortion feature. The baseline uses a hierarchical reordering model with five orientation types, including monotone and swap, described in §2, as well as two discontinuous orientations, distinguishing if the previous phrase is to the left or right of the current phrase. Finally, monotone global indicates that all previous phrases can be combined into a single hierarchical block. The baseline includes a modified Kneser-Ney word-based language model trained on the target-side of the parallel data, which is described below. Log-linear weights are estimated with MERT (Och, 2003). We regard the 1-best output of the phrase-based decoder with the hierarchical reordering model as the baseline accuracy.

Evaluation. We use training and test data from the WMT 2012 campaign and report results on French-English and German-English translation (Callison-Burch et al., 2012). Translation models are estimated on 102M words of parallel data for French-English and 91M words for German-English; between 7.5-8.2M words are newswire, depending on the language pair, and the remainder are parliamentary proceedings. All discriminative reordering models are trained on the newswire subset since we found this portion of the data to be most useful in initial experiments. We evaluate on six newswire domain test sets from 2008, 2010 to 2013 as well as the 2010 system combination test set containing between 2034 to 3003 sentences. Log-linear weights are estimated on the 2009 data set comprising 2525 sentences. We evaluate using BLEU with a single reference.

Discriminative Reordering Model. We use 100-best lists generated by the phrase-based decoder to train the discriminative reordering model. The n-best lists are generated by ten systems, each trained on 90% of the available data in order to decode the remaining 10%. The purpose of this procedure is to avoid a bias introduced by generating n-best lists for sentences on which the translation model was previously trained.[7] Unless otherwise

[5] Phrase-local features allow pre-computation which results in significant speed-ups at run-time. Cherry (2013) shows that local features are responsible for most of his gains.

[6] Although, our model is likely to learn significantly fewer parameters since many phrase, orientation pairs will only be seen in the word-aligned data but not in actual machine translation output.

[7] Later, we found that the bias has only a negligible effect on end-to-end accuracy since we obtained very similar results when decoding with a system trained on all data. This setting increased the training data BLEU score from 27.5 to 37.8. We used a maximum source and target phrase length of 7 words.

	dev	2008	2010	sc2010	2011	2012	2013	AllTest	FeatTypes
noRM	23.37	20.18	24.24	24.18	24.83	24.23	24.85	23.93	-
HRM (baseline)	24.11	20.85	24.92	24.83	25.68	25.11	25.76	24.72	-
SparseHRMLocal	25.24	21.26	25.99	25.93	26.98	26.34	26.77	25.77	4,407
SparseHRM	25.29	21.43	26.17	26.14	26.99	26.63	27.01	25.95	9,463
+UncommonWords	25.32	21.76	26.30	26.29	27.15	26.77	27.18	26.12	897,537
+BiPhrases	25.46	21.67	26.19	26.19	27.55	27.07	27.41	26.26	3,043,053

Table 1: French-English results of expected BLEU trained sparse reordering models compared to no reordering model at all (noRM) and the likelihood trained baseline hierarchical reordering model (HRM) on WMT test sets; sc2010 is the 2010 system combination test set. FeatTypes is the number of different types and AllTest is the average BLEU score over all the test sets, weighted by corpus size. All results for our sparse reordering models include a likelihood-trained hierarchical reordering model.

	dev	2008	2010	sc2010	2011	2012	2013	AllTest	FeatTypes
noRM	18.54	19.28	20.14	20.01	18.90	18.87	21.60	19.81	-
HRM (baseline)	19.35	19.96	20.87	20.66	19.60	19.80	22.48	20.58	-
SparseHRMLocal	19.89	19.86	21.11	20.84	20.04	20.21	22.93	20.88	4,410
SparseHRM	19.83	20.27	21.26	21.05	20.22	20.44	23.17	21.11	9,477
+UncommonWords	20.06	20.35	21.45	21.31	20.28	20.55	23.30	21.24	1,136,248
+BiPhrases	20.09	20.33	21.62	21.47	20.66	20.75	23.27	21.40	3,640,693

