Maximum Entropy Based Phrase Reordering for Hierarchical Phrase-based Translation

Zhongjun He Yao Meng Hao Yu
Fujitsu R&D Center CO., LTD.
15/F, Tower A, Ocean International Center, 56 Dongsihuan Zhong Rd.
Chaoyang District, Beijing, 100025, China
{hezhongjun, mengyao, yu}@cn.fujitsu.com

Abstract

Hierarchical phrase-based (HPB) translation provides a powerful mechanism to capture both short and long distance phrase reorderings. However, the phrase reorderings lack of contextual information in conventional HPB systems. This paper proposes a context-dependent phrase reordering approach that uses the maximum entropy (MaxEnt) model to help the HPB decoder select appropriate reordering patterns. We classify translation rules into several reordering patterns, and build a MaxEnt model for each pattern based on various contextual features. We integrate the MaxEnt models into the HPB model. Experimental results show that our approach achieves significant improvements over a standard HPB system on large-scale translation tasks. On Chinese-to-English translation, the absolute improvements in BLEU (case-insensitive) range from 1.2 to 2.1.

1 Introduction

The hierarchical phrase-based (HPB) model (Chiang, 2005; Chiang, 2007) has been widely adopted in statistical machine translation (SMT). It utilizes synchronous context free grammar (SCFG) rules to perform translation. Typically, there are three types of rules (see Table 1): *phrasal* rule, a phrase pair consisting of consecutive words; *hierarchical* rule, a hierarchical phrase pair consisting of both words and variables; and *glue* rule, which is used to merge phrases serially. Phrasal rule captures short distance reorderings within phrases, while hierarchical rule captures long distance reorderings between phrases. Therefore, the HPB model outperforms conventional phrase-based models on phrase reorderings.

Type	Constituent		Examples
	Word	Variable	
PR	$\surd$	-	$X \rightarrow \langle$之一, one of$\rangle$
HR	$\surd$	$\surd$	$X \rightarrow \langle X$的, of$X\rangle$
GR	-	$\surd$	$S \rightarrow \langle SX, SX\rangle$

Table 1: A classification of grammar rules for the HPB model. PR = phrasal rule, HR = hierarchical rule, GR = glue rule.

However, HPB translation suffers from a limitation, in that the phrase reorderings lack of contextual information, such as the surrounding words of a phrase and the content of sub-phrases that represented by variables. Consider the following two hierarchical rules in translating a Chinese sentence into English:

$$X \rightarrow \langle X_1 \text{ 的 } X_2,\ X_1 \text{ 's } X_2 \rangle \quad (1)$$

$$X \rightarrow \langle X_1 \text{ 的 } X_2,\ X_2 X_1 \rangle \quad (2)$$

和 俄罗斯 的 谈判
with Russia 's talks
talks with Russia

Both pattern-match the source sentence, but produce quite different phrase reorderings. The first rule generates a monotone translation, while the second rule swaps the source phrases covered by X_1 and X_2 on the target side. During decoding, the first

Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 555–563,
MIT, Massachusetts, USA, 9-11 October 2010.

rule is more likely to be used, as it occurs more frequently in a training corpus. However, the example is not a noun possessive case because the sub-phrase covered by X_1 is not a noun but a prepositional phrase. Thus, without considering information of sub-phrases, the decoder may make errors on phrase reordering.

Contextual information has been widely used to improve translation performance. It is helpful to reduce ambiguity, thus guide the decoder to choose correct translation for a source text. Several researchers observed that word sense disambiguation improves translation quality on lexical translation (Carpuat and Wu, 2007; Chan et al., 2007). These methods utilized contextual features to determine the correct meaning of a source word, thus help an SMT system choose an appropriate target translation.

Zens and Ney (2006) and Xiong et al. (2006) utilized contextual information to improve phrase reordering. They addressed phrase reordering as a two-class classification problem that translating neighboring phrases serially or inversely. They built a maximum entropy (MaxEnt) classifier based on boundary words to predict the order of neighboring phrases.

