Discriminative Reordering Model Adaptation via Structural Learning

Biao Zhang[1], **Jinsong Su**[1]*, **Deyi Xiong**[2], **Hong Duan**[1] and **Junfeng Yao**[1]
Xiamen University, Xiamen, China 361005[1]
Soochow University, Suzhou, China 215006[2]
zb@stu.xmu.edu.cn, {jssu, hduan, yao0010}@xmu.edu.cn
dyxiong@suda.edu.cn

Abstract

Reordering model adaptation remains a big challenge in statistical machine translation because reordering patterns of translation units often vary dramatically from one domain to another. In this paper, we propose a novel adaptive discriminative reordering model (DRM) based on *structural learning*, which can capture correspondences among reordering features from two different domains. Exploiting both in-domain and out-of-domain monolingual corpora, our model learns a shared feature representation for cross-domain phrase reordering. Incorporating features of this representation, the DRM trained on out-of-domain corpus generalizes better to in-domain data. Experiment results on the NIST Chinese-English translation task show that our approach significantly outperforms a variety of baselines.

1 Introduction

Reordering model, as one of essential components of statistical machine translation (SMT), has become an active area of research in recent years [Zens and Ney, 2006; Xiong *et al.*, 2006; Galley and Manning, 2008; Feng *et al.*, 2013; Li *et al.*, 2013; Alrajeh and Niranjan, 2014]. Particularly, discriminative reordering model (DRM) has attracted wide attention due to its advantage of easy fusion of various reordering features. However, similar to other discriminative models, a DRM trained on out-of-domain (e.g. *newswire*) corpus often performs badly on in-domain (e.g. *weblog*) data. This is because reordering patterns of translation units differs dramatically from one domain to another [Chen *et al.*, 2013]. Therefore, building an adaptive DRM is crucial for SMT systems in practice.

A great variety of approaches have been proposed for language and translation model adaptation [Ueffing *et al.*, 2007; Koehn and Schroeder, 2007; Bertoldi and Federico, 2009; Matsoukas *et al.*, 2009; Foster *et al.*, 2010; Axelrod *et al.*, 2011; Phillips and Brown, 2011; Su *et al.*, 2012; Sennrich, 2012; Duh *et al.*, 2013; Cui *et al.*, 2013; Chen *et al.*, 2013; Hoang and Sima'an, 2014]. However, study on reordering model adaptation is very limited. In this respect, Chen et al. [2013] employed a linear mixture model that is trained on multiple parallel training corpus. Although the model achieves a significant improvement, they mainly focus on the exploitation of parallel corpora, of which many language pairs are short. They do not utilize monolingual corpora that are easier to be accessible.

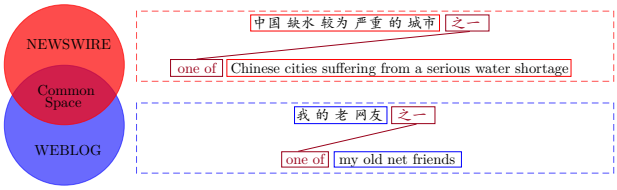

Figure 1: Reordering examples of "*one of*(target side)" and "之一(source side)" for structural learning in domain *newswire* and domain *weblog*, indicated in red and blue color respectively. The ellipse rendered with purple color shows the common space that both domains share, and the phrases with purple color have similar reordering patterns in both domains.

Although reordering patterns and features differ in different domains, some of them are shared across domains. As illustrated in Figure 1, reordering features in domain *weblog* (e.g. *net friends*) differ extremely from those in domain *newswire* (e.g. *water shortage*). If we train a reordering model with the latter features, the performance in the former feature space is expected to be poor, since many features are not observed. Despite of this domain gap, the feature "*one of*" on the target side is a strong indicator for *inverted* orientation in both domains. To some extent, these domain-independent features can bridge the gap between different domains, from which a shared reordering sub-space can be induced, as shown by "*Common Space*" in Figure 1. Exploiting these domain-independent features (which we call *pivot features* following previous work) will, no doubt, be beneficial for reordering model adaptation.

In this paper, we propose a novel adaptive DRM for phrase-based SMT based on *structural learning* [Ando and Zhang, 2005; Blitzer *et al.*, 2006]. Different from previous work, we directly exploit correlations of reordering features from in-domain and out-of-domain monolingual corpora, and further learn a common feature representation for phrase reorder-

*Corresponding author.

ing. Incorporating features of this common representation, the DRM trained on out-of-domain corpus is able to generalize better to in-domain data. Particularly,

- First, we select domain-shared pivot features according to two criteria: 1) frequency in monolingual corpora of different domains, and 2) information gain ratio in out-of-domain parallel corpus;
- Second, we construct reordering-related auxiliary problems with pivot features, and learn a transformation matrix to common space from monolingual corpora using structural learning;
- Finally, we train a new reordering model on out-of-domain bilingual corpora with the combination of original and transformed features.