Table 2: German-English results of expected BLEU trained sparse reordering models (cf. Table 1).

mentioned, we train our reordering model on the news portion of the parallel data, corresponding to 136K-150K sentences, depending on the language pair. We tuned the various hyper-parameters on a held-out set, including the learning rate, for which we found a simple setting of 0.1 to be useful. To prevent overfitting, we experimented with ℓ_2 regularization, but found that it did not improve test accuracy. We also tuned the probability scaling parameter γ (Eq. 6) but found $\gamma = 1$ to be very good among other settings. We evaluate the performance on a held-out validation set during training and stop whenever the objective changes less than a factor of 0.0003. For our **PRO experiments**, we tuned three hyper-parameters controlling ℓ_2 regularization, sentence-level BLEU smoothing, and length. The latter is important to eliminate PRO's tendency to produce too short translations (Nakov et al., 2012).

7.1 Scaling the Feature Set

We first compare our baseline, a likelihood trained hierarchical reordering model (HRM; Galley & Manning, 2008), to various expected BLEU trained models, starting with SparseHRMLocal, inspired by Cherry (2013) and compare it to SparseHRM+BiPhrases, a set that is three orders of magnitudes larger.

Our results on French-English translation (Table 1) and German-English translation (Table 2) show that the expected BLEU trained models scale to millions of features and that we outperform the baseline by up to 2.0 BLEU on newstest2012 for French-English and by up to 1.1 BLEU on newstest2011 for German-English.[8] Increasing the size of the feature set improves accuracy across the board: The average accuracy over all test sets improves from 1.0 BLEU for the most basic feature set to 1.5 BLEU for the largest feature set on French-English and from 0.3 BLEU to 0.8 BLEU on German-English.[9] The most comparable setting to Cherry (2013) is the feature set SparseHRM, which we outperform by up to 0.5 BLEU on French-English and by 0.3 BLEU on average on both language pairs, demonstrating the benefit of being able to effectively train large feature sets. Furthermore, the increase in the number of features does not affect runtime, since most

[8] Different to the setups of Galley & Manning (2008) and Cherry (2013) our WMT evaluation framework uses only one instead of four references, which makes our BLEU score improvements not directly comparable.

[9] We attribute smaller improvements on German-English to the low distortion limit of only six words of our system and the more difficult reordering patterns when translating from German which may require more elaborate features.

features can be pre-computed and stored in the phrase-table, only requiring a constant time table-lookup, similar to traditional reordering models.

Another appeal of our approach is that training is very fast given a set of n-best lists for the training data. The SparseHRM model with 4,407 features is trained in only 26 minutes, while the SparseHRM+BiPhrases model with over three million parameters can be trained in just over two hours (136K sentences and 100 epochs in both cases). We attribute this to the training regime (§4), which does not iteratively re-decode the training data for expected BLEU training.[10]

7.2 Varying Training Set Size

Previous work on sparse reordering models was restricted to small data sets (Cherry, 2013) due to the limited ability of standard machine translation optimizers to handle more than a few thousand sentences. In particular, recent attempts to scale the margin-infused relaxation algorithm, a variation which was also used by Cherry (2013), to larger data sets showed that more data does not necessarily help to improve test set accuracy for large feature sets (Eidelman et al., 2013).

In the next set of experiments, we shed light on the advantage of training discriminative reordering models with expected BLEU on large training sets. Specifically, we start off by estimating a reordering model on only 2,000 sentences, similar to the size of the development set used by Cherry (2013), and incrementally increase the amount of training data to nearly three hundred thousand sentences. To avoid overfitting to small data sets we experiment with our most basic feature set SparseHRM-Local, comprising of just over 4,400 types.

For this experiment only, we measure accuracy in a *re-ranking* framework for faster experimentation where we use the 100-best output of the baseline system relying on a likelihood-based hierarchical reordering model. We re-estimate the log-linear weights by running a further iteration of MERT on the n-best list of the development set which is augmented by scores corresponding to the discriminative reordering model. The weights of those features are initially set to one and we use 20 random restarts for MERT. At test time we rescore the 100-best list of the test set using the new set of log-linear weights learned previously.

