Improving Reordering for Statistical Machine Translation with Smoothed Priors and Syntactic Features

Bing Xiang, Niyu Ge, and Abraham Ittycheriah
IBM T. J. Watson Research Center
Yorktown Heights, NY 10598
{bxiang,niyuge,abei}@us.ibm.com

Abstract

In this paper we propose several novel approaches to improve phrase reordering for statistical machine translation in the framework of maximum-entropy-based modeling. A smoothed prior probability is introduced to take into account the distortion effect in the priors. In addition to that we propose multiple novel distortion features based on syntactic parsing. A new metric is also introduced to measure the effect of distortion in the translation hypotheses. We show that both smoothed priors and syntax-based features help to significantly improve the reordering and hence the translation performance on a large-scale Chinese-to-English machine translation task.

1 Introduction

Over the past decade, statistical machine translation (SMT) has evolved into an attractive area in natural language processing. SMT takes a source sequence, $S = [s_1\ s_2\ \ldots\ s_K]$ from the source language, and generates a target sequence, $T^* = [t_1\ t_2\ \ldots\ t_L]$, by finding the most likely translation given by:

$$T^* = \arg\max_T p(T|S) \quad (1)$$

In most of the existing approaches, following (Brown et al., 1993), Eq. (1) is factored using the source-channel model into

$$T^* = \arg\max_T p(S|T)p^\lambda(T), \quad (2)$$

where the two models, the translation model, $p(S|T)$, and the language model (LM), $p(T)$, are estimated separately: the former using a parallel corpus and a hidden alignment model and the latter using a typically much larger monolingual corpus. The weighting factor λ is typically tuned on a development test set by optimizing a translation accuracy criterion such as BLEU (Papineni et al., 2002).

In recent years, among all the proposed approaches, the phrase-based method has become the widely adopted one in SMT due to its capability of capturing local context information from adjacent words. Word order in the translation output relies on how the phrases are reordered based on both language model scores and distortion cost/penalty (Koehn et al., 2003), among all the features utilized in a maximum-entropy (log-linear) model (Och and Ney, 2002). The distortion cost utilized during the decoding is usually a penalty linearly proportional to the number of words in the source sentence that are skipped in a translation path.

In this paper, we propose several novel approaches to improve reordering in the phrase-based translation with a maximum-entropy model. In Section 2, we review the previous work that focused on the distortion and phrase reordering in SMT. In Section 3, we briefly review the baseline of this work. In Section 4, we introduce a smoothed prior probability by taking into account the distortions in the priors. In Section 5, we present multiple novel distortion features based on syntactic parsing. A new distortion evaluation metric is proposed in Section 6 and experimental results on a large-scale Chinese-English machine translation task are reported in Section 7. Section 8 concludes the paper.

Proceedings of SSST-5, Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation, pages 61–69,
ACL HLT 2011, Portland, Oregon, USA, June 2011.

2 Previous Work

Significant amount of research has been conducted in the past on the word reordering problem in SMT. In (Brown et al., 1993) IBM Models 3 through 5 model reordering based on the surface word information. For example, Model 4 attempts to assign target-language positions to source-language words by modeling $d(j|i, K, L)$ where j is the target-language position, i is the source-language position, K and L are respectively source and target sentence lengths. These models are not effective in modeling reordering because they do not have enough context and lack structural information.

Phrase-based SMT systems such as (Koehn et al., 2003) move from using words as translation units to using phrases. One of the advantages of phrase-based SMT systems is that the local reordering is inherent in the phrase translations. However, phrase-based SMT systems capture reordering instances and not reordering phenomena. It has trouble to produce the right translation order if the training data does not contain the specific phrase pairs. For example, phrases do not capture the phenomenon that Arabic adjectives and nouns need to be reordered.

