Handling phrase reorderings for machine translation

Yizhao Ni, Craig J. Saunders[*], Sandor Szedmak and Mahesan Niranjan
ISIS Group
School of Electronics and Computer Science
University of Southampton
Southampton, SO17 1BJ
United Kingdom
yn05r@ecs.soton.ac.uk, craig.saunders@xrce.xerox.com,
{ss03v,mn}@ecs.soton.ac.uk

Abstract

We propose a distance phrase reordering model (DPR) for statistical machine translation (SMT), where the aim is to capture phrase reorderings using a structure learning framework. On both the reordering classification and a Chinese-to-English translation task, we show improved performance over a baseline SMT system.

1 Introduction

Word or phrase reordering is a common problem in bilingual translations arising from different grammatical structures. For example, in Chinese the expression of the date follows “Year/Month/Date”, while when translated into English, “Month/Date/Year” is often the correct grammar. In general, the fluency of machine translations can be greatly improved by obtaining the correct word order in the target language.

As the reordering problem is computationally expensive, a word distance-based reordering model is commonly used among SMT decoders (Koehn, 2004), in which the costs of phrase movements are linearly proportional to the reordering distance. Although this model is simple and efficient, the content independence makes it difficult to capture many distant phrase reordering caused by the grammar. To tackle the problem, (Koehn et al., 2005) developed a *lexicalized reordering model* that attempted to learn the phrase reordering based on content. The model learns the local orientation (e.g. “monotone” order or “switching” order) probabilities for each bilingual phrase pair using Maximum Likelihood Estimation (MLE). These orientation probabilities are then integrated into an SMT decoder to help finding a Viterbi–best local orientation sequence. Improvements by this model have been reported in (Koehn et al., 2005). However, the amount of the training data for each bilingual phrase is so small that the model usually suffers from the data sparseness problem. Adopting the idea of predicting the orientation, (Zens and Ney, 2006) started exploiting the context and grammar which may relate to phrase reorderings. In general, a *Maximum Entropy* (ME) framework is utilized and the feature parameters are tuned by a discriminative model. However, the training times for ME models are usually relatively high, especially when the output classes (i.e. phrase reordering orientations) increase.

Alternative to the ME framework, we propose using a classification scheme here for phrase reorderings and employs a structure learning framework. Our results confirm that this distance phrase reordering model (DPR) can lead to improved performance with a reasonable time efficiency.

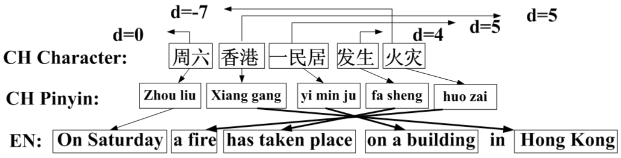

Figure 1: The phrase reordering distance d.

2 Distance phrase reordering (DPR)

We adopt a discriminative model to capture the frequent distant reordering which we call *distance phrase reordering*. An ideal model would consider every position as a class and predict the position of the next phrase, although in practice we must consider a limited set of classes (denoted as Ω). Using the reordering distance d (see Figure 1) as defined by (Koehn et al., 2005), we extend the two class model in (Xiong et al., 2006) to multiple classes (e.g. three–class setup $\Omega = \{d < 0, d = 0, d > 0\}$; or five–class setup $\Omega = \{d \leq -5, -5 < d < 0, d = 0, 0 < d < 5, d \geq 5\}$). Note that the more

[*] the author’s new address: Xerox Research Centre Europe 6, Chemin de Maupertuis, 38240 Meylan France.

Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pages 241–244,
Suntec, Singapore, 4 August 2009. ©2009 ACL and AFNLP

classes it has, the closer it is to the ideal model, but the smaller amount of training samples it would receive for each class.

