Improving A Lexicalized Hierarchical Reordering Model Using Maximum Entropy

Vinh Van Nguyen
Lac Viet research Lab
LacViet Computing Corp
185 Giang Vo, Hanoi, VIETNAM
vinhnv2000@gmail.com

Akira Shimazu, Minh Le Nguyen
Japan Advanced Institute of Science and Technology
1-1, Asahidai, Nomi, Ishikawa, 923-1292, JAPAN
{shimazu, nguyenml}@jaist.ac.jp

Thai Phuong Nguyen
College of Technology, VNU
144 Xuan Thuy, Hanoi, VIETNAM
thainp@vnu.edu.vn

Abstract

In this paper, we present a reordering model based on Maximum Entropy. This model is extended from a hierarchical reordering model with PBSMT (Galley and Manning, 2008), which integrates syntactic information directly in decoder as features of MaxEnt model. The advantages of this model are (1) maintaining the strength of phrase based approach with a hierarchical reordering model, (2) many kinds of linguistic information integrated in PBSMT as arbitrary features of MaxEntropy model. The experiment results with English-Vietnamese pair showed that our approach achieves improvements over the system which use a lexical hierarchical reordering model (Galley and Manning, 2008).

1 Introduction

The emergence of phrase-based statistical translation (Koehn et al., 2003) has been one of the major developments in statistical approaches to translation. In PBSMT, translation of phrases (word contiguous sequences) instead of single words has some advantages, such as a robustness in word selection and local word reordering. The experiment results show that our approach achieves significant improvements over the baseline system.

Recently, in (Tillmann, 2004; Koehn et al., 2007), the lexicalized reordering models (LRMs) have been descried that it tries to predict the orientation of a phrase pair based on previous adjacent target phrase. These models distinguish three orientations of a current phrase pair with respect to the previous target phrase: (1) *monotone* (M) - the previous source phrase is previously adjacent to the current source phrase, (2) *swap* (S) - the previous source phrase is next adjacent to the current source phrase, and (3) *discontinuous* (D) - Not *monotone* or *swap*. Figure 1(1) shows an example where such a model effectively swaps the adjective phrase *"nice new"* with a noun *"house"*, and the phrase *"a"* remains in monotone order with respect to the previous phrase *"This is"*. Those lexicalized reordering models showed that improvement over PBSMT. However, those models tackled local re-orderings of neighboring phrases because they usually are fail to capture long distance reordering. In Figure 1(2), orientation of phrase *"Tom's"* should swap with the rest of the noun phrase, however, LRMs predict this orientation to discontinuous(D).

Galley and Manning (2008) extended the above models, proposed a hierarchical phrase reordering model (HRM). Their model bases on a hierarchical structure which enables phrase movements that are more complex than swaps between adjacent phrases. In Figure 1(2), their model enable to treat the adjacent phrase *"two"* and *"blue books"* as one single phrase, and the displacement of *"Tom's"* with respect to this phrase can be treated as a swap(S), demonstrated by blue color S. Similarly, orientation of "." is changed from (D) to (M). However, their model have several weaknesses as follows:

- This model estimates probabilities based on relative-frequency approach, which can suffer from the data sparseness problem. One of reasons is most of the phrase examples occur only once in the training corpus (96.5% the phrase

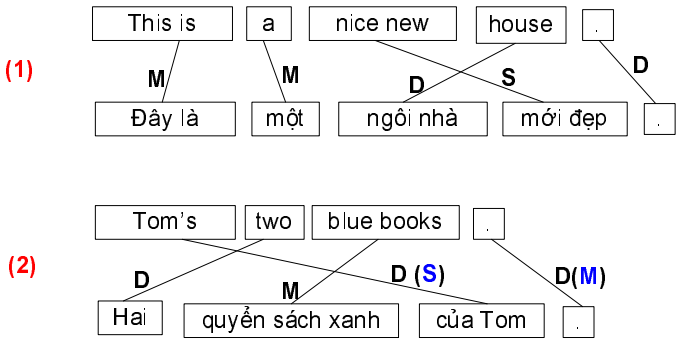

Figure 1: Phase orientation (monotone, swap, discontinuous) for English-Vietnamese translation.

examples occur only once in the training corpus "Genenal" for our experiments).

- This model do not use any linguistic information. This is poor context for predicting orientation and estimating probabilities.