He et al. (2008) presented a lexicalized rule selection model to improve both lexical translation and phrase reordering for HPB translation. They built a MaxEnt model for each ambiguous source side based on contextual features. The method was also successfully applied to improve syntax-based SMT translation (Liu et al., 2008), using more sophisticated syntactical features. Shen et al. (2008) integrated various contextual and linguistic features into an HPB system, using surrounding words and dependency information for building context and dependency language models, respectively.

In this paper, we focus on improving phrase reordering for HPB translation. We classify SCFG rules into several reordering patterns consisting of two variables X and F (or E) [1], such as X_1FX_2 and X_2EX_1. We treat phrase reordering as a classification problem and build a MaxEnt model for each source reordering pattern based on various contextual features. We propose a method to integrate the MaxEnt models into an HPB system. Specifically:

- For hierarchical rules, we classify the source-side and the target-side into 7 and 17 reordering patterns, respectively. Target reordering patterns are treated as possible labels. We then build a classifier for each source pattern to predict phrase reorderings. This is different from He et al. (2008), in which they built a classifier for each ambiguous hierarchical source-side. Therefore, the training examples for each MaxEnt model is small and the model maybe unstable. Here, we classify source hierarchical phrases into 7 reordering patterns according to the arrangement of words and variables. We can obtain sufficient samples for each MaxEnt model from large-scale bilingual corpus.

- For glue rules, we extend the HPB model by using bracketing transduction grammar (BTG) (Wu, 1996) instead of the monotone glue rule. By doing this, there are two options for the decoder to merge phrases: serial or inverse. We then build a classifier for glue rules to predict reorderings of neighboring phrases, analogous to Xiong et al. (2006).

- We integrate the MaxEnt based phrase reordering models as features into the HPB model (Chiang, 2005). The feature weights can be tuned together with other feature functions by MERT algorithm (Och, 2003).

Experimental results show that the presented method achieves significant improvement over the baseline. On Chinese-to-English translation tasks of NIST evluation, improvements in BLEU (case-insensitive) are 1.2 on MT06 GALE set, 1.8 on MT06 NIST set, and 2.1 on MT08.

The rest of the paper is structured as follows: Section 2 describes the MaxEnt based phrase reordering method. Section 3 integrates the MaxEnt models into the translation model. In Section 4, we report experimental results. We analyze the presented method and experimental results in Section 5 and conclude in Section 6.

[1] We use F and E to represent source and target words, respectively.

Source phrase	Target phrase
X 和	X and
	with X
	between X and

Figure 1: A source hierarchical phrase and its corresponding target translation.

Source pattern	Target pattern
XF	XE
FX	EX
FXF	EXE

Table 2: A classification of the source side and the target side for the hierarchical rule that contains one variable.

Source pattern	Target pattern
	X_1EX_2
	X_2EX_1
	X_1X_2E
	X_2X_1E
	EX_1X_2
X_1FX_2	EX_2X_1
X_1FX_2F	X_1EX_2E
FX_1FX_2	X_2EX_1E
FX_1FX_2F	EX_1X_2E
	EX_2X_1E
	EX_1EX_2
	EX_2EX_1
	EX_1EX_2E
	EX_2EX_1E

Table 3: A classification of the source side and the target side for the hierarchical rule that contains two variables.

2 MaxEnt based Phrase Reordering

We regard phrase reordering as a pattern classification problem. A reordering pattern indicates an arrangement of words and variables. Although there are a large amount of hierarchical rules may be extracted from bilingual corpus, these rules can be classified into several reordering patterns (Section 2.1). In addition, we extend the HPB model with BTG, that adding an inverted glue rule to merge phrases inversely (Section 2.2). Therefore, the glue rules are classified into two patterns: serial or inverse. We then build a MaxEnt phrase reordering (MEPR) classifier for each source reordering pattern (Section 2.3). In Section 2.4, we describe contextual features.