Instead of directly modeling the distance of word movement, some phrase-level reordering models indicate how to move phrases, also called orientations. Orientations typically apply to the adjacent phrases. Two adjacent phrases can be either placed monotonically (sometimes called straight) or swapped (non-monotonically or inverted). In (Och and Ney, 2004; Tillmann, 2004; Kumar and Byrne, 2005; Al-Onaizan and Papineni, 2006; Xiong et al., 2006; Zens and Ney, 2006; Ni et al., 2009), people presented models that use lexical features from the phrases to predict their orientations. These models are very powerful in predicting local phrase placements. In (Galley and Manning, 2008) a hierarchical orientation model is introduced that captures some non-local phrase reordering by a shift reduce algorithm. Because of the heavy use of lexical features, these models tend to suffer from data sparseness problems.

Syntax information has been used for reordering, such as in (Xia and McCord, 2004; Collins et al., 2005; Wang et al., 2007; Li et al., 2007; Chang et al., 2009). More recently, in (Ge, 2010) a probabilistic reordering model is presented to model directly the source translation sequence and explicitly assign probabilities to the reordering of the source input with no restrictions on gap, length or adjacency. The reordering model is used to generate a reordering lattice which encodes many reordering and their costs (negative log probability). Another recent work is (Green et al., 2010), which estimates future linear distortion cost and presents a discriminative distortion model that predicts word movement during translation based on multiple features.

This work differentiates itself from all the previous work on the phrase reordering as the following. Firstly, we propose a smoothed distortion prior probability in the maximum-entropy-based MT framework. It not only takes into account the distortion in the prior, but also alleviates the data sparseness problem. Secondly, we propose multiple syntactic features based on the source-side parse tree to capture the reordering phenomena between two different languages. The correct reordering patterns will be automatically favored during the decoding, due to the higher weights obtained through the maximum entropy training on the parallel data. Finally, we also introduce a new metric to quantify the effect on the distortions in different systems. The experiments on a Chinese-English MT task show that these proposed approaches additively improve both the distortion and translation performance significantly.

3 Maximum-Entropy Model for MT

In this section we give a brief review of a special maximum-entropy (ME) model as introduced in (Ittycheriah and Roukos, 2007). The model has the following form,

$$p(\mathbf{t}, j|\mathbf{s}) = \frac{p_0(\mathbf{t}, j|\mathbf{s})}{Z} exp \sum_i \lambda_i \phi_i(\mathbf{t}, j, \mathbf{s}), \quad (3)$$

where $\mathbf{s}$ is a source phrase, and $\mathbf{t}$ is a target phrase. j is the jump distance from the previously translated source word to the current source word. During training j can vary widely due to automatic word alignment in the parallel corpus. To limit the sparseness created by long jumps, j is capped to a window of source words (-5 to 5 words) around the last translated source word. Jumps outside the window are treated as being to the edge of the window. In

Eq. (3), p_0 is a prior distribution, Z is a normalizing term, and $\phi_i(\mathbf{t}, j, \mathbf{s})$ are the features of the model, each being a binary question asked about the source and target streams. The feature weights λ_i can be estimated with the Improved Iterative Scaling (IIS) algorithm.

Several categories of features have been proposed:

- Lexical features that examine source word, target word and jump;
- Lexical context features that examine the previous and next source words, and also the previous two target words;
- Segmentation features based on morphological analysis;
- Part-of-speech (POS) features that collect the syntactic information from the source and target words;
- Coverage features that examine the coverage status of the source words to the left and to the right. They fire only if the left source is open (untranslated) or the right source is closed.

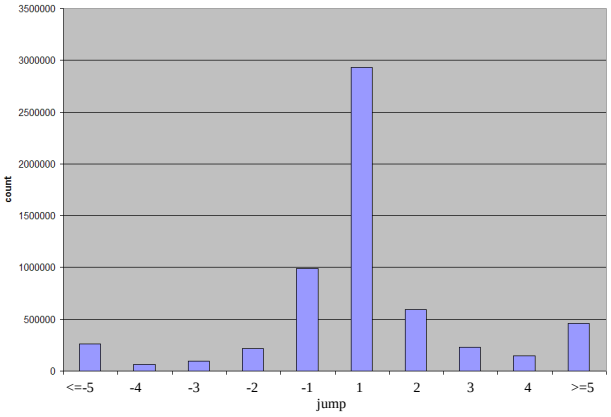

Figure 1: Counts of jumps for words with POS *NN*.