We incorporate the adaptive reordering model into SMT and conduct experiments on the NIST Chinese-English translation task to evaluate its effectiveness. Results on the NIST 08 web part show that our model achieves a significant improvement over the baseline methods. The main contributions of our work are in two folds:

- We employ structural learning (Section 3.2) to DRM, and propose reordering-specific *pivot feature selection* (Section 3.1) and *predictor construction* (Section 3.1) methods using large-scale monolingual domain data. The utilization of structural learning to DRM, to the best of our knowledge, has never been investigated before.
- We conduct a series of experiments to demonstrate the effectiveness of our model and investigate the influence of various factors on translation quality (See Section 4).

2 Maximum Entropy Based Reordering Model

Many researchers have introduced different DRMs for phrase reordering [Xiong *et al.*, 2006; He *et al.*, 2010; Huck *et al.*, 2012; Alrajeh and Niranjan, 2014]. Without loss of generality, here we focus on the maximum entropy based reordering model (MERM) proposed by Xiong et al. [2006].

Generally, MERM considers phrase reordering as a classification problem, under bracketing transduction grammar (BTG) [Wu, 1997]. There are three types of rules in BTG:

$$A \rightarrow [A^1, A^2] \quad (1)$$

$$A \rightarrow \langle A^1, A^2 \rangle \quad (2)$$

$$A \rightarrow f/e \quad (3)$$

where reordering rules (1) and (2) are used to merge two neighboring blocks in a straight or inverted order, respectively. Lexical rule (3) is used to translate a source phrase f into a target phrase e. Correspondingly, there are only two labels, $o \in \{straight, inverted\}$, in MERM.

Different from relative-frequency based approaches which bind reorderings to individual concrete phrases, MERM is able to integrate arbitrary and overlapping features extracted from data. The probability that two neighboring blocks A^1 and A^2 are combined in the orientation o is computed as follows:

$$p(o|A^1, A^2) = \frac{exp\left(\sum_i \theta_i h_i(o, A^1, A^2)\right)}{\sum_{o'} exp\left(\sum_i \theta_i h_i(o', A^1, A^2)\right)} \quad (4)$$

where $h_i \in \{0, 1\}$ are the feature functions and θ_i are weights of the model features. In our paper, we investigate three kinds of features, all of which are located at the boundaries of A^1 and A^2 following Xiong et al. [2006]. Specifically, for each block, we define the following features:

- **N-gram:** unigrams and bigrams on the left and right boundary of a block[1].
- **Word Class:** word classes of unigrams.
- **PoS:** part of speech tags of unigrams (source side only)

Consider the phrase "*my old net friends*" in Figure 1. N-gram features are given here: A^2.Left.*my*, A^2.Left.*my_old*, A^2.Right.*friends*, A^2.Right.*net_friends*, where *Left* and *Right* indicate the left-side and right-side boundary, respectively.

Despite the ability of incorporating arbitrary features, this model often performs badly on the test data that are different from training data in terms of their domains. We will discuss how to address this issue with structural learning in the next section.

3 Our Model

Given an out-of-domain parallel corpus consisting of monolingual corpora from the source language $\mathcal{D}^{out}_{m,f}$ and the target language $\mathcal{D}^{out}_{m,e}$, and in-domain monolingual corpora for the source language $\mathcal{D}^{in}_{m,f}$ and the target language $\mathcal{D}^{in}_{m,e}$, we aim at learning a common feature representation that is meaningful across different domains using only unlabeled data from both in-domain and out-of-domain corpora. The key idea is to leverage monolingual corpora to generate auxiliary problems for each language that are useful for discovering important correspondences among reordering features from different domains. Intuitively, if these created auxiliary problems are similar or at least related to the reordering problem, we will benefit from solving them.

Let us revisit the example shown in Figure 1. The reordering instance in red dash box comes from *newswire* and blue dash box *weblog*. Although the context vary greatly, the target-side bigram "*one of*" strongly indicates that the neighbor blocks are in *inverted* order in whichever domain. Intuitively, these features connect domain-specific features. Therefore we would like to learn these domain-independent features. It is worth noting that these features can be learned from unlabeled monolingual data, thus breaking down the bilingual corpus barrier.

We first elaborate auxiliary problem in Section 3.1. Then, we detail structural learning in Section 3.2, and present how we learn an adaptive DRM via structural learning.

Due to the space limit, we only describe the learning procedure for the target language (English), omitting the procedure for the source language (Chinese), which can be implemented in a similar way.

3.1 Auxiliary Classification Problems

Pivot Feature Selection

Pivot features are features that occur frequently and behave similarly in corpora of different domains [Blitzer *et al.*, 2006].