2.1 Reordering Pattern Classification for Hierarchical Rule

Hierarchical rule, consisting of both words and variables, is of great importance for the HPB model. During decoding, words are used for lexical translation, and variables capture phrase reordering. We may learn millions of hierarchical rules from a bilingual corpus. Although these rules are different from each other, they can be classified into several reordering patterns according to the arrangement of variables and words.

In this paper, we follow the constraint as described in (Chiang, 2005) that a hierarchical rule can have at most two variables and they cannot be adjacent on the source side. We use "X" to represent the variable, and "F" and "E" to represent word strings in source and target language, respectively. Therefore, in a hierarchical rule, E is the lexical translation of F, while the order of X and E contains phrase reordering information.

For the hierarchical rule that contains one variable (see Figure 1 for example), both the source and the target phrases can be classified into three patterns (Table 2). To reduce the complexity of classification, we do not distinguish the order of word strings. For example, we consider "e_1Xe_2" and "e_2Xe_1" as the same pattern "EXE", because the target words are determined by lexical translation of source words. Our focus is the order between X and E. During decoding the phrases covered by X are dynamically changed and the contextual information of these phrases is ignored for pattern-matching of hierarchical rules.

Analogously, for the hierarchical rule that contains two variables, the source phrases are classified into 4 patterns, while the target phrases are classified into 14 patterns, as shown in Table 3. The pattern number on the source side is less than that on the target side, because on the source side, "X_1" always appears before "X_2", and they cannot be adjacent.

2.2 Reordering Pattern Classification for Glue Rule

The HPB model used glue rule to combine phrases serially. The reason is that in some cases, there are no valid translation rules that cover a source span. Therefore, the glue rule provides a default monotone combination of phrases in order to complete a translation. This is not sufficient because in certain cases, the order of phrases may be inverted on the target-side.

In this paper, we extend the glue rule with BTG (Wu, 1996), which consists of three types of rules:

$$X \rightarrow \langle \tilde{f}, \tilde{e} \rangle \tag{3}$$

$$X \rightarrow \langle X_1 X_2, X_1 X_2 \rangle \tag{4}$$

$$X \rightarrow \langle X_1 X_2, X_2 X_1 \rangle \tag{5}$$

Rule 3 is a phrasal rule that translates a source phrase $\tilde{f}$ into a target phrase $\tilde{e}$. Rule 4 merges two consecutive phrases in monotone order, while Rule 5 merges them in inverted order. During decoding, the decoder first uses Rule 3 to produce phrase translation, and then iteratively uses Rule 4 and 5 to merge two neighboring phrases into a larger phrase until the whole sentence is covered.

We replace the original glue rules in the HPB model with BTG rules (see Table 4). We believe that the extended HPB model can benefit from BTG in the following aspects:

- In the HPB model, as we mentioned, hierarchical rules are constrained in that nonterminals cannot be adjacent on the source side, i.e., the source side cannot contain "X_1X_2". One reason is that it will heavily increase the rule table size. The other reason is that it can cause a spurious ambiguity problem (Chiang, 2005). The inverted glue rule in BTG, however, can solve this problem.

- In the HPB model, only a monotone glue rule is provided to merge phrases serially. In the extended HPB model, the combination of phrases is classified into two types: monotone and inverse.

Analogous to Xiong et al. (2006), to perform context-dependent phrase reordering, we build a MaxEnt based classifier for glue rules to predict the order of two neighboring phrases. In this paper, we utilize more contextual features.

Glue Rule	Extended Glue Rule
$S \rightarrow \langle X, X \rangle$	$S \rightarrow \langle X, X \rangle$
$S \rightarrow \langle SX, SX \rangle$	$X \rightarrow \langle X_1 X_2, X_1 X_2 \rangle$
-	$X \rightarrow \langle X_1 X_2, X_2 X_1 \rangle$

Table 4: Extending the glue rules in the HPB model with BTG.