4 Distortion Priors

Generally the prior distribution in Eq. (3) can contain any information we know about the future.

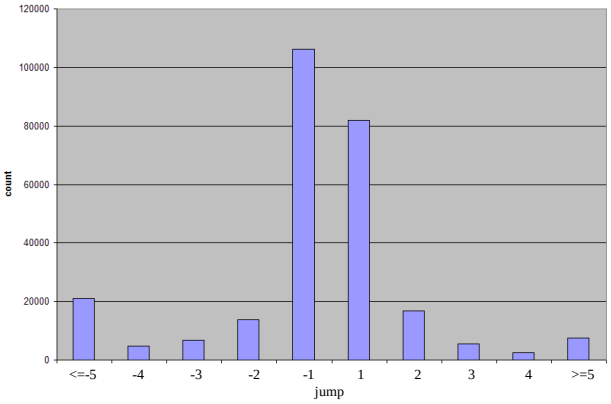

Figure 2: Counts of jumps for words with POS *NT*.

In (Ittycheriah and Roukos, 2007), the normalized phrase count is utilized as the prior, i.e.

$$p_0(\mathbf{t}, j|\mathbf{s}) \approx \frac{1}{l} p_0(\mathbf{t}|\mathbf{s}) = \frac{C(\mathbf{s}, \mathbf{t})}{l * C(\mathbf{s})} \quad (4)$$

where l is the jump window size (a constant), $C(\mathbf{s}, \mathbf{t})$ is the co-ocurrence count of phrase pair ($\mathbf{s}$, $\mathbf{t}$), and $C(\mathbf{s})$ is the source phrase count of $\mathbf{s}$. It can be seen that distortion j is not taken into account in Eq. (4). The contribution of distortion solely comes from the features. In this work, we estimate the prior probability with distortion included,

$$p_0(\mathbf{t}, j|\mathbf{s}) = p_0(\mathbf{t}|\mathbf{s}) p(j|\mathbf{s}, \mathbf{t}) \quad (5)$$

where $p(j|\mathbf{s}, \mathbf{t})$ is the distortion probability for a given phrase pair ($\mathbf{s}$, $\mathbf{t}$).

Due to the sparseness issue in the estimation of $p(j|\mathbf{s}, \mathbf{t})$, we choose to smooth it with the global distortion probability through

$$p(j|\mathbf{s}, \mathbf{t}) = \alpha p_l(j|\mathbf{s}, \mathbf{t}) + (1 - \alpha) p_g(j), \quad (6)$$

where p_l is the local distortion probability estimated based on the counts of jumps for each phrase pair in the training, p_g is the global distortion probability estimated on all the training data, and α is the interpolation weight. In this work, p_g is estimated based on either source POS (if it's a single-word source phrase) or source phrase size (if it's more than one word long), as shown below.

$$p_g(j) = \begin{cases} P_g(j|POS), & \text{if } |\mathbf{s}| = 1 \\ P_g(j||\mathbf{s}|), & \text{if } |\mathbf{s}| > 1 \end{cases} \quad (7)$$

In this way, the system can differentiate the distortion distributions for single source words with different POS tags, such as adjectives versus nouns. And in the meantime, we also differentiate the distortion distribution with different source phrase lengths. We show several examples of the jump distributions in Fig. 1 and 2 collected from 1M sentence pairs in a Chinese-to-English parallel corpus with automatic parsing and word alignment. Fig. 1 shows the count histogram for single-word phrases with POS tag as *NN*. The distortion with $j = 1$, i.e. monotone, dominates the distribution with the highest count. The reordering with $j = -1$ has the second highest count. Such pattern is shared by most of the other POS tags. However, Fig. 2 shows that the distribution of jumps for *NT* is quite different from *NN*. The jump with $j = -1$ is actually the most dominant, with higher counts than monotone translation. This is due to the different order in English when translating Chinese temporal nouns.