2.1 Reordering Probability model and training algorithm

Given a (source, target) phrase pair $(\bar{f}_j, \bar{e}_i)$ with $\bar{f}_j = [f_{j_l}, \ldots, f_{j_r}]$ and $\bar{e}_i = [e_{i_l}, \ldots, e_{i_r}]$, the distance phrase reordering probability has the form

$$p(o|\bar{f}_j, \bar{e}_i) := \frac{h\big(\mathbf{w}_o^T \phi(\bar{f}_j, \bar{e}_i)\big)}{\sum_{o' \in \Omega} h\big(\mathbf{w}_{o'}^T \phi(\bar{f}_j, \bar{e}_i)\big)} \qquad (1)$$

where $\mathbf{w}_o = [w_{o,0}, \ldots, w_{o,dim(\phi)}]^T$ is the weight vector measuring features' contribution to an orientation $o \in \Omega$, ϕ is the feature vector and h is a pre-defined monotonic function. As the reordering orientations tend to be interdependent, learning $\{\mathbf{w}_o\}_{o \in \Omega}$ is more than a multi–class classification problem. Take the five–class setup for example, if an example in class $d \leq -5$ is classified in class $-5 < d < 5$, intuitively the loss should be smaller than when it is classified in class $d > 5$. The output (orientation) domain has an inherent structure and the model should respect it. Hence, we utilize the structure learning framework proposed in (Taskar et al., 2003) which is equivalent to minimising the sum of the classification errors

$$\min_{\mathbf{w}} \frac{1}{N} \sum_{n=1}^{N} \rho(o, \bar{f}_j^n, \bar{e}_i^n, \mathbf{w}) + \frac{\lambda}{2} \|\mathbf{w}\|^2 \qquad (2)$$

where $\lambda \geq 0$ is a regularisation parameter,

$$\begin{aligned} \rho(o, \bar{f}_j, \bar{e}_i, \mathbf{w}) = \quad & \max\{0, \max_{o' \neq o}[\triangle(o, o') + \\ & \mathbf{w}_{o'}^T \phi(\bar{f}_j, \bar{e}_i)] - \mathbf{w}_o^T \phi(\bar{f}_j, \bar{e}_i)\} \end{aligned}$$

is a structured margin loss function with

$$\triangle(o, o') = \begin{cases} 0 & \text{if } o = o' \\ 0.5 & \text{if } o \text{ and } o' \text{ are close in } \Omega \\ 1 & \text{else} \end{cases}$$

measuring the distance between pseudo orientation o' and the true one o. Theoretically, this loss requires that orientation o' which are "far away" from the true one o must be classified with a large margin while nearby candidates are allowed to be classified with a smaller margin. At training time, we used a perceptron–based structure learning (PSL) algorithm to learn $\{\mathbf{w}_o\}_{o \in \Omega}$ which is shown in Table 1.

```
Input: The samples $\{o, \phi(\bar{f}_j, \bar{e}_i)\}_{n=1}^{N}$, step size $\eta$
Initialization: $k = 0$; $\mathbf{w}_{o,k} = \mathbf{0} \quad \forall o \in \Omega$;
Repeat
  for $n = 1, 2, \ldots, N$ do
    for $o' \neq o$ get
      $V = \max_{o'} \{\triangle(o, o') + \mathbf{w}_{o',k}^T \phi(\bar{f}_j, \bar{e}_i)\}$
      $o^* = \arg\max_{o'} \{\triangle(o, o') + \mathbf{w}_{o',k}^T \phi(\bar{f}_j, \bar{e}_i)\}$
    if $w_{o,k}^T \phi(\bar{f}_j, \bar{e}_i) < V$ then
      $\mathbf{w}_{o,k+1} = \mathbf{w}_{o,k} + \eta \phi(\bar{f}_j, \bar{e}_i)$
      $\mathbf{w}_{o^*,k+1} = \mathbf{w}_{o^*,k} - \eta \phi(\bar{f}_j, \bar{e}_i)$
  $k = k + 1$
until converge
Output: $\mathbf{w}_{o,k+1} \quad \forall o \in \Omega$
```

Table 1: Perceptron-based structure learning.