2.3 The MaxEnt based Phrase Reordering Classifier

As described above, we classified phrase reorderings into several patterns. Therefore, phrase reordering can be regarded as a classification problem: for each source reordering pattern, we treat the corresponding target reordering patterns as labels.

We build a general classification model within the MaxEnt framework:

$$P_{me}(T_\gamma | T_\alpha, \alpha, \gamma) = \frac{exp(\sum_i \lambda_i h_i(\gamma, \alpha, f(X), e(X)))}{\sum_{T_\gamma} exp(\sum_i \lambda_i h_i(\gamma, \alpha, f(X), e(X)))} \tag{6}$$

where, α and γ are the source and target side, respectively. T_α/T_γ is the reordering pattern of α/γ. $f(X)$ and $e(X)$ are the phrases that covered by X one the source and target side, respectively. Given a source phrase, the model predicts a target reordering pattern, considering various contextual features (Section 2.4).

According to the classification of reordering patterns, there are 3 kinds of classifiers:

- P_{me}^{hr1} includes 3 classifiers for the hierarchical rules that contain 1 variable. Each of the classifier has 3 labels;

- P_{me}^{hr2} includes 4 classifiers for the hierarchical rules that contain 2 variables. Each of the classifier has 14 labels;

- P_{me}^{gr} includes 1 classifier for the glue rules. The classifier has 2 labels that predict a monotone or inverse order for two neighboring phrases. This classifier is analogous to (Xiong et al., 2006).

There are 8 classifiers in total. This is much fewer than the classifiers in He et al. (2008), in which a classifier was built for each ambiguous hierarchical source side. In this way, a classifier may face the risk that there are not enough samples for training a stable MaxEnt model. While our approach is more generic, rather than training a MaxEnt model for a specific hierarchical source side, we train a model for a source reordering pattern. Thus, we reduce the number of classifiers and can extract large training examples for each classifier.

2.4 Feature definition

For a reordering pattern pair $\langle T_\alpha, T_\gamma \rangle$, we design three feature functions for phrase reordering classifiers:

- Source lexical feature, including boundary words and neighboring words. Boundary words are the left and right word of the source phrases covered by $f(X)$, while neighboring words are the words that immediately to the left and right of a source phrase $f(\alpha)$;
- Part-of-Speech (POS) feature, POS tags of the boundary and neighboring words on the source side.
- Target lexical feature, the boundary words of the target phrases covered by $e(X)$.

These features can be extracted together with translation rules from bilingual corpus. However, since the hierarchical rule does not allow for adjacent variables on the source side, we extract features for P_{me}^{gr} by using the method described in Xiong et al. (2006). We train the classifiers with a MaxEnt trainer (Zhang, 2004).

3 Integrating the MEPR Classifier into the HPB Model

The HPB model is built within the standard log-linear framework (Och and Ney, 2002):

$$Pr(e|f) \propto \sum_i \lambda_i h_i(\alpha, \gamma) \quad (7)$$

where $h_i(\alpha, \gamma)$ is a feature function and λ_i is the weight of h_i. The HPB model has the following features: translation probabilities $p(\gamma|\alpha)$ and $p(\alpha|\gamma)$, lexical weights $p_w(\gamma|\alpha)$ and $p_w(\alpha|\gamma)$, word penalty, phrase penalty, glue rule penalty, and a target n-gram language model.

To integrate the MEPR classifiers into the translation model, the features of the log-linear model are changed as follows:

- We add the MEPR classifier as a feature function to predict reordering pattern:

$$h_{me}(T_\gamma|T_\alpha) = \sum P_{me}(T_\gamma|T_\alpha, \alpha, \gamma) \quad (8)$$

  During decoding, we first classify each source phrase into one of the 8 source reordering patterns and then use the corresponding MEPR classifier to predict the possible target reordering pattern. Therefore, the contextual information guides the decoder to perform phrase reordering.