[1] A *unigram* refers to a word and a *bigram* refers to two adjacent words.

Chinese	English
A^2.Left.先生, A^2.Left.内_的	A^2.Right.of, A^1.Right.one_of
A^1.Left.之间, A^2.Left.之一	A^1.Left.of_the, A^1.Left.to_this
A^2.Left.女士, A^2.Left.中_所	A^2.Right.end_of, A^2.Right.after

Table 1: Some top pivot features ranked by IGR for both the source (Chinese) and target (English) languages.

The selection of pivot features is a crucial step in structural learning, since they are used to establish the correlation between two different domains. Similar to Blitzer et al. [2006] and Prettenhofer and Stein [2010], we select pivot features according to the following two criteria:

- Pivot features should occur frequently in both in-domain and out-of-domain corpora;
- Pivot features should be useful for reordering prediction.

Although pivot features can be n-grams, word classes or PoS tags, we only focus on n-gram features in this paper. In order to meet the first criterion, we extract frequent unigrams and bigrams from $\mathcal{D}^{in}_{m,e}$ and $\mathcal{D}^{out}_{m,e}$ with a frequency threshold ϵ_1, and annotate them with corresponding positions (A^2.*Right*, A^2.*Left*, A^1.*Right*, A^1.*Left*). These annotated n-gram features are used as our candidate pivot features. They ensure that correlations between different domains can be estimated accurately. With regard to the second criterion, we calculate *information gain ratio* (IGR) for candidate features that satisfy the first criterion on reordering examples extracted from the out-of-domain parallel corpus and select top m candidates as final pivot features according to their IGR values. This way can ensure that only those correlations that are useful for discriminative learning will be modeled.

Each pivot feature for reordering consists of an n-gram (e.g. *one_of*) and a position (e.g. A^1.*Right*). We give some top pivot features selected in our model in Table 1. Most of them, as expected, are function words that often encode grammatical relations among phrases within a sentence and greatly affect phrase reordering. For example, the target-side feature "*one of*" with position A^1.*Right* in Figure 1 satisfies the two characteristics mentioned above: (1) it strongly indicates the *inverted* reordering orientation, and (2) it occurs frequently in both *newswire* and *weblog* domain. Thus, it is an ideal pivot feature for order predicting.

Pivot Feature Predictors

After obtaining pivot features, we conduct an auxiliary classification task on monolingual corpora $\mathcal{D}^{in}_{m,e}$ and $\mathcal{D}^{out}_{m,e}$. For each pivot feature f_p, we construct a binary classifier to predict whether it occurs in a reordering instance. Since reordering instances in SMT are extracted from parallel corpus, a serious challenge here is to extract related pivot features as well as reordering-specific training examples from monolingual corpus.

To tackle this challenge, we adopt an imitation strategy to extract *pseudo reordering instances* from monolingual corpora according to their reordering behaviors. Basically, for a monolingual sentence, we try to select two neighboring phrases within a limited window size s_l (left phrase) and s_r (right phrase) to simulate the two adjacent blocks in reordering. If a sentence contains the n-gram in f_p, we directly select neighboring phrases according to the n-gram (e.g. *one of*) and position (e.g. A^1.*Right*) in f_p, and treat this instance as a positive instance. An example is illustrated in Figure 2. Suppose that f_p is "A^1.Right.*one_of*", we will choose the third instance as our positive one.

A^2.Right: ... this ⟨place has⟩⟨become one of⟩ the mainland 's ...
A^2.Left: ... this place ⟨has become⟩⟨one of the⟩ mainland 's ...
A^1.Right: ... this place has ⟨become one of⟩⟨the mainland⟩ 's ...
A^1.Left: ... this place has become ⟨one of the⟩⟨mainland 's⟩ ...

Figure 2: Training instance extraction from example sentence "*Today this place has become one of the mainland 's largest ocean-going container freight ports .*" for pivot n-gram "*one of*". Each instance corresponds to a pivot label, which is A^2.*Right*, A^2.*Left*, A^1.*Right*, A^1.*Left* from top to bottom. $\langle\cdot\rangle$ represents a monolingual phrase, and we set $s_l = s_r = 2$ for better illustration.