2.1.1 Feature Extraction and Application

Following (Zens and Ney, 2006), we consider different kinds of information extracted from the phrase environment (see Table 2), where given a sequence s (e.g. $s = [f_{j_l - z}, \ldots, f_{j_l}]$), the features selected are $\phi_u(s_p^{|u|}) = \delta(s_p^{|u|}, u)$, with the indicator function $\delta(\cdot, \cdot)$, $p = \{j_l - z, \ldots, j_r + z\}$ and string $s_p^{|u|} = [f_p, \ldots, f_{p+|u|}]$. Hence, the phrase features are distinguished by both the content u and its start position p. For example, the left side context features for phrase pair (xiang gang, Hong Kong) in Figure 1 are $\{\delta(s_0^1, \text{"zhou"}), \delta(s_1^1, \text{"liu"}), \delta(s_0^2, \text{"zhou liu"})\}$. As required by the algorithm, we then *normalise* the feature vector $\bar{\phi}_t = \frac{\phi_t}{\|\phi\|}$.

To train the DPR model, the training samples $\{(\bar{f}_j^n, \bar{e}_i^n)\}_{n=1}^N$ are extracted following the phrase pair extraction procedure in (Koehn et al., 2005) and form the sample pool, where the instances having the same source phrase $\bar{f}_j$ are considered to be from the same cluster. A sub-DPR model is then trained for each cluster using the PSL algorithm. During the decoding, the DPR model finds the corresponding sub-DPR model for a source phrase $\bar{f}_j$ and generates the reordering probability for each orientation class using equation (1).

3 Experiments

Experiments used the Hong Kong Laws corpus[1] (Chinese-to-English), where sentences of lengths between 1 and 100 words were extracted and the ratio of source/target lengths was no more than $2:1$. The training and test sizes are $50,290$ and $1,000$ respectively.

[1] This bilingual Chinese-English corpus consists of mainly legal and documentary texts from Hong Kong. The corpus is aligned at the sentence level which are collected and revised manually by the author. The full corpus will be released soon.

	Features for source phrase $\bar{f}_j$	Features for target phrase $\bar{e}_i$
Context	Source word n–grams within a window (length z) around the phrase edge $[j_l]$ and $[j_r]$	Target word n–grams of the phrase $[e_{i_l}, \dots, e_{i_r}]$
Syntactic	Source word class tag n-grams within a window (length z) around the phrase edge $[j_l]$ and $[j_r]$	Target word class tag n-grams of the phrase $[e_{i_l}, \dots, e_{i_r}]$

Table 2: The environment for the feature extraction. The word class tags are provided by MOSES.

3.1 Classification Experiments

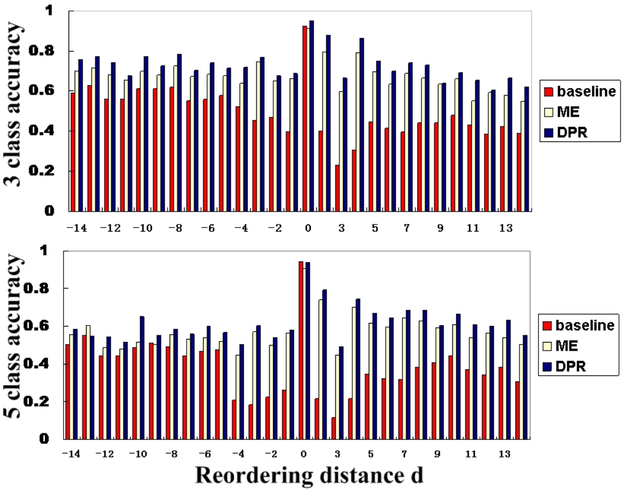

Figure 2: Classification results with respect to d.