In this paper, we focus on studying the improvement of the lexical reordering model. We extend the hierarchical phrase reordering model (Galley and Manning, 2008) to a new model using Maximum Entropy model for predicting orientation and estimating probabilities. We can integrate POS information, syntactic information into our framework. Moreover, probabilities is more exact and smooth because they are estimated directly from Maximum Entropy model. The experiment results with English-Vietnamese pair show that our approach achieves improvements over the system which use a lexical hierarchical reordering model (Galley and Manning, 2008).

The rest of this paper is structured as follows. Section 2 reviews related work. Section 3 briefly introduces PBSMT with lexicalized reordering models. Section 4 presents lexicalized reordering model using maximum entropy and the definition of features for integrating linguistic information into Maximum Entropy. Section 5 describes and discusses the experimental results. Finally, conclusions are given in Section 6.

2 Related Work

Decoding in PBSMT built target sentence from left to right. From current hypothesis, it is important to identify source phrase which need be translated. Several researcher (Tillmann, 2004; Koehn et al., 2005) proposed a powerful model called lexicalized reordering model for predicting orientation of source phrase as descried above. Lexicalized reordering model learns local orientations (monotone or swap or discontinue) with probabilities for each bilingual phrase from training data.

(Xiong et al., 2006; Zens and Hey, 2006) applied Maximum Entropy (ME) model for phrase reordering. They used ME for estimating distortion probability. However, estimation is local, because the next phrase only depends on the current phrase. So, as a result, their systems are not robust to unseen phrases.

Galley and Manning (2008) extended the above models, proposed a hierarchical phrase reordering model (HRM). Their model is a more powerful model because this model bases on a hierarchical structure which enables phrase movements that are more complex than swaps between adjacent phrases. However, the limitation of their model is the sparseness data problem and the poor of context information because their model estimates and learns orientations only based on training data.

Our model is most similar to (Galley and Manning, 2008).

3 Lexicalized Reordering Models

The limitation of distance based distortion modeling are stated in lexical distortion models (Tillmann, 2004; Koehn et al., 2005), which directly learn the probabilities for a given phrase being reordering relative to adjacent phrases.

Given a source sentence f, which is to be translated into a target sentence e. The current state-of-

the-art phrase based systems are log-linear models of the conditional probability $Pr(f|e$:

$$Pr(f|e) = \frac{\exp \sum_i \lambda_i h_i(e,f)}{\sum_{e'} \exp \lambda_i h_i(e',f)} \quad (1)$$

where the $h_i(e,f)$ are arbitrary feature functions over sentence pairs; the λ are weights on feature functions $h_i(e,f)$. The decoder searches for the most probable translation $\hat{e}$ according to the following equation:

$$\hat{e} = \underset{e}{\operatorname{argmax}} \left\{ \exp \sum_i \lambda_i h_i(e,f) \right\} \quad (2)$$

The features include lexicalized reordering models, which are parameterized as follows: given an source sentence f, a sequence of target language phrases $e = (\bar{e}_1, \ldots, \bar{e}_n)$ currently hypothesized by the decoder, and phrase alignment $a = (a_1, \ldots, a_n)$ that defines a source $\bar{f}_{a_i}$ for each translated phrase $\bar{e}_i$, those models estimate the probability of a sequence of orientation $o = (o_1, \ldots, o_n)$ as follows:

$$Pr(o|e,f) = \prod_{i=1}^{n} p(o_i|\bar{e}_i, \bar{f}_{a_i}) \quad (3)$$

in which, each o_i takes values over the set of possible orientation $\Delta = M, S, D$. When collecting phrase pairs, can classify them into these three categories based on:

- $o_i = M$ if $a_i - a_{i-1} = 1$
- $o_i = S$ if $a_i - a_{i-1} = -1$
- $o_i = S$ if ($a_i - a_{i-1} \neq 1$ and $a_i - a_{i-1} \neq -1$)

At decoding step, we adapt the approach of Moses, which assign three distinct parameters $(\lambda_m, \lambda_s, \lambda_d)$ for the three feature functions:

- $f_m = \sum_{i=1}^{n} logp(o_i = M|\bar{e}_i, \bar{f}_{a_i})$
- $f_s = \sum_{i=1}^{n} logp(o_i = S|\bar{e}_i, \bar{f}_{a_i})$
- $f_d = \sum_{i=1}^{n} logp(o_i = D|\bar{e}_i, \bar{f}_{a_i})$

In order to integrate $p(o_i|\bar{e}_i, \bar{f}_{a_i})$ into formulation 1 in decoding, we need to compute those probabilities. A simple way based on relative-frequency approach computes those probabilities as follows:

$$p(o_i|\bar{e}_i, \bar{f}_{a_i}) = \frac{Count(o_i, \bar{e}_i, \bar{f}_{a_i})}{\sum_o Count(o, \bar{e}_i, \bar{f}_{a_i})} \quad (4)$$

where $Count(x)$ is a number of times of x which occur into the training data.