For a sentence $\mathbb{S}$ without the n-gram in f_p, we construct a *pseudo pivot feature* f'_p with the position of f_p and an n-gram chose from $\mathbb{S}$. A negative instance is further extracted from $\mathbb{S}$ with f'_p. In regard to the n-gram, we first extract a frequent n-gram set $\mathcal{S}^{in}$ and $\mathcal{S}^{out}$ from $\mathcal{D}^{in}_{m,e}$ and $\mathcal{D}^{out}_{m,e}$ respectively according to a threshold ϵ_2. Afterwards we randomly pick an n-gram from $\mathcal{A}$: the intersection of $\mathbb{S}$ and $\mathcal{S}^{in}$ - $\mathcal{S}^{out}$ for in-domain sentence (or $\mathbb{S}$ and $\mathcal{S}^{out}$ - $\mathcal{S}^{in}$ for out-of-domain sentence). If $\mathcal{A}$ is empty, we choose the n-gram randomly from $\mathcal{B}$: the intersection of $\mathbb{S}$ and $\mathcal{S}^{in} \cap \mathcal{S}^{out}$. If $\mathcal{B}$ is also empty, we simply select an n-gram randomly from $\mathbb{S}$. Thus, the selected n-gram will be not only frequent, but also domain-specific, which ensures the accuracy of learned correlations.

Features for a pivot predictor are the same as those in MERM defined in Section 2, except for the pivot feature itself. In practice, negative and positive training examples are extremely imbalanced. To overcome this problem, we randomly select training instances to approximately make the ratio of negative to positive examples 2:1, as implemented in [Hernault *et al.*, 2011].

3.2 Adaptive DRM Based on Structural Learning

Assume that we have a training set consisting of T reordering examples $\{h_t, o_t\}_{t=1}^{T}$, where h_t is a reordering feature vector with dimension d and orientation o_t; we also have a large-scale unlabeled corpus $\{h_j\}$ collected from $\mathcal{D}^{out}_{m,e}$ and $\mathcal{D}^{in}_{m,e}$, where $h_j \in \mathcal{R}^d$ is the feature vector extracted from monolingual sentence. Based on structural learning, we learn an adaptive DRM in the following steps.

In **Step 1**, we first choose m pivot features which have similar effects on phrase movements in different domains (Section 3.1), and then define a predictor for each pivot feature based on simple linear classifier (Section 3.1). For each predictor p_l, $l = 1 \ldots m$, we learn the optimal weight vector $\hat{w}_l$ in the following way

$$\hat{w}_l = \text{argmin}_w \left(\sum_j L(p_l(h_j), w \cdot h_j) + \lambda \|w\|^2 \right) \quad (5)$$

where L is a loss function and λ a regularization coefficient, and $p_l(h_j)$ indicates the binary label for h_j.

In **Step 2**, through the above-mentioned predictors, we induce a mapping θ from the original reordering feature spaces of both domains to a shared, low-dimensional real-valued feature space. In the specific implementation, we first create a matrix $W = [\hat{w}_1 \dots \hat{w}_m]$ in a column-per-column fashion, and then perform a singular value decomposition (SVD) on this matrix to reduce its dimension. In this way, we can learn a compact representation of the space of auxiliary classifiers.

Note that typically DRM employ several types of heterogeneous features, such as phrases, word classes and PoS tags. Hence we perform a localized dimension reduction for each type of feature. For each feature type $\mathbf{f}_i$, $i \in [1 \dots n]$ with start position s_i and end position e_i in the feature space, we create a feature-type-specific structural parameter matrix θ_i so that,

$$U_i, D_i, V_i^T = SVD(W_{[s_i:e_i,:]}) \quad (6)$$

$$\theta_i = U_i^T{}_{[1:k,:]} \quad (7)$$

where $W_i = U_i D_i V_i^T$. Assume that the diagonal elements of D_i are arranged in decreasing order. The rows of θ_i are given by the first k rows of U_i^T, which are the left singular vectors corresponding to the largest k singular values of W_i. Therefore, the complete structural parameter matrix $\theta = [\theta_1 \dots \theta_n]$ has dimension $k \times d$, and it encodes the structure learned by the auxiliary tasks in a low-dimensional common space.

In **Step 3**, both the original features h and transformed features θh are used simultaneously to construct the reordering classifier. We project each training and test feature vector of DRM onto θ, and obtain a set of k new structural features, which are then appended to their original feature vector. Formally, we extend the feature set in the following way:

$$\{(\begin{bmatrix} h_t \\ \theta h_t \end{bmatrix}, o_t)_{t=1}^T\} \quad (8)$$

In **Step 4**, we rescale extended features to unit length, $\frac{\theta h_t}{\|\theta h_t\|}$.

Compared with DRM trained only with original features, the model trained on adjusted feature set would put much more emphasis on the transformed features. Because these features are shared across domains, our model will benefit more from them and generalize better when transferring to in-domain data. Note that in our model, we have two kinds of transformed features (from the source and target language respectively). We empirically concatenate them together with the original features h into a single feature vector, which is feed into the MERM for reordering.

4 Experiments

We conduct experiments on the NIST Chinese-to-English translation task to evaluate our model. After a brief description of the data preparation and the experimental setup, we first studied the effects of various factors on the reordering classification results, and then investigated the effectiveness of our method.