- We split the "glue rule penalty" into two features: monotone glue rule number and inverted glue rule number. These features reflect preference of the decoder for using monotone or inverted glue rules.

The advantage of our extension method is that the weights of the new features can be tuned together with the other features by MERT algorithm (Och, 2003).

We utilize a standard CKY algorithm for decoding. Given a source sentence, the decoder searches the best derivation from the bottom to top. For a source span $[j_1, j_2]$, the decoder uses three kinds of rules: translation rules produce lexical translation and phrase reordering (for hierarchical rules), monotone rule merges any neighboring sub-spans $[j_1, k]$ and $[k+1, j_2]$ serially, and inverted rule swap them. Note that when the decoder uses the monotone and inverted glue rule to combine sub-spans, it merges phrases that do not contain variables. Because the CKY algorithm guarantees that the sub spans $[j_1, k]$ and $[k+1, j_2]$ have been translated before $[j_1, j_2]$.

4 Experiments

We carried out experiments on four systems:

- HPB: replication of the Hiero system (Chiang, 2005);
- HPB+MEHR: HPB with MaxEnt based classifier for hierarchical rules, as described in Section 2.1;
- HPB+MEGR: HPB with MaxEnt based classifier for glue rules, as described in Section 2.2;
- HPB+MER: HPB with MaxEnt based classifier for both hierarchical and glue rules.

All systems were tuned on NIST MT03 and tested on MT06 and MT08. The evaluation metric was BLEU (Papineni et al., 2002) with case-insensitive matching of n-grams, where $n = 4$.

We evaluated our approach on Chinese-to-English translation. The training data contained 77M Chinese words and 81M English words. These data come from 17 corpora: LDC2002E18, LDC2002L27, LDC2002T01, LDC2003E07, LDC2003E14, LDC2004T07, LDC2005E83, LDC2005T06, LDC2005T10, LDC2005T34, LDC2006E24, LDC2006E26, LDC2006E34, LDC2006E86, LDC2006E92, LDC2006E93, LDC2004T08 (HK_News, HK_Hansards).

To obtain word alignments, we first ran GIZA++ (Och and Ney, 2000) in both translation directions and then refined the results using the "grow-diag-final" method (Koehn et al., 2003). For the language model, we used the SRI Language Modeling Toolkit (Stolcke, 2002) to train two 4-gram models on the Xinhua portion of the GigaWord corpus and the English side of the training corpus.

4.1 Statistical Information of Rules

Hierarchical Rules

We extracted 162M translation rules from the training corpus. Among them, there were 127M hierarchical rules, which contained 85M hierarchical source phrases. We classified these source phrases into 7 patterns as described in Section 2.1. Table 5 shows the statistical information. We observed that the most frequent source pattern is "FXF", which accounted for 46.1% of the total. Interestingly, "X_1FX_2", accounting for 3.7%, was the least frequent pattern. Table 6 and Table 7 show the distributions of reordering patterns for hierarchical source phrases that contain one and two variables, respectively. From both the tables, we observed that for Chinese-to-English translation, the most frequent "reordering" pattern for a source phrase is monotone translation (bold font in the tables).

Source Pattern	Percentage (%)
XF	9.7
FX	9.7
FXF	46.1
X_1FX_2	3.7
X_1FX_2F	11.9
FX_1FX_2	11.8
FX_1FX_2F	7.1

Table 5: Statistical information of reordering pattern classification for hierarchical source phrases.

#	Source		
Target (%)	FX	XF	FXF
EX	**82.8**	7	4.6
XE	6.4	**82.4**	2.9
EXE	10.8	10.6	**92.5**

Table 6: Percentage of target reordering pattern for each source pattern containing one variable.