5 Distortion Features

Although the maximum entropy translation model has an explicit indicator of distortion, j, built into the features, we discuss in this section some novel features that try to capture the distortion phenomena of translation. These features are questions about the parse tree of the source language and in particular about the local parse node neighborhood of the current source word being translated. Figure 3 shows an example sentence from the Chinese-English Parallel Treebank (LDC2009E83) and the source language parse is displayed on the left. The features below can be viewed as either being within a parse node or asking about the coverage status of neighborhood nodes.

Since these features are asking about the current coverage, they are specific to a path in the search lattice during the decoding phase of translation. Training these features is done by evaluating on the path defined by the automatic word alignment of the parallel corpus sentence.

5.1 Parse Tree Modifications

The 'de' construction in Chinese is by now famous. In order to ask more coherent questions about the parse neighborhood, we modify the parse structures to "raise" the 'de' structure. The parse trees annotated by the LDC have a structure as shown in Fig. 4. After raising the 'de' structure we obtain the tree in Fig. 5.

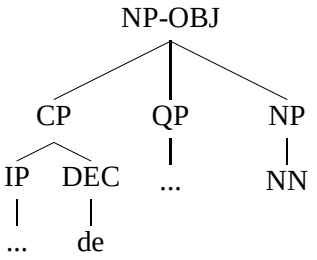

Figure 4: Original parse tree from LDC.

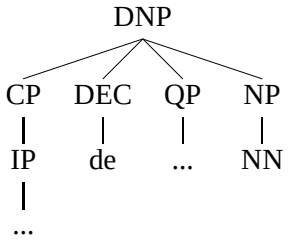

Figure 5: The parse tree after transformation.

The transformation has been applied to the example shown in Figure 3. The resulting flat structure facilitates the parse sibling feature discussed below.

5.2 Parse Coverage Feature

The first set of new features we will introduce is the source parse coverage feature. This feature is interior to a source parse node and asks if the leaves under this parse node are covered (translated) or not so far. The feature has the following components:

ϕ_i(*SourceWord, TargetWord, SourceParseParent, jump, Coverage*).

Unary parents in the source parse tree are excluded since the feature has no ambiguity in coverage. In Figure 3, the 'PP' node above position 5 has two children, P, NP. When translating source position 6, this feature indicates that the PP node has a leaf that is already covered.

5.3 Parse Sibling Feature

The second set of new features is the source parse sibling feature. This feature asks whether the neigh-

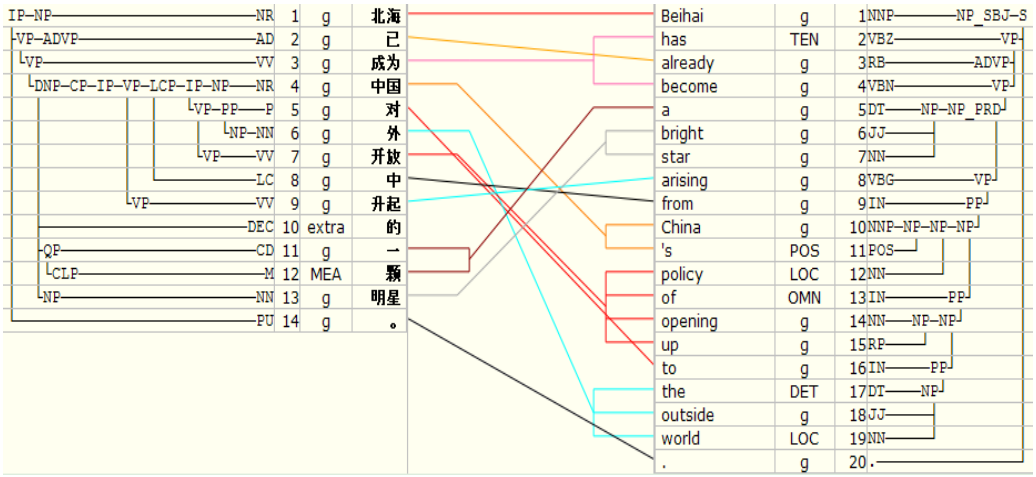

Figure 3: Chinese-English example.