Data	Sent#	Word#
dev	1048/4192	22.4K/86.3K
tst	666/2664	30.3K/62.4K
train(out)	1M	25.2M/29M
train(in)	6.5M/4.2M	159.9M/69.6M

Table 2: Statistics of the experiment data sets. **dev**=development set, **tst**=test set, **train(out)**=out-of-domain parallel corpus, and **train(in)**=in-domain monolingual corpora. Note that in **dev** and **tst**, each source sentence has four reference translations.

4.1 Data Preparation

The out-of-domain (*newswire*) training corpus comes from the FBIS corpus and Hansards part of LDC2004T07 corpus. We used the Chinese Sohu weblog in 2009[2] and the English Blog Authorship corpus[3] as the in-domain (*weblog*) monolingual corpora in the source language and target language, respectively. For these monolingual data, we first preprocessed them with Stanford NLP toolkit[4], and then filtered noisy blog documents and those consisting of short sentences.

Because some important parameters such as the pivot feature number m and reduced dimension k directly influence the performance of our approach, we used the web part of NIST 06 MT evaluation test data as our development set to obtain the optimal parameters, and the web part of the 2008 NIST MT test data, as the final test set. The statistics of the various data sets are shown in Table 2.

4.2 Setup

We word-aligned the training corpus using *GIZA++*[5] with the option "*grow-diag-final-and*". For the parallel corpus and monolingual corpora, we obtained word classes using the *mkcls* tool[6] with 50 classes. After getting the original features and the transformed features, we used the binary logistic regression (maximum entropy model) implemented in the *Classias* toolkit[7] to train the MERM.

We used *SRILM* toolkit[8] to train a 4-gram language model on the Xinhua portion of Gigaword corpus. Our decoder is a state-of-the-art SMT system which adapts BTG to phrasal translation and equips itself with a MERM [Xiong *et al.*, 2006]. During decoding, we performed *minimum-error-rate* (MERT) training to tune feature weights of the log-linear model [Och and Ney, 2003]. The translation quality is evaluated by case-sensitive BLEU-4 [Papineni *et al.*, 2002] and NIST [Doddington, 2002]. Finally, we conducted paired bootstrap sampling [Koehn, 2004] to test the significance in BLEU score differences.

We chose the modified Huber loss [Ando and Zhang, 2005] as the loss function, and used stochastic gradient descent

[2] http://blog.sohu.com/

[3] http://u.cs.biu.ac.il/ koppel/BlogCorpus.html

[4] http://nlp.stanford.edu/software/corenlp.shtml

[5] http://www.fjoch.com/GIZA++.html

[6] http://www-i6.informatik.rwthaachen.de/Colleagues/och/software/mkcls.html

[7] http://www.chokkan.org/software/classias/

[8] http://www.speech.sri.com/projects/srilm/download.html

Length		2	3	4	5	6	7	8	9	10	11	12	13	14	15
dev	straight#	23474	11406	5269	3188	3188	1982	1638	1403	1028	804	764	806	526	498
	inverted#	980	1223	1353	1299	1213	939	1018	682	558	436	416	309	146	174
tst	straight#	16262	8564	4136	2784	1722	1572	1397	1089	967	714	622	593	439	338
	inverted#	836	1143	1084	925	739	783	536	411	310	215	150	120	100	134

Table 3: Phrase orientation distribution over source-side lengths of reordering examples extracted from **dev** (the development set) and **tst** (the test set).

(SGD) algorithm to train auxiliary predictors. Following Prettenhofer and Stein [2010], we empirically set the number of iterations and the regularization parameter as 10^6 and 10^{-5} for SGD, respectively. We empirically set $\epsilon_1 = \epsilon_2 = 1000$, $s_l = s_r = 3$, and used the Lanczos algorithm implemented by *SVDLIBC*[9] to compute the SVD of the dense parameter matrix W, similar to Blitzer et al. [2006]. In this process, negative values in W were set as 0 to yield a sparse representation.

4.3 Result and Analysis

Various Factors on Classification Results

From **Step 2** in Section 3.2, we know that the pivot feature number m and reduced dimensionality k play important roles in our method. The former determines the quality of the correlation between different domains, and the latter controls the representation of transformed features. In this section, we try different settings for these parameters based on experiments of reordering classification on the development set.

In our implementation, we first word-aligned all 4192 parallel sentences in the development set together with original training corpus, and then extracted reordering examples with their source-side lengths ranging from 2 to 15. Table 3 summaries distribution of phrase orientations (*straight* or *inverted*) over different lengths. From development set, we find that: 1) all distributions of the straight and inverted orientation are imbalanced, especially when the length is small; 2) the number of instances shrink with the growth of length. Inspired by these observations, we chose F-measure as our metric for reordering classification (phrase orientation prediction) accuracy, and conducted experiments separately on each length for classification.