Glue Rules

To train a MaxEnt classifier for glue rules, we extracted 65.8M reordering (monotone and inverse) instances from the training data, using the algorithm described in Xiong et al. (2006). There were 63M monotone instances, accounting for 95.7%. Although instances of inverse reordering accounted for 4.3%, they are important for phrase reordering.

4.2 Results

Table 8 shows the BLEU scores and decoding speed of the four systems on MT06 (GALE set and NIST set) and MT08. From the table, we made the following observations:

#	Source			
Target (%)	FX_1FX_2	FX_1FX_2F	X_1FX_2	X_1FX_2F
EX_1EX_2	**78.1**	3.6	4.6	1.2
EX_1EX_2E	2.1	**75.9**	0.1	1.6
EX_1X_2	6.8	0.1	2.8	0.1
EX_1X_2E	1.8	11.2	0.1	2
EX_2EX_1	2.8	1.4	2	1.2
EX_2EX_1E	1.4	2.3	0.7	1.1
EX_2X_1	0.9	0.1	2.2	0.2
EX_2X_1E	1	1.1	0.9	1.0
X_1EX_2	1.9	0.1	**71.2**	3.3
X_1EX_2E	0.7	2.1	6	**78.4**
X_1X_2E	0.1	0.1	2.8	5.9
X_2EX_1	0.9	0.4	1.6	0.7
X_2EX_1E	1.5	1.5	2.6	2.4
X_2X_1E	0.1	0.04	2.2	0.8

Table 7: Percentage of target reordering pattern for each source pattern containing two variables.

System	Test Data			Speed
	06G	06N	08	
HPB	14.19	33.93	25.85	8.7
HPB+MEHR	14.76	34.95	26.56	3.2
HPB+MEGR	15.09	35.72	27.34	2.7
HPB+MER	15.42	35.80	27.94	1.7

Table 8: BLEU percentage scores and translation speed (words/second) on test data. G=GALE set, N=NIST set. All improvements are statistically significant ($p < 0.01$). Note that MT06G has one reference for each source sentence, while the MT06N and MT08 have four references.

- The HPB+MEHR system achieved significant improvements on all test sets compared to the HPB system. The absolute increases in BLEU scores ranging from 0.6 (on 06G) to 1.0 (on 06N) percentage points. This indicates that the ME based reordering for hierarchical rules improves translation performance.

- The HPB+MEGR system achieved significant improvements over the HPB system. The absolute increases in BLEU scores ranging from 0.9 (on 06G) to 1.8 (on 06N) percentage points. The HPB+MEGR system overcomes the shortcoming of the HPB system by using both monotone glue rule and inverted glue rule, which merging phrases serially and inversely, respectively. Furthermore, the HPB+MEGR system outperformed the HPB+MEHR system.

- The HPB+MER system achieved the best performances on all test sets, with absolute increases of BLEU scores ranging from 1.2 (on 06G) to 2.1 (on 08). The system combining with ME based reordering for both hierarchical and glue rules, outperformed both the HPB+MEHR and HPB+MEGR systems.

- In addition, we found that the decoder takes more time after adding the MEPR models (the *speed* column of Table 8). The average translation speed of HPB+MER (1.7 words/second) is about 5 times slower than the HPB system (8.7 words/second). One reason is that the MEPR models utilized contextual information to compute classification scores. Another reason is that adding inverted glue rules increases search space.

5 Analysis

Experiments showed that the presented approach achieved significant gains on BLEU scores. Furthermore, we sought to explore what would happen after integrating the MEPR classifiers into the translation model. We compared the outputs of HPB and HPB+MER and observed that the translation performance are improved on phrase reordering. For example, the translations of a source sentence in MT08 are as follows [2]:

- **Src:** 韩国$_1$ 政府$_2$ 上个月$_3$ 底$_4$ 开始$_5$ 启动$_6$ 对$_7$ 朝鲜$_8$ 提供$_9$ 40万$_{10}$ 吨$_{11}$ 大米$_{12}$ 的$_{13}$ 援助$_{14}$ 计划$_{15}$