We used GIZA++ to produce alignments, enabling us to compare using a DPR model against a baseline lexicalized reordering model (Koehn et al., 2005) that uses MLE orientation prediction and a discriminative model (Zens and Ney, 2006) that utilizes an ME framework. Two orientation classification tasks are carried out: one with three–class setup and one with five–class setup. We discarded points that had long distance reordering ($|d| > 15$) to avoid some alignment errors cause by GIZA++ (representing less than 5% of the data). This resulted in data sizes shown in Table 3. The classification performance is measured by an overall precision across all classes and the class-specific F1 measures and the experiments are are repeated three times to asses variance.

Table 4 depicts the classification results obtained, where we observed consistent improvements for the DPR model over the baseline and the ME models. When the number of classes (orientations) increases, the average relative improvements of DPR for the switching classes (i.e. $d \neq 0$) increase from 41.6% to 83.2% over the baseline and from 7.8% to 14.2% over the ME model, which implies a potential benefit of structure learning. Figure 2 further demonstrate the average accuracy for each reordering distance d. It shows that even for long distance reordering, the DPR model still performs well, while the MLE baseline usually performs badly (more than half examples are classified incorrectly). With so many classification errors, the effect of this baseline in an SMT system is in doubt, even with a powerful language model. At training time, training a DPR model is much faster than training an ME model (both algorithms are coded in Python), especially when the number of classes increase. This is because the generative iterative scaling algorithm of an ME model requires going through all examples twice at each round: one is for updating the conditional distributions $p(o|\bar{f}_j, \bar{e}_i)$ and the other is for updating $\{\mathbf{w}_o\}_{o\in\Omega}$. Alternatively, the PSL algorithm only goes through all examples once at each round, making it faster and more applicable for larger data sets.

3.2 Translation experiments

We now test the effect of the DPR model in an MT system, using MOSES (Koehn et al., 2005) as a baseline system. To keep the comparison fair, our MT system just replaces MOSES's reordering models with DPR while sharing all other models (i.e. phrase translation probability model, 4-gram language model (A. Stolcke, 2002) and beam search decoder). As in classification experiments the three-class setup shows better results in switching classes, we use this setup in DPR. In detail, all consistent phrases are extracted from the training sentence pairs and form the sample pool. The three-class DPR model is then trained by the PSL algorithm and the function $h(z) = \exp(z)$ is applied to equation (1) to transform the prediction scores. Contrasting the direct use of the reordering probabilities used in (Zens and Ney, 2006), we utilize the probabilities to adjust the word distance–based reordering cost, where the reordering cost of a sentence is computed as $P_o(\mathbf{f}, \mathbf{e}) =$

Settings	three–class setup			five–class setup				
Classes	$d < 0$	$d = 0$	$d > 0$	$d \leq -5$	$-5 < d < 0$	$d = 0$	$0 < d < 5$	$d \geq 5$
Train	$181,583$	$755,854$	$181,279$	$82,677$	$98,907$	$755,854$	$64,881$	$116,398$
Test	$5,025$	$21,106$	$5,075$	$2,239$	$2,786$	$21,120$	$1,447$	$3,629$

Table 3: Data statistics for the classification experiments.

System	three–class setup task				
	Precision	$d < 0$	$d = 0$	$d > 0$	Training time (hours)
Lexicalized	77.1 ± 0.1	55.7 ± 0.1	86.5 ± 0.1	49.2 ± 0.3	1.0
ME	83.7 ± 0.3	67.9 ± 0.3	90.8 ± 0.3	69.2 ± 0.1	58.6
DPR	$\mathbf{86.7} \pm 0.1$	$\mathbf{73.3} \pm 0.1$	$\mathbf{92.5} \pm 0.2$	$\mathbf{74.6} \pm 0.5$	27.0