boring parse node has been covered or not. The feature includes two types:

ϕ_i(*SourceWord, TargetWord, SourceParseSibling, jump, SiblingCoverage, SiblingOrientation*)

and

ϕ_i(*SourcePOS, TargetPOS, SourceParseSibling, jump, SiblingCoverage, SiblingOrientation*).

Some example features for the first type are shown in Table 1, where $\alpha_i = e^{\lambda_i}$. The coverage status (Cov) of the parse sibling node indicates if the node is covered completely (1), partially (2) or not covered (0). In order to capture the relationship of the neighborhood node, we indicate the orientation which can be either of {left (-1), right (1)}. Given the example shown in Figure 3, at source position 10, the system can now ask about the 'CP' structure to the left and the 'QP' and 'NP' structures to the right. An α_i of greater than 1.0 (meaning $\lambda_i > 0$) indicates that the feature increases the probability of the related target block. From these examples, it's clear that the system prefers to produce an empty translation for the Chinese word "de" when the 'QP' and 'NP' nodes to the right of it are already covered (the first two features in Table 1) and when the 'CP' node to left is still uncovered (the third feature). The last feature in the table shows α_i for the case when 'CP' has already been covered.

These features are able to capture neighborhoods that are much larger than the original baseline model which only asked questions about the immediate lexical neighborhood of the current source word.

Cnt	α_i	Tgt	Src	Parse Node	Cov	Orientation
18065	2.06	e_0	de	QP	1	1
366153	1.99	e_0	de	NP	1	1
143433	3.41	e_0	de	CP	0	-1
99297	1.05	e_0	de	CP	1	-1

Table 1: Parse Sibling Word Features (e_0 represents empty target).

6 A New Distortion Evaluation Metric

MT performance is usually measured by such metric as BLEU which measures the MT output as a whole including word choice and reordering. It is useful to measure these components separately. Unigram BLEU (BLEUn1) measures the precision of word choice. We need a metric for measuring reordering accuracy. The naive way of counting accuracy at every source position does not account for the case of the phrasal movement. If a phrase is moved to the wrong place, every source word in the phrase would be penalized whereas a more reasonable metric would penalize the phrase movement only once if the phrase boundary is correct.

We propose the following pair-wise distortion metric. From an MT output, we first extract the source visit sequence:

Hyp:$\{h_1, h_2, \ldots h_n\}$

where h_i are the visit order of the source sentence. From the reference, we extract the true visit sequence:

Ref:$\{r_{1,}r_{2,}\ldots r_n\}$

The Pair-wise Distortion metric PDscore can be computed as follows:

$$PDscore(\overrightarrow{H}) = \sum_{i=1}^{n} \frac{I(h_i = r_j \wedge h_{i-1} = r_{j-1})}{n} \quad (8)$$

It measures how often the translation output gets the pair-wise source visit order correct. We notice that an MT metric named LRscore was proposed in (Birch and Osborne, 2010). It computes the distance between two word order sequences, which is different from the metric we proposed here.

7 Experiments

7.1 Data and Baseline

We conduct a set of experiments on a Chinese-to-English MT task. The training data includes the UN parallel corpus and LDC-released parallel corpora, with about 11M sentence pairs, 320M words in total (counted at the English side). To evaluate the smoothed distortion priors and different features, we use an internal data set as the development set and the NIST MT08 evaluation set as the test set, which includes 76 documents (691 sentences) in newswire and 33 documents (666 sentences) in weblog, both with 4 sets of references for each sentence. Instead of using all the training data, we sample the training corpus based on the dev/test set to train the system more efficiently. The most recent and good-quality corpora are sampled first. For the given test set, we obtain the first 20 instances of n-grams (length from 1 to 15) from the test that occur in the training universe and the resulting sentences then form the training sample. In the end, 1M sentence pairs are selected for the sampled training for each genre of the MT08 test set.