We calculate $p(o_i|\bar{e}_i, \bar{f}_{a_i})$ based on a previous phrase alignment a_{i-1} of a_i. We assume that a_i have m previous phrase alignments. Let a_{i-1}^k $(k = 1, \ldots, m)$ be k-th previous phrase alignment of a_i, we have:

$$p(o_i|\bar{e}_i, \bar{f}_{a_i}) = \frac{\sum_{k=1}^{m} Count(o_i, \bar{e}_i, \bar{f}_{a_i}, a_{i-1}^k)}{\sum_{k=1}^{m} \sum_o Count(o, \bar{e}_i, \bar{f}_{a_i}, a_{i-1}^k)} \quad (5)$$

However, above way meets several limitations as described in Section 1. It is very reasonable to use maximum entropy model to integrate features to predict reordering of phrases. Under the Maximum Entropy, we define:

$$p(o_i|\bar{e}_i, \bar{f}_{a_i}) = \frac{exp(\sum_j \theta_j h_j(o_i, \bar{e}_i, \bar{f}_{a_i}))}{\sum_o exp(\sum_j \theta_j h_j(o, \bar{e}_i, \bar{f}_{a_i}))} \quad (6)$$

where the Kronecker function h_i which takes values over 0,1 are model features and the θ_i are weight of the model features which can be trained by different methods (Sha and Pereira, 2003).

4 Lexicalized Reordering Model into PBSMT using Maximum Entropy

4.1 Model

In this section, we focus on using contextual information to help the HRM compute probabilities and prediction orientation of phrases. We consider the orientation of phrases as a multi-class classification task: the orientation of phrases (M,S,D) is label. Thus during decoding, a good way to tackle the classification problem is the maximum entropy approach:

$$p(o_i|\bar{e}_i, \bar{f}_{a_i}) = \frac{exp(\sum_j \theta_j h_j(o_i, \bar{e}_i, \bar{f}_{a_i}))}{\sum_o exp(\sum_j \theta_j h_j(o, \bar{e}_i, \bar{f}_{a_i}))} \quad (7)$$

We use linguistic information of source phrases to integrate HRM. When this model predict orientation of source phrases, linguistic information such as

POS tagger, syntax help usefully to decide orientation of phrases. To avoid a complicated linguistic information, each source phrase, we use three linguistic information from source syntactic subtree (subsume from phrase):

1. Head word of phrase (HW)
2. The part of speech tag of head word (TG)
3. Syntactic label of phrase (SL)

During the process of extracting features, we must annotated billing phrases given a source sentence and its parse tree. The implementation of annotating labels of phrases is as follows:

- if subtree st spans exactly a phrase p then we get (HW, TG, SL) from subtree to a phrase.
- if subtree does not span a phrase p (a phrase p is non-syntactic) then we choose the smallest subtree sst subsume phrase p. We get (HW, TG, SL) from subtree sst to a phrase p (if $HW \notin p$, we choose the first word of a phrase p as HW).

Features

We calculate $p(o_i|\bar{e}_i, \bar{f}_{a_i})$ based on a previous phrase alignment a_{i-1} of a_i. With each phrase pair $(\bar{e}_i, \bar{f}_{a_i})$, we have m examples of phrase pairs $(\bar{e}_i, \bar{f}_{a_i}, a_{i-1}^k, a_i)$ $(k = 1, \ldots, m)$. Each example of phrase pairs, we extract features based on (HW, TG, SL) of $\bar{f}_{a_i}$ and $\bar{f}_{a_{i-1}^k}$ for our Maximum Entropy-based reordering model. We use two temples of features: single features and combine features. Each phrases have three linguistic elements. Therefore, examples of phrase pairs have nine features in total. In other words, each phrase pair $(\bar{e}_i, \bar{f}_{a_i})$ have $9m$ features in total.
For example with six single features:

$$h_1(o, \bar{e}_i, \bar{f}_{a_i}, a_{i-1}^k, a_i) = \begin{cases} 1 & \text{HW}(\bar{f}_{a_i}) = w_1, o = o_i \\ 0 & \text{otherwise} \end{cases}$$