We trained two reordering models: one is the conventional model with only original reordering features; the other is the improved model with the original and transformed reordering features. Considering the effect of different factors, we implemented our approach using different numbers of pivot features (from 100 to 400 with an increment of 100 each time) and reduced dimensions (from 60 to 120 with an increment of 20 each time) following previous work on structural learning [Blitzer *et al.*, 2006; 2007; Prettenhofer and Stein, 2010].

Results are shown in Figure 3, from which we observe that the utilization of the transformed features can improve reordering significantly under most settings. The F-measure for both our model and baseline decreases as phrase length increases. This indicates that the reordering problem is more serious for long phrases. However, compared with the baseline, our adaptive model falls more slowly, and maintains better generalization in domain *weblog*. Intuitively, our model

[9]http://tedlab.mit.edu/~dr/SVDLIBC/

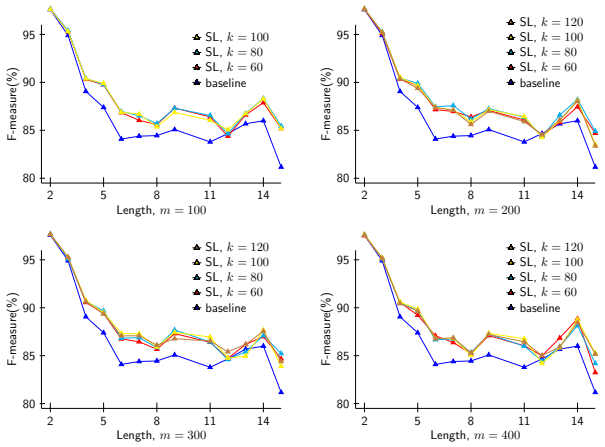

Figure 3: Reordering classification (orientation prediction) accuracy (F-measure) on the development set with different numbers of pivot features m and reduced dimensions k. **SL**=our model with structural learning. Note that there are no $k = 120$ when $m = 100$, since $k \leq m$.

System	m	k (F-measure%)			
		60	80	100	120
Baseline	-	92.88			
Ours	100	93.63	93.70	93.65	-
	200	93.61	93.75	93.69	93.62
	300	93.61	93.67	93.72	93.69
	400	93.52	93.57	93.65	93.61

Table 4: Reordering classification accuracy (F-measure) on the development set with all-length instances, and different numbers of pivot features m and reduced dimensions k

performs better than the baseline when phrase length ranges from 5 to 11.

We further show overall performance of all reordering instances in Table 4. On the one hand, our model is consistently better than the baseline. The absolute improvement is small, this is mainly due to the large proportion of short phrases of which the F-measure in the baseline is already very high (over 90%). On the other hand, we do not see consistent improvement with the growth of the pivot feature number and dimensionality. This suggests that our model is insensitive to the change of m and k in a relatively large range. Such insensitivity resonates with the finding of Ando and Zhang [2005].

Roughly, our model performs well when the pivot feature number ranges from 200 to 300, and the dimensionality from 80 to 100. Considering the tradeoff between performance

and efficiency, we conducted SMT experiments setting the reduced dimension to 80 and the pivot feature number to 200.

Translation Results

With the optimal hyper-parameter configuration found on the development set, we evaluated our method on the Chinese-to-English translation task on the test set. In addition to to baseline system, we also compared against another two methods that also focus on the exploitation of monolingual resources.

- **Additional Language Model:** Since a huge amount of in-domain monolingual corpora are available, we trained an additional language model with this monolingual resource, and expected better adaptation of this additional language model to the in-domain data.
- **Transductive Learning Model:** Inspired by the *transductive learning* method [Ueffing *et al.*, 2007], we retrained our reordering model. We used importance sampling to select 10,0000 sentence pairs over 20-best lists, based on length-normalized sentence scores.

Table 5 summaries the results. The enhanced system consistently outperforms the baselines. Using our method, the BLEU/NIST scores of the SMT system are 17.71/6.2121, which obtain 1.59/0.6914, 0.41/0.4628 and 0.76/0.4234 Bleu/NIST points over the baseline system, the system with additional language model and the system using transductive learning, respectively. Our approach outperforms the other three methods, and all these improvements are statistically significant (p<0.01, p<0.05 and p<0.01, respectively) verified by the significance test tool developed by Zhang et al. [2004].

We further aligned all 2664 parallel sentences on the test set together with the original training corpus, and extracted reordering examples to see how different models perform. The data statistics is also given in Table 3, and the classification result is shown in Figure 4. Our model outperforms both the baseline and TLM, and performs well for long phrase reordering prediction. During decoding, our model will greatly benefit from the higher prediction accuracy of long phrase reordering. This further demonstrates the advantage of our model over baselines in that our model can generalize better to in-domain data.