A 5-gram language model is trained from the English Gigaword corpus and the English portion of the parallel corpus used in the translation model training. The Chinese parse trees are produced by a maximum entropy based parser (Ratnaparkhi, 1997). The baseline decoder is a phrase-based decoder that employs both normal phrases and also non-contiguous phrases. The value of maximum skip is set to 9 in all the experiments. The smoothing parameter α for distortion prior is set to 0.9 empirically based on the results on the development set.

7.2 Distortion Evaluation

We evaluate the MT distortion using the metric in Eq. (8) on two hand-aligned test sets. Test-278 includes 278 held-out sentences. Test-52 contains the first 52 sentences from the MT08 Newswire set, with the Chinese input sentences manually aligned to the first set of reference translations. From the hand alignment, we extract the true source visit sequence and this is the reference.

The evaluation results are in Table 2. It is shown that the smoothed distortion prior, parse coverage feature and parse sibling feature each provides improvement on the PDscore on Test-278 and Test-52. The final system scores are 2 to 3 points absolute higher than the baseline scores. The state visit sequence in the final system is closer to the true visit sequence than that of the baseline. This indicates the advantage of using both parse-based syntactic features and also the smoothed prior that takes into account of the distortion effect. We also provide an upper-bound in the last row by computing the PDscore between the first and second set of references for Test-52. The number shows the agreement between two human translators in terms of PDscore is around 71%.

System	Test-278	Test-52
ME Baseline	44.58	48.96
+Prior	45.12	49.22
+COV	45.00	49.03
+SIB	45.43	49.20
+COV+SIB	46.16	49.45
+Prior+COV+SIB	47.68	51.04
Ref1 vs. Ref2	-	70.99

Table 2: Distortion accuracy PDscore (Prior:smoothed distortion prior; COV:parse coverage feature; SIB:parse sibling feature).

7.3 Translation Results

Translation results on the MT08 Newswire set and MT08 Weblog set are listed in Table 3 and Table 4 respectively. The MT performance is measured with the widely adopted BLEU and TER (Snover et al., 2006) metrics. We also compare the results from different configurations with a normal phrase-based

System	Number of Features	BLEU	TER
PBT	n/a	29.71	59.40
ME	9,008,382	32.12	56.78
+Prior	9,008,382	32.46	56.41
+COV	9,202,431	32.48	56.50
+SIB	10,088,487	32.73	56.26
+COV+SIB	10,282,536	32.94	55.97
+Prior+COV+SIB	10,282,536	33.15	55.62

Table 3: MT results on MT08 Newswire set (PBT:normal phrase-based MT; ME:Maximum-entropy baseline; Prior:smoothed distortion prior; COV:parse coverage feature; SIB:parse sibling feature).

System	Number of Features	BLEU	TER
PBT	n/a	20.07	62.90
ME	9,192,617	22.42	60.36
+Prior	9,192,617	22.70	60.11
+COV	9,306,967	22.69	60.14
+SIB	9,847,445	22.91	59.92
+COV+SIB	9,961,795	23.04	59.78
+Prior+COV+SIB	9,961,795	23.25	59.56

Table 4: MT results on MT08 Weblog set (PBT:normal phrase-based MT; ME:Maximum-entropy baseline; Prior:smoothed distortion prior; COV:parse coverage feature; SIB:parse sibling feature).

SMT system (Koehn et al., 2003) that is trained on the same training data. The number of features used in the systems are listed in the tables.