$$h_2(o, \bar{e}_i, \bar{f}_{a_i}, a_{i-1}^k, a_i) = \begin{cases} 1 & \text{TG}(\bar{f}_{a_i}) = tg_1, o = o_i \\ 0 & \text{otherwise} \end{cases}$$

$$h_3(o, \bar{e}_i, \bar{f}_{a_i}, a_{i-1}^k, a_i) = \begin{cases} 1 & \text{SL}(\bar{f}_{a_i}) = sl_1, o = o_i \\ 0 & \text{otherwise} \end{cases}$$

$$h_4(o, \bar{e}_i, \bar{f}_{a_i}, a_{i-1}^k, a_i) = \begin{cases} 1 & \text{HW}(\bar{f}_{a_{i-1}^k}) = w_1, o = o_i \\ 0 & \text{otherwise} \end{cases}$$

$$h_5(o, \bar{e}_i, \bar{f}_{a_i}, a_{i-1}^k, a_i) = \begin{cases} 1 & \text{TG}(\bar{f}_{a_{i-1}^k}) = tg_1, o = o_i \\ 0 & \text{otherwise} \end{cases}$$

$$h_6(o, \bar{e}_i, \bar{f}_{a_i}, a_{i-1}^k, a_i) = \begin{cases} 1 & \text{SL}(\bar{f}_{a_{i-1}^k}) = sl_1, o = o_i \\ 0 & \text{otherwise} \end{cases}$$

For example with three combine features:

$$h_7(o, \bar{e}_i, \bar{f}_{a_i}, a_{i-1}^k, a_i) = \begin{cases} 1 & HW(\bar{f}_{a_i}) = w_1, HW(\bar{f}_{a_{i-1}^k}) = w_2, o = o_i \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

$$h_8(o, \bar{e}_i, \bar{f}_{a_i}, a_{i-1}^k, a_i) = \begin{cases} 1 & TG(\bar{f}_{a_i}) = tg_1, TG(\bar{f}_{a_{i-1}^k}) = tg_2, o = o_i \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

$$h_9(o, \bar{e}_i, \bar{f}_{a_i}, a_{i-1}^k, a_i) = \begin{cases} 1 & SL(\bar{f}_{a_i}) = sl_1, SL(\bar{f}_{a_{i-1}^k}) = sl_2, o = o_i \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

4.2 Training

First, each phrase pair, we extract examples of phrase pairs. Second, we extract features from those examples. Finally, we compute $p(o_i|\bar{e}_i, \bar{f}_{a_i})$ using Maximum Entropy model. We assume that phrase e_i spans the word range $s, \ldots, t$ in the target sentence e and that the phrase $\bar{f}_{a_i}$ spans the range $u, \ldots, v$ in the source sentence f. All examples of phrase pairs in this paper are extracted according to the phrase-extract algorithm (Och and Ney, 2004), with maximum length set to 8.

We identify orientation of phrases using hierarchical orientation model described in (Galley and Manning, 2008). This model analyzes alignments beyond adjacent phrases. Specifically, orientation is set to $o_i = M$ if the phrase extract algorithm is able to extract a phrase pair at $(s-1, u-1)$ given no constraint on maximum phrase length (if orientation of phrase pair e_1, f_1 at $(s-1, u-1)$ is M then orientation of a phrase pair e_2, f_2 (e_2 is sub-phrase of e_1) at $s-1$ is M). Orientation is S if the same is true at $(s-1, v+1)$, and orientation is (D) otherwise.

We induce features as descried in Section 4.1 from examples of phrase pairs described above. Then we use the open source toolkit for Maximum Entropy[1] to train Maximum Entropy model for reordering model. We set the iteration number to 100 and Gaussian prior to 1.

4.3 Decoding

In the decoding process, we need to find $\hat{e}$ according to formulation 1. We develop our decoder PBSMT which adapts Pharaoh decoder (Koehn, 2004). To integrate HRM model into decoding, we compute reordering score with HRM model. In other words, we identify $p(o_i|\bar{e}_i, \bar{f}_{a_i})$. For computing those probabilities, the model must identify contiguous blocks-monotone (M) or swap (S) that may be merged into hierarchical blocks. We adapt the way in (Galley and Manning, 2008; Zhang et al., 2008), we use an instance of the shift-reduce parsing algorithm, and relies on a stack (Stk) of source substring that have already been translated. Each time the decoder adds a new block to the current translation hypothesis, it shifts the source language indices of the block into S, then repeatedly tries reducing the top two elements of S if they are contiguous. We need not to store target language indices into the stack because the decoder proceeds left to right, and thus successive blocks are always contiguous according to the target language.