System	Tst	
	BLEU	NIST
Baseline	16.12	5.5207
ALM	17.30	5.7493
TLM	16.95	5.7887
Our Model	17.71	6.2121

Table 5: BLEU and NIST scores on the test set under the condition m=200, k=80. **ALM**= additional language model, and **TLM**=transductive learning model.

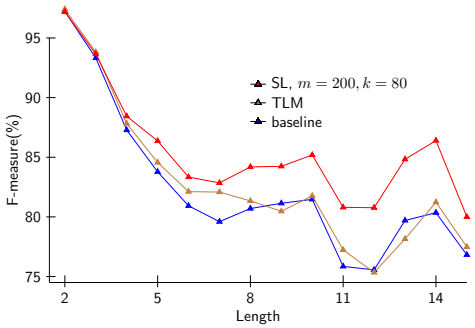

Figure 4: Reordering classification F-measure on the test set with all-length instances.

5 Related Work

Domain adaptation for SMT has attracted considerable attentions in recent years. Previous studies have mainly focused on the adaptation of translation model and language model, while we pay attention to the adaptation of a discriminative reordering model, which has attracted little attention before.

Previous methods either focus on the collection of in-domain bilingual data, or exploit relations between different domains. Some researchers use *transductive learning* to translate in-domain monolingual data with an SMT system, and use these translated data as additional training data [Ueffing *et al.*, 2007; Bertoldi and Federico, 2009]. Data selection approaches [Axelrod *et al.*, 2011; Duh *et al.*, 2013] are also used to search for bilingual sentence pairs that are similar to the in-domain data set. The key idea of these approaches is to create in-domain sentence pairs for SMT systems.

Different from the above-mentioned work, many other researchers focus on the studies of assigning appropriate weights to sentences or phrase pairs from training corpus [Matsoukas *et al.*, 2009; Foster *et al.*, 2010; Sennrich, 2012; Hoang and Sima'an, 2014]. With respect to the adaptation of reordering model, Chen et al. [2013] adopts linear mixture model technique. The correlations between different domains are explored in their reordering model, which, different from ours, is based on parallel corpora.

Compared to the above approaches, we directly exploit correlations between different domains by utilizing monolingual data without collecting parallel corpus. The idea of our method is related to Cui et al. [2013], who explore multi-task learning to jointly adapt SMT models to multiple domains by leveraging the common knowledge shared by different domains. The difference is that we focus on the adaptation of reordering model, and adapt structural learning to DRM.

6 Conclusion

In this paper, we present a structural learning based adaptive DRM for phrase-based SMT. We first apply structural learning on the monolingual data from different domains to obtain correlations between reordering features of different domains. A MaxEnt-based DRM is trained on the out-of-domain corpus that incorporates the learned correlation information. Experiment results have shown that our approach achieves significant improvements over a variety of baselines.

In the future, we plan to continue our work in the following aspects. First, since our approach is general to other DRMs, we will apply it on other SMT systems. Second, we will explore better methods to optimize important parameters such as the pivot feature number and the reduced dimension.

Acknowledgments

The authors were supported by National Natural Science Foundation of China (Grant Nos 61303082 and 61403269), Natural Science Foundation of Jiangsu Province (Grant No. BK20140355), Natural Science Foundation of Fujian Province (Grant No. 2013J01250), Research Fund for the Doctoral Program of Higher Education of China (No. 20120121120046), and the Special and Major Subject Project of the Industrial Science and Technology in Fujian Province 2013 (Grant No. 2013HZ0004-1), and 2014 Key Project of Anhui Science and Technology Bureau (Grant No. 1301021018). We also thank the anonymous reviewers for their insightful comments.

References

[Alrajeh and Niranjan, 2014] Abdullah Alrajeh and Mahesan Niranjan. Large-scale reordering model for statistical machine translation using dual multinomial logistic regression. In *Proc. EMNLP*, 2014.

[Ando and Zhang, 2005] Rie Kubota Ando and Tong Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *J. Mach. Learn. Res.*, 2005.

[Axelrod *et al.*, 2011] Amittai Axelrod, Xiaodong He, and Jianfeng Gao. Domain adaptation via pseudo in-domain data selection. In *Proc. EMNLP*, 2011.

[Bertoldi and Federico, 2009] Nicola Bertoldi and Marcello Federico. Domain adaptation for statistical machine translation with monolingual resources. In *Proc. StatMT*, 2009.

[Blitzer *et al.*, 2006] John Blitzer, Ryan McDonald, and Fernando Pereira. Domain adaptation with structural correspondence learning. In *Proc. EMNLP*, 2006.