We start from the maximum-entropy baseline, a system implemented similarly as in (Ittycheriah and Roukos, 2007). It utilizes multiple features as listed in Section 3, including lexical reordering features, and produces an already significantly better performance than the normal phrase-based MT system (PBT). It is around 2.5 points better in both BLEU and TER than the PBT baseline. By adding smoothed priors, parse coverage features or parse sibling features each separately, the MT performance is improved by 0.3 to 0.6. The parse sibling feature alone provides the largest individual contribution. When adding both types of new features, the improvement is around 0.6 to 0.8 on two genres. Finally, applying all three results in the best performance (the last row). On the Newswire set, the final system is more than 3 points better than the PBT baseline and 1 point better than the ME baseline. On the Weblog set, it is more than 3 points better than PBT and 0.8 better than the ME baseline. All the MT results above are statistically significant with p-value < 0.0001 by using the tool described in (Zhang and Vogel, 2004).

7.4 Analysis

To better understand the distortion and translation results, we take a closer look at the parse-based features. In Table 5, we list the most frequent parse sibling features that are related to the Chinese phrases with "PP VV" structures. It is known that in Chinese usually the preposition phrases ("PP") are written/spoken before the verbs ("VV"), with a different order from English. Table 5 shows how such reordering phenomenon is captured by the parse sibling features. Recall that when α_i is greater than 1, the system prefers the reordering with that feature fired. When α_i is smaller than 1, the system will penalize the corresponding translation order during the decoding search. When the coverage is equal to 1, it means "PP" has been translated before translating current "VV". As shown in the table, those features with coverage equal to 1 have α_i lower than 1, which will result in penalties on incorrect translation orders.

In Fig. 6, we show the comparison between the

Count	α_i	j	TgtPOS	SrcPOS	ParseSib Node	Cov	Orien-tation
3052	1.10	5	VBD	VV	PP	0	-1
2662	1.10	-1	VBD	VV	PP	0	-1
2134	1.25	4	VBD	VV	PP	0	-1
50	0.73	5	VBD	VV	PP	1	-1
39	0.84	-5	VBD	VV	PP	1	-1
18	0.95	-2	VBD	VV	PP	1	-1

Table 5: Parse Sibling Word Features related to Chinese "PP VV".

Src1	瑞士 科学院 冰川 专家 长期 跟踪 研究 发现 , 1850 年 至 2005 年间 , 瑞士 的 1800 余 条 冰川 **正(were) 以(at) 年均(annual) 3% 的 速度(rate) 缩减 (shrinking)**。
Ref	a long-term follow-up research by glacier experts at the swiss academy of sciences found that from 1850 to 2005 the 1,800 plus glaciers in switzerland **were shrinking at an annual rate of 3 %** .
Baseline	the swiss academy of sciences glacier experts long-term follow-up study found that from 2005 to 1850 , with **an average** of more than 1800 glaciers in switzerland is the **reduced rate of 3 %** .
New	the swiss academy of sciences glacier experts long-term follow-up study found that from 1850 to 2005 , more than 1800 of swiss glaciers **shrinking at an annual rate of 3 %**.
Src2	但 在 此 同时 塔利班 组织 说 , 另 一 名 **遭到(had been) 绑架(kidnapped) 的(who) 德国(german) 人质(hostage)** 身体 非常 虚弱 , 开始 陷入 昏迷 并 失去 意识 。
Ref	but at the same time the taliban said that another **german hostage who had been kidnapped** was in extremely poor health , and had started to become comatose and to lose consciousness .
Baseline	but at the same time , another one **was kidnapped** by the taliban of the **german hostage** body very weak , began to fall into a coma and lost consciousness .
New	but at the same time , the taliban said that the body of another **german hostage who was kidnapped** very weak , began to fall into a coma and lost consciousness .

Figure 6: Chinese-English MT examples(Baseline:Maximum-entropy baseline; New:System with smoothed priors and syntactic features).

ME baseline output and those from the improved system with the parse-based features and smoothed distortion priors. The differences are highlighted in bold for easy understanding. The first example shows that the new system fixes the order for "PP VV", while the second one shows the fix for the translation of "CP de NP". This is consistent with the features we showed in Table 1 and 5. The new features help to translate the Chinese text in the right order.