For example: A given source sentence in English "Do you know what time the film begins ?" and translation sentence in Vietnamese "Ban biet bo_phim bat_dau may gio khong ?". We demonstrate the steps for this translation process. Figure 3 describes an example of the execution of this algorithm for the translation output shown in Figure 2, which is implemented by a PBSMT decoder integrating hierarchical reordering model. The first column shows target phrases which the decoder proceeds left to right. Implementation column includes shift (S), reduce (R), and accept (A) for operating the stack Stk. The source and stack columns contain source language spans (the word ranges of source phrases in source sentence), which is the information needed to determine whether two given blocks are contiguous. o_i column shows the label is predicted by the hierarchical model by comparing the current block to the hierarchical phrase that is at the top of the stack. The decoder successively pushes source-language spans [2-2], [3-3], which are successively merged into [2-3], and correspond to monotone orientations. It then encounters a discontinuity that prevents the next block [6-7] from being merged with [2-3]. Next, the decoder merged [8-8] with [6-7] into [6-8] with monotone orientation, and then merged [4-5] with [6-8] into [4-8] with swap orientation. As the decoder reaches the last phrase of the sentence (*"khong"*), corresponding to source-language spans [1-1] which is successively merged with [2-8], yielding a stack that contains only [1-8].

5 Experiments

5.1 Data sets

We conducted the experiments with English-Vietnamese pair. We used the English-Vietnamese corpus, which was collected from daily newspapers (named "General") (Nguyen et al., 2007). This corpora, which includes $55,341$ sentences, are split into training sets, development test sets, the test sets. Data sets are described in Tables 1 and corpus statistics are shown in Table 2.

5.2 BLEU score

We carried out the experiments on a PC with Pentium IV processor 3.4Gz, RAM memory 2GB. We ran GIZA++ (Och and Ney, 2003) on the training corpus in both directions using its default setting, and applied the refinement rule "grow-diag-final" (Koehn et al., 2003) to obtain a single many-to-many word alignment for each sentence pair. For learning language models, we used the SRILM toolkit (Stol-

[1] http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html

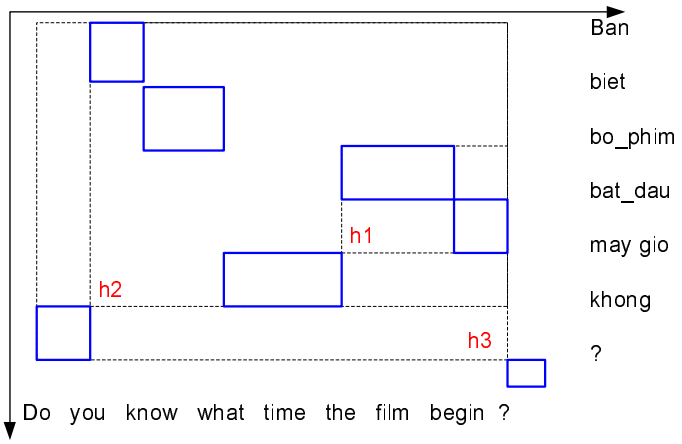

Figure 2: Hierarchical phrase h_1 and h_2 show that *"may gio"* and *"khong"* have a *swap* orientation. Whereas, $h3$ shows that *"?"* is *monotone* orientation.

Target phrase	Source spans	Implementation	o_i	Stack (Stk)
Ban	[2-2]	S	M	
biet	[3-3]	R	M	[2-2]
bo_phim	[6-7]	S	D	[2-3]
bat_dau	[8-8]	R	M	[6-7], [2-3]
may gio	[4-5]	R	S	[6-8], [2-3]
khong	[1-1]	R,R	S	[2-8]
?	[9-9]	R,A	M	[1-8]

Figure 3: The shift-reduce parsing algorithm for identifying hierarchical blocks with example in Figure 2.

cke, 2002). For MT evaluation, we used the BLEU measure (Papineni et al., 2002) calculated by the NIST script version 11b. For parsing the training set of English sentences, we used a the state-of-the-art statistical English (Charniak, 2000). Then we identify a triple (HW, TG, SL) of examples of phrase pairs according to the way described in Section 4.1.