System	five–class setup task						
	Precision	$d \leq -5$	$-5 < d < 0$	$d = 0$	$0 < d < 5$	$d \geq 5$	Training Time (hours)
Lexicalized	74.3 ± 0.1	44.9 ± 0.2	32.0 ± 1.5	86.4 ± 0.1	29.2 ± 1.7	46.2 ± 0.8	1.3
ME	80.0 ± 0.2	52.1 ± 0.1	54.7 ± 0.7	90.4 ± 0.2	63.9 ± 0.1	61.8 ± 0.1	83.6
DPR	$\mathbf{84.6 \pm 0.1}$	$\mathbf{60.0} \pm 0.7$	$\mathbf{61.4} \pm 0.1$	$\mathbf{92.6} \pm 0.2$	$\mathbf{75.4} \pm 0.6$	$\mathbf{68.8} \pm 0.5$	29.2

Table 4: Overall precision and class-specific F1 scores [%] using different number of orientation classes. Bold numbers refer to the best results.

$\exp\{-\sum_{m} \frac{d_m}{\beta p(o|\bar{f}_{jm},\bar{e}_{im})}\}$ with tuning parameter β. This distance–sensitive expression is able to fill the deficiency of the three–class setup of DPR and is verified to produce better results. For parameter tuning, minimum-error-rating training (F. J. Och, 2003) is used in both systems. Note that there are 7 parameters needed tuning in MOSES's reordering models, while only 1 requires tuning in DPR. The translation performance is evaluated by four MT measurements used in (Koehn et al., 2005).

Table 5 shows the translation results, where we observe consistent improvements on most evaluations. Indeed both systems produced similar word accuracy, but our MT system does better in phrase reordering and produces more fluent translations.

4 Conclusions and Future work

We have proposed a distance phrase reordering model using a structure learning framework. The classification tasks have shown that DPR is better in capturing the phrase reorderings over the lexicalized reordering model and the ME model. Moreover, compared with ME DPR is much faster and more applicable to larger data sets. Translation experiments carried out on the Chinese-to-English task show that DPR gives more fluent translation results, which verifies its effectiveness. For future work, we aim at improving the prediction accuracy for the five-class setup using a richer feature set before applying it to an MT system, as DPR can be more powerful if it is able to provide more precise phrase position for the decoder. We will also apply DPR on a larger data set to test its performance as well as its time efficiency.

Tasks	Measure	MOSES	DPR
CH–EN	BLEU [%]	44.7 ± 1.2	$\mathbf{47.1} \pm 1.3$
	word accuracy	$\mathbf{76.5} \pm 0.6$	76.1 ± 1.5
	NIST	8.82 ± 0.11	$\mathbf{9.04} \pm 0.26$
	METEOR [%]	66.1 ± 0.8	$\mathbf{66.4} \pm 1.1$

Table 5: Four evaluations for the MT experiments. Bold numbers refer to the best results.

References

P. Koehn. 2004. Pharaoh: a beam search decoder for phrase–based statistical machine translation models. In *Proc. of AMTA 2004*, Washington DC, October.

P. Koehn, A. Axelrod, A. B. Mayne, C. Callison–Burch, M. Osborne and D. Talbot. 2005. Edinburgh system description for the 2005 IWSLT speech translation evaluation. In *Proc. of IWSLT*, Pittsburgh, PA.

F. J. Och. 2003. SRILM - An Extensible Language Modeling Toolkit. In *Proc. Intl. Conf. Spoken Language Processing*, Colorado, September.

A. Stolcke. 2002. Minimum error rate training in statistical machine translation. In *Proc. ACL*, Japan.

B. Taskar, C. Guestrin, and D.Koller. 2003. Max–margin Markov networks. In *Proc. NIPS*, Vancouver, Canada, December.

D. Xiong, Q. Liu and S. Lin. 2006. Maximum Entropy Based Phrase Reordering Model for Statistical Machine Translation. In *Proc. of ACL*, Sydney, July.

R. Zens and H. Ney. 2006. Discriminative Reordering Models for Statistical Machine Translation. In *Proc. of ACL*, pages 55–63, New York City, June.