8 Conclusion

In this paper we have presented several novel approaches that improved phrase reordering in the framework of maximum entropy based translation. A smoothed prior probability was proposed to take into account the distortions in the priors. Several novel distortion features were presented based on the syntactic parsing. A new metric PDscore was also introduced to measure the effect of distortion in the translation hypotheses. We showed that both smoothed prior and syntax-based features additively improved the distortion and also the translation performance significantly on a large-scale Chinese-English machine translation task. How to further take advantage of the syntactic information to improve the reordering in SMT will continue to be an interesting topic in the future.

Acknowledgments

We would like to acknowledge the support of DARPA under Grant HR0011-08-C-0110 for fund-

ing part of this work. The views, opinions, and/or findings contained in this article/presentation are those of the author/presenter and should not be interpreted as representing the official views or policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the Department of Defense.

References

Yaser Al-Onaizan and Kishore Papineni. 2006. Distortion models for statistical machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, pages 529–536, Sydney, Australia.

Alexandra Birch and Miles Osborne. 2010. Lrscore for evaluating lexical and reordering quality in mt. In *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR*.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–311.

Pi-Chuan Chang, Huihsin Tseng, Dan Jurafsky, and Christopher D. Manning. 2009. Discriminative reordering with chinese grammatical relations features. In *Proceedings of the Third Workshop on Syntax and Structure in Statistical Translation*.

Michael Collins, Philipp Koehn, and Ivona Kučerová. 2005. Clause restructuring for statistical machine translation. In *Proceedings of ACL*, pages 531–540.

Michel Galley and Christoph D. Manning. 2008. A simple and effective hierarchical phrase reordering model. In *Proceedings of the EMNLP*.

Niyu Ge. 2010. A direct syntax-driven reordering model for phrase-based machine translation. In *Proceedings of HLT-NAACL*, pages 849–857.

Spence Green, Michel Galley, and Christopher D. Manning. 2010. Improved models of distortion cost for statistical machine translation. In *Proceedings of HLT-NAACL*.

Abraham Ittycheriah and Salim Roukos. 2007. Direct translation model 2. In *Proceedings HLT/NAACL*, pages 57–64, April.

Philipp Koehn, Franz Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of NAACL/HLT*.

Shankar Kumar and William Byrne. 2005. Local phrase reordering models for statistical machine translation. In *Proceedings of HLT/EMNLP*, pages 161–168.

Chi-Ho Li, Dongdong Zhang, Mu Li, Ming Zhou, Minghui Li, and Yi Guan. 2007. A probabilistic approach to syntax-based reordering for statistical machine translation. In *Proceedings of ACL*.

Yizhao Ni, Craig J.Saunders, Sandor Szedmak, and Mahesan Niranjan. 2009. Handling phrase reorderings for machine translation. In *Proceedings of ACL*.

Franz-Josef Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translations. In *40th Annual Meeting of the ACL*, pages 295–302, Philadelphia, PA, July.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Weijing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318.

Adwait Ratnaparkhi. 1997. A linear observed time statistical parser based on maximum entropy models. In *Proceedings of EMNLP*, pages 1–10.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*.

Christoph Tillmann. 2004. A unigram orientation model for statistical machine translation. In *Proceedings of HLT-NAACL*.

Chao Wang, Michael Collins, and Philipp Koehn. 2007. Chinese syntactic reordering for statistical machine translation. In *Proceedings of EMNLP*, pages 737–745.

Fei Xia and Michael McCord. 2004. Improving a statistical mt system with automatically learned rewrite patterns. In *Proceedings of COLING*.

Deyi Xiong, Qun Liu, and Shouxun Lin. 2006. Maximum entropy based phrase reordering model for statistical machine translation. In *Proceedings of ACL*.

Richard Zens and Hermann Ney. 2006. Discriminative reordering models for statistical machine translation. In *Proceedings of the Workshop on Statistical Machine Translation*.

Ying Zhang and Stephan Vogel. 2004. Measuring confidence intervals for the machine translation evaluation metrics. In *Proceedings of The 10th International Conference on Theoretical and Methodological Issues in Machine Translation*.