- **Ref:** At the end$_4$ of last$_3$ month$_3$, the South$_1$ Korean$_1$ government$_2$ began$_5$ a plan$_{15}$ to provide$_9$ 400,000$_{10}$ tonnes$_{11}$ of rice$_{12}$ as aid$_{14}$ to North$_8$ Korea$_8$

- **HPB:** South Korean government late last month to start with 400,000 tons of rice aid to the DPRK

- **HPB+MER**: Start at the end of last month, South Korean government plans to provide 400,000 tons of rice in aid to the DPRK

The most obvious error that the baseline system makes is the order of the time expression "*上个月底, the end of last month*", which should be either at the beginning or the end on target side. However, the baseline produced a monotone translation by using the rule "*韩国 政府 X_1, South Korean government X_1*". The HPB+MER system, however, moved the time expression to the beginning of the sentence by using the rule "*韩国 政府 X_1, X_1 South Korean government*'. The reason is that the MaxEnt phrase reordering classifier uses the contextual features (e.g. the boundary words) of the phrase covered by X_1 to predict the phrase reordering as X_1E for the source phrase FX_1.

[2]The co-indexes of the words in the source and reference sentence indicate word alignments.

6 Conclusions and Future Work

In this paper, we have proposed a MaxEnt based phrase reordering approach to help the HPB decoder select reordering patterns. We classified hierarchical rules into 7 reordering patterns on the source side and 17 reordering patterns on the target side. In addition, we introduced BTG to enhance the reordering of neighboring phrases and classified the glue rules into two patterns. We trained a MaxEnt classifier for each reordering pattern and integrated it into a standard HPB system. Experimental results showed that the proposed approach achieved significant improvements over the baseline. The absolute improvements in BLEU range from 1.2 to 2.1.

MaxEnt based phrase reordering provides a mechanism to incorporate various features into the translation model. In this paper, we only use a few feature sets based on standard contextual word and POS tags. We believe that additional features will further improve translation performance. Such features could include syntactical features (Chiang et al., 2009). In the future, we will carry out experiments on deeper features and evaluate the effects of different feature sets.

References

Marine Carpuat and Dekai Wu. 2007. Improving statistical machine translation using word sense disambiguation. In *Proceedings of EMNLP-CoNLL 2007*, pages 61–72.

Yee Seng Chan, Hwee Tou Ng, and David Chiang. 2007. Word sense disambiguation improves statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 33–40.

David Chiang, Wei Wang, and Kevin Knight. 2009. 11,001 new features for statistical machine translation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies 2009 Conference*, page 218 – 226.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 263–270.

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

Zhongjun He, Qun Liu, and Shouxun Lin. 2008. Improving statistical machine translation using lexical-

ized rule selection. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 321–328.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48–54.

Qun Liu, Zhongjun He, Yang Liu, and Shouxun Lin. 2008. Maximum entropy based rule selection model for syntax-based statistical machine translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, page 89 – 97.

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 440–447.

Franz Josef Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 295–302.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Libin Shen, Jinxi Xu, Bing Zhang, Spyros Matsoukas, and Ralph Weischedel. 2008. Effective use of linguistic and contextual information for statistical machine translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 72–80.

Andreas Stolcke. 2002. SRILM – An extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken language Processing*, volume 2, pages 901–904.

Dekai Wu. 1996. A polynomial-time algorithm for statistical machine translation. In *Proceedings of the Thirty-Fourth Annual Meeting of the Association for Computational Linguistics*.

Deyi Xiong, Qun Liu, and Shouxun Lin. 2006. Maximum entropy based phrase reordering model for statistical machine translation. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics*, pages 521–528.

Richard Zens and Hermann Ney. 2006. Discriminative reordering models for statistical machine translation. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 55–63.

Le Zhang. 2004. Maximum entropy modeling toolkit for python and c++. available at http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html.