[Blitzer *et al.*, 2007] John Blitzer, Mark Dredze, and Fernando Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proc. ACL*, 2007.

[Chen *et al.*, 2013] Boxing Chen, George Foster, and Roland Kuhn. Adaptation of reordering models for statistical machine translation. In *Proc. NAACL-HLT*, 2013.

[Cui *et al.*, 2013] Lei Cui, Xilun Chen, Dongdong Zhang, Shujie Liu, Mu Li, and Ming Zhou. Multi-domain adaptation for SMT using multi-task learning. In *Proc. EMNLP*, 2013.

[Doddington, 2002] George Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proc. HLT*, 2002.

[Duh *et al.*, 2013] Kevin Duh, Graham Neubig, Katsuhito Sudoh, and Hajime Tsukada. Adaptation data selection using neural language models: Experiments in machine translation. In *Proc. ACL*, 2013.

[Feng *et al.*, 2013] Minwei Feng, Jan-Thorsten Peter, and Hermann Ney. Advancements in reordering models for statistical machine translation. In *Proc. ACL*, 2013.

[Foster *et al.*, 2010] George Foster, Cyril Goutte, and Roland Kuhn. Discriminative instance weighting for domain adaptation in statistical machine translation. In *Proc. EMNLP*, 2010.

[Galley and Manning, 2008] Michel Galley and Christopher D. Manning. A simple and effective hierarchical phrase reordering model. In *Proc. EMNLP*, 2008.

[He *et al.*, 2010] Zhongjun He, Yao Meng, and Hao Yu. Maximum entropy based phrase reordering for hierarchical phrase-based translation. In *Proc. EMNLP*, 2010.

[Hernault *et al.*, 2011] Hugo Hernault, Danushka Bollegala, and Mitsuru Ishizuka. Semi-supervised discourse relation classification with structural learning. In *Proc. CICLing*. 2011.

[Hoang and Sima'an, 2014] Cuong Hoang and Khalil Sima'an. Latent domain phrase-based models for adaptation. In *Proc. EMNLP*, 2014.

[Huck *et al.*, 2012] Matthias Huck, Stephan Peitz, Markus Freitag, and Hermann Ney. Discriminative reordering extensions for hierarchical phrase-based machine translation. In *Proc. EAMT*, 2012.

[Koehn and Schroeder, 2007] Philipp Koehn and Josh Schroeder. Experiments in domain adaptation for statistical machine translation. In *Proc. StatMT*, 2007.

[Koehn, 2004] Philipp Koehn. Statistical significance tests for machine translation evaluation. In *Proc. EMNLP*, 2004.

[Li *et al.*, 2013] Peng Li, Yang Liu, and Maosong Sun. Recursive autoencoders for ITG-based translation. In *Proc. EMNLP*, 2013.

[Matsoukas *et al.*, 2009] Spyros Matsoukas, Antti-Veikko I. Rosti, and Bing Zhang. Discriminative corpus weight estimation for machine translation. In *Proc. EMNLP*, 2009.

[Och and Ney, 2003] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Comput. Linguist.*, 2003.

[Papineni *et al.*, 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proc. ACL*, 2002.

[Phillips and Brown, 2011] Aaron B. Phillips and Ralf D. Brown. Training machine translation with a second-order taylor approximation of weighted translation instances. In *MT Summit.*, 2011.

[Prettenhofer and Stein, 2010] Peter Prettenhofer and Benno Stein. Cross-language text classification using structural correspondence learning. In *Proc. ACL*, 2010.

[Sennrich, 2012] Rico Sennrich. Mixture-modeling with unsupervised clusters for domain adaptation in statistical machine translation. In *Proc. EAMT*, 2012.

[Su *et al.*, 2012] Jinsong Su, Hua Wu, Haifeng Wang, Yidong Chen, Xiaodong Shi, Huailin Dong, and Qun Liu. Translation model adaptation for statistical machine translation with monolingual topic information. In *Proc. ACL*, 2012.

[Ueffing *et al.*, 2007] Nicola Ueffing, Gholamreza Haffari, and Anoop Sarkar. Transductive learning for statistical machine translation. In *Proc. ACL*, 2007.

[Wu, 1997] Dekai Wu. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Comput. Linguist.*, 1997.

[Xiong *et al.*, 2006] Deyi Xiong, Qun Liu, and Shouxun Lin. Maximum entropy based phrase reordering model for statistical machine translation. In *Proc. ACL*, 2006.

[Zens and Ney, 2006] Richard Zens and Hermann Ney. Discriminative reordering models for statistical machine translation. In *Proc. StatMT*, 2006.

[Zhang and Vogel, 2004] Ying Zhang and Stephan Vogel. Measuring confidence intervals for the machine translation evaluation metrics. In *Proc. TMI*, 2004.