[10] We would expect better accuracy when iteratively decoding the training data but did not do so in this study for efficiency reasons.

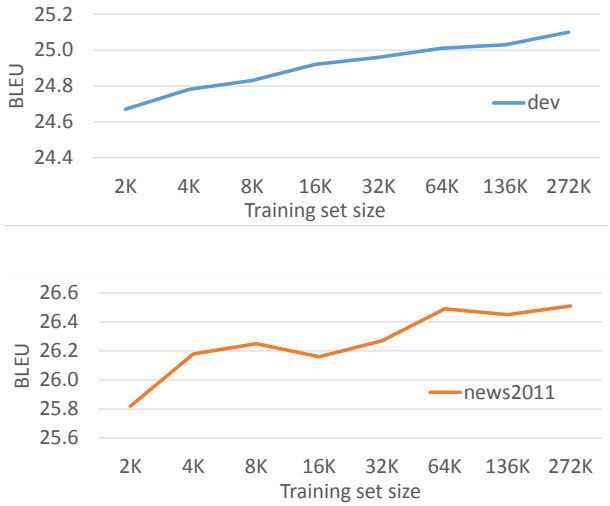

Figure 3: Effect of increasing the training set size from 2,000 to 272,000 sentences measured on the dev set (top) and news2011 (bottom) in an n-best list rescoring setting.

Figure 3 confirms that more training data increases accuracy and that the best model requires a substantially larger amount of training data than what is typically used for maximum BLEU training. We expect an even steeper curve for larger feature sets where more parameters need to be estimated and where the amount of training data is likely to have an even larger effect.

7.3 Likelihood versus BLEU Optimization

Previous research has shown that directly training a reordering model for BLEU can vastly outperform a likelihood trained maximum entropy reordering model (Cherry, 2013). However, the two approaches do not only differ in the objectives used, but also in the type of training data. The maximum entropy reordering model is trained on a word-aligned corpus, trying to learn human reordering patterns, whereas the sparse reordering model is trained on machine translation output, trying to learn from the mistakes made by the actual system. It is therefore not clear how much either one contributes to good accuracy.

Our next experiment teases those two aspects apart and clearly shows the effect of the objective function. Specifically, we compare the traditionally used conditional log-likelihood (CLL) objective to expected BLEU on the French-English translation task in a small feature condition (SparseHRM) of about 9K features and

	dev	2008	2010	sc2010	2011	2012	2013	AllTest
noRM	23.37	20.18	24.24	24.18	24.83	24.23	24.85	23.93
HRM (baseline)	24.11	20.85	24.92	24.83	25.68	25.11	25.76	24.72
SparseHRM (CLL)	24.28	21.02	25.11	25.10	25.92	25.24	25.76	24.88
SparseHRM (xBLEU)	25.29	21.43	26.17	26.14	26.99	26.63	27.01	25.95
SparseHRM+BiPhrases (CLL)	24.42	21.17	25.12	25.00	25.86	25.36	26.18	24.98
SparseHRM+BiPhrases (xBLEU)	25.46	21.67	26.19	26.19	27.55	27.07	27.41	26.26

Table 3: French-English results comparing the baseline hierarchical reordering model (HRM) to sparse reordering model trained towards conditional log-likelihood (CLL) and expected BLEU (xBLEU).

	dev	2008	2010	sc2010	2011	2012	2013	AllTest
PRO	24.05	20.90	25.42	25.28	25.79	25.09	26.07	24.94
xBLEU	25.24	21.26	25.99	25.93	26.98	26.34	26.77	25.77

Table 4: French-English results on the SparseHRMLocal feature set when when trained with pair-wise ranked optimization (PRO) and expected BLEU (xBLEU).

a large feature setting of over 3M features (SparseHRM+BiPhrases). In the CLL setting, we maximize the likelihood of the hypothesis with the highest BLEU score in the n-best list of each training sentence.