The translation results are presented in Table 3. The baseline system is a non-monotone translation system, in which the decoder does reordering on the target language side (we adapted the beam search decoding algorithm (Koehn, 2004)). Additionally, we also compare our systems with two systems: (1) the state of the art PBSMT system - Moses (Koehn et al., 2007), which uses a lexicalized reordering model; (2) the HRM system, which uses a lexicalized hierarchical reordering model (Galley and Manning, 2008). The system which use our method named MEM. The BLEU score of HRM and MEM systems are 35.39 and 36.14 absolute points, which improved by 0.64 points and 1.39 points compared with the Moses system. The BLEU scores of MEM system improved by 0.75 points compared with the HRM system.

Our method is effective (improvement over HRM model with 0.75 point). Because a number of of examples of phrase pairs which occur at least 10 times is 0.1% and a number of examples of phrase pairs which occur once is 96.5%, relative-frequency based probabilities with HRM model causes errors. For affirming our method effective, we plan to carry out experiments for our method with a large corpus, such as English-Japanese.

6 Conclusion

In this paper, we extend a hierarchical phrase reordering model (Galley and Manning, 2008), which propose a framework for predicting orientation and estimating probabilities base on Maximum Entropy model. We can integrate POS information, syntactic information into our framework. The experiment results with English-Vietnamese pair show that our

Table 1: Corpora and data sets (sentences)

Corpus	Sentence pairs	Training set	Dev set	Test set
General	55,341	54,642	200	499

Table 2: Corpus statistics of English-Vietnamese translation task.

		English	Vietnamese
Training	Sentences	54,642	
	Average sentence length	11.2	10.6
	Words	614,578	580,754
	Vocabulary	23,804	24,097
Test	Sentences	499	
	Average sentence length	11.2	10.5
	Words	5620	6240
	Vocabulary	1844	1851

Table 3: Translation performance for the English-Vietnamese task

Corpus	**Method**	**BLEU score**
General	Baseline	34.07
	Moses	34.75
	HRM	35.39
	Our method (MEM)	36.14

approach achieves improvements over the system which use a lexical hierarchical reordering model (Galley and Manning, 2008). In future, we also plan to experiment with a set of the richer features described in (Liu et al., 2008).

References

Eugene Charniak. 2000. A maximum-entropy-inspired parser. In *Proceedings of the ANLP-NAACL 2000*, pages 132–139.

Michel Galley and Christopher D. Manning. 2008. A simple and effective hierarchical phrase reordering model. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 848–856, Honolulu, Hawaii, October. Association for Computational Linguistics.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of HLT-NAACL 2003*, pages 127–133. Edmonton, Canada.

Philipp Koehn, Amittai Axelrod, Alexandra Birch Mayne, Chris Callison-Burch, Miles Osborne, David Talbot, and Michael White. 2005. Edinburgh system description for the 2005 nist mt evaluation. In *Proceedings of Machine Translation Evaluation Workshop 2005*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL, Demonstration Session*.

Philipp Koehn. 2004. Pharaoh: A beam search decoder for phrase-based statistical machine translation models. In *Proceedings of AMTA*, pages 115–124.

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*, pages 89–97, Honolulu, Hawaii, October. Association for Computational Linguistics.

Phuong Thai Nguyen, Akira Shimazu, Le-Minh Nguyen,

and Van-Vinh Nguyen. 2007. A syntactic transformation model for statistical machine translation. *International Journal of Computer Processing of Oriental Languages (IJCPOL)*, 20(2):1–20.

Franz J. Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

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

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

Fei Sha and Fernando Pereira. 2003. Shallow parsing with conditional random fields. In *Proceedings of HLT-NAACL03*, pages 213–220.

Andreas Stolcke. 2002. Srilm - an extensible language modeling toolkit. In *Proceedings of International Conference on Spoken Language Processing*, volume 29, pages 901–904.

Christoph Tillmann. 2004. A unigram orientation model for statistical machine translation. In Daniel Marcu Susan Dumais and Salim Roukos, editors, *Proceedings of HLT-NAACL 2004: Short Papers*, pages 101–104, Boston, Massachusetts, USA, May 2 - May 7. Association for Computational Linguistics.

Deyi Xiong, Qun Lui, and Shouxun Lin. 2006. Maximum entropy based phrase reordering model for statistical machine translation. In *Proceedings of ACL'06*, pages 521–528.

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

Hao Zhang, Daniel Gildea, and David Chiang. 2008. Extracting synchronous grammar rules from word-level alignments in linear time. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 1081–1088, Manchester, UK, August. Coling 2008 Organizing Committee.