Our results (Table 3) show that CLL training achieves only a fraction of the gains yielded by the expected BLEU objective. For SparseHRM, CLL improves the baseline by less than 0.2 BLEU on average across all test sets, whereas expected BLEU achieves 1.2 BLEU. Increasing the number of features to 3M (SparseHRM+BiPhrases) results in a slightly better average gain of 0.3 BLEU for CLL but but expected BLEU still achieves a much higher improvement of 1.5 BLEU. Because our gains with likelihood training are similar to what Cherry (2013) reported for his maximum entropy model, we conclude that the objective function is the most important factor to achieving good accuracy.

7.4 Comparison to PRO

In our final experiment we compare expected BLEU training to pair-wise ranked optimization (PRO), a popular off the shelf trainer for machine translation models with large feature sets (Hopkins and May, 2011).[11] Previous work has shown that PRO does not scale to truly large feature sets with millions of types (Yu et al., 2013) and we therefore restrict ourselves to our smallest set (SparseHRMLocal) of just over 4.4K features. We train PRO on the development set comprising of 2,525 sentences, a setup that is commonly used by standard machine translation optimizers. In this setting, PRO directly learns weights for the baseline features (§7) as well as the 4.4K indicator features corresponding to the sparse reordering model. For expected BLEU training we use the full 136K sentences from the training data. The results (Table 4) demonstrate that expected BLEU outperforms a typical setup commonly used to train large feature sets.

8 Conclusion and Future Work

The expected BLEU objective is a simple and effective approach to train large-scale discriminative reordering models. We have demonstrated that it scales to millions of features, which is orders of magnitudes larger than other modern machine translation optimizers can currently handle.

Empirically, our sparse reordering model improves machine translation accuracy across the board, outperforming a strong hierarchical lexicalized reordering model by up to 2.0 BLEU on a French to English WMT2012 setup, where the baseline was trained on over two million sentence pairs. We have shown that scaling to large training sets is crucial to good performance and that the best performance is reached when hundreds of thousands of training sentences are used. Furthermore, we demonstrate that task-specific training towards expected BLEU is much more effective than optimizing conditional log-likelihood as

[11] MIRA is another popular optimizer but as previously mentioned, even the best publicly available implementation does not scale to large training sets (Eidelman et al., 2013).

is usually done. We attribute this to the fact that likelihood is a strict zero-one loss that does not assign credit to partially correct solutions, whereas expected BLEU does.

In future work we plan to extend expected BLEU training to lattices and to evaluate the effect of estimating weights for the dense baseline features as well. Our current training procedure (Gao and He, 2013; Gao et al., 2014) decodes the training data only once. In future work, we would like to compare this to repeated decoding as done by conventional optimization methods as well as other large-scale discriminative training approaches (Yu et al., 2013). We expect this to yield additional accuracy gains.

Acknowledgements

We would like to thank Arul Menezes and Xiaodong He for helpful discussion related to this work and the three anonymous reviewers for their comments.

References

Léon Bottou. 2004. Stochastic learning. In Olivier Bousquet and Ulrike von Luxburg, editors, *Advanced Lectures in Machine Learning*, Lecture Notes in Artificial Intelligence, pages 146–168. Springer Verlag, Berlin.

Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479, Dec.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 Workshop on Statistical Machine Translation. In *Proc. of WMT*, pages 10–51. Association for Computational Linguistics, June.

Colin Cherry. 2013. Improved Reordering for Phrase-Based Translation using Sparse Features. In *Proc. of NAACL*, pages 9–14. Association for Computational Linguistics, June.

David Chiang, Kevin Knight, and Wei Wang. 2009. 11,001 New Features for Statistical Machine Translation. In *Proc. of NAACL*, pages 218–226. Association for Computational Linguistics, June.

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228.

Vladimir Eidelman, Ke Wu, Ferhan Ture1, Philip Resnik, and Jimmy Lin. 2013. Mr. MIRA: Open-Source Large-Margin Structured Learning on MapReduce. In *Proc. of ACL*, pages 199–204. Association for Computational Linguistics, August.

Michel Galley and Christopher D. Manning. 2008. A Simple and Effective Hierarchical Phrase Reordering Model. In *Proc. of EMNLP*, pages 848–856.

Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. 2004. What's in a translation rule? In *Proc. of HLT-NAACL*, pages 273–280, Boston, MA, USA, May.

Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and Ignacio Thayer. 2006. Scalable Inference and Training of Context-Rich Syntactic Translation Models. In *Proc. of ACL*, pages 961–968, Sydney, Australia, June.

Jianfeng Gao and Xiaodong He. 2013. Training MRF-Based Phrase Translation Models using Gradient Ascent. In *Proc. of NAACL-HLT*, pages 450–459. Association for Computational Linguistics, June.

Jianfeng Gao, Xiaodong He, Scott Wen tau Yih, and Li Deng. 2014. Learning Continuous Phrase Representations for Translation Modeling. In *Proc. of ACL*. Association for Computational Linguistics, June.

Spence Green, Daniel Cer, and Christopher Manning. 2014. An Empirical Comparison of Features and Tuning for Phrase-based Machine Translation. In *Proc. of WMT*. Association for Computational Linguistics, June.

Xiaodong He and Li Deng. 2012. Maximum Expected BLEU Training of Phrase and Lexicon Translation Models. In *Proc. of ACL*, pages 8–14. Association for Computational Linguistics, July.

Mark Hopkins and Jonathan May. 2011. Tuning as ranking. In *Proc. of EMNLP*. Association for Computational Linguistics, July.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In *Proc. of HLT-NAACL*, pages 127–133, Edmonton, Canada, May.

Philipp Koehn, Amittai Axelrod, Alexandra Birch Mayne, Chris Callison-Burch, Miles Osborne, and David Talbot. 2005. Edinburgh System Description for the 2005 IWSLT Speech Translation Evaluation. In *Proc. of IWSLT*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proc. of ACL Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, Jun.

Percy Liang, Alexandre Bouchard-Côté, Ben Taskar, and Dan Klein. 2006. An end-to-end discriminative approach to machine translation. In *Proc. of ACL-COLING*, pages 761–768, Jul.

Preslav Nakov, Francisco Guzman, and Stephan Vogel. 2012. Optimizing for Sentence-Level BLEU+1 Yields Short Translations. In *Proc. of COLING*. Association for Computational Linguistics.

Vinh Van Nguyen, Akira Shimazu, Minh Le Nguyen, and Thai Phuong Nguyen. 2009. Improving A Lexicalized Hierarchical Reordering Model Using Maximum Entropy. In *MT Summit XII*. Association for Computational Linguistics, August.

Franz Josef Och and Hermann Ney. 2004. The alignment template approach to machine translation. *Computational Linguistics*, 30(4):417–449, June.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proc. of ACL*, pages 160–167, Sapporo, Japan, July.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. of ACL*, pages 311–318, Philadelphia, PA, USA, Jul.

Antti-Veikko I Rosti, Bing Zhang, Spyros Matsoukas, and Richard Schwartz. 2010. BBN System Description for WMT10 System Combination Task. In *Proc. of WMT*, pages 321–326. Association for Computational Linguistics, July.

Antti-Veikko I Rosti, Bing Zhang, Spyros Matsoukas, and Richard Schwartz. 2011. Expected BLEU Training for Graphs: BBN System Description for WMT11 System Combination Task. In *Proc. of WMT*, pages 159–165. Association for Computational Linguistics, July.

Christoph Tillmann. 2003. A Unigram Orientation Model for Statistical Machine Translation. In *Proc. of NAACL*, pages 106–108. Association for Computational Linguistics, June.

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

Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–403.

Deyi Xiong, Qun Liu, and Shouxun Lin. 2006. Maximum entropy based phrase reordering model for statistical machine translation. In *Proc. of ACL-COLING*, pages 521–528, Sydney, Jul.

Heng Yu, Liang Huang, Haitao Mi, and Kai Zhao. 2013. Max-Violation Perceptron and Forced Decoding for Scalable MT Training. In *Proc. of EMNLP*, pages 1112–1123. Association for Computational Linguistics, October.