Function Word Generation in Statistical Machine Translation Systems*

Lei Cui†, **Dongdong Zhang**‡, **Mu Li**‡, and **Ming Zhou**‡

†School of Computer Science and Technology
Harbin Institute of Technology, Harbin, China
leicui@hit.edu.cn
‡Microsoft Research Asia
Beijing, China
{dozhang,muli,mingzhou}@microsoft.com

Abstract

Function words play an important role in sentence structures and express grammatical relationships with other words. Most statistical machine translation (SMT) systems do not pay enough attention to translations of function words which are noisy due to data sparseness and word alignment errors. In this paper, a novel method is designed to separate the generation of target function words from target content words in SMT decoding. With this method, the target function words are deleted before the translation modeling while in SMT decoding they are inserted back into the translations. To guide the target function words insertion, a new statistical model is proposed and integrated into the log-linear model for SMT, which can lead to better reordering and partial hypotheses ranking. The experimental results show that our approach improves the SMT performance significantly on Chinese-English translation task.

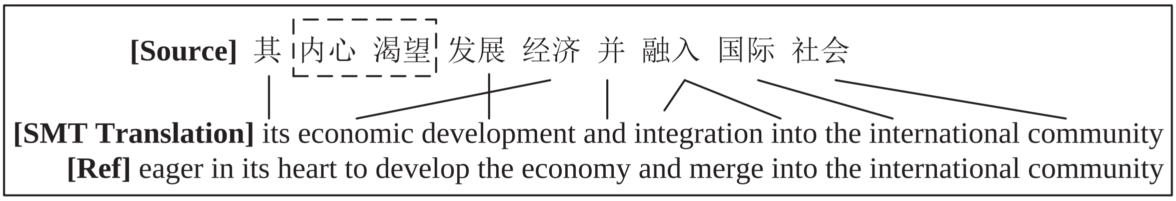
1 Introduction

Function words belong to a relatively closed set with very high frequencies in a language compared to content words such as nouns, verbs, adjectives and most adverbs. They often play an important role in sentence structures and express grammatical relationships with other words within a sentence, but have few semantic meanings or have multiple meanings. In most SMT systems (Koehn et al., 2003; Och and Ney, 2004; Chiang, 2007), function words are processed in the same way as content words. Their translation knowledge is automatically learnt via word alignment followed by translation modeling. Such training procedure ignores the specialties of function words. In practice, many function words do not have the exact counterparts in the other language and will not align to any words (i.e. align to NULL) in the results of word alignment. Furthermore, due to the high frequencies of function words, they could be associated with any content words to form bilingual phrases which might be quite noisy.

Consequently, many target function words may be either missing or inappropriately generated in the translations. This not only degrades the readability but also impairs the quality of content word translations. For example, in Figure 1, the translation of a Chinese phrase “内心 渴望” is missing because no appropriate target function words are generated. Both of the source words “内心” (heart) and “渴望” (eager) are translated into NULL according to our phrase translation table, where all the translation pairs are ranked based on their probabilities automatically learnt from our training corpus. Although the probabilities of the NULL hypotheses are quite low, it is found in our phrase table that if no appropriate target function words connect those hypotheses which have relatively high probabilities, the language model prefers the shorter hypotheses and assigns a higher probability to the translation ”its economic development” rather than other translations.

However, if the target function word ”in” could be generated between the translations of “内心” (heart) and “渴望” (eager), the SMT decoder

*This work has been done while the first author was visiting Microsoft Research Asia.



[Source] 其 内心 渴望 发展 经济 并 融入 国际 社会

[SMT Translation] its economic development and integration into the international community

[Ref] eager in its heart to develop the economy and merge into the international community

Figure 1: The translation of source phrase is missing in the SMT output.

	其	内心	渴望	发展 经济
Hypothesis 1	its ||| 0.36	mind ||| 0.26	eager ||| 0.33	economic development ||| 0.23
Hypothesis 2	NULL ||| 0.21	the heart ||| 0.18	desire ||| 0.20	develop the economy ||| 0.15
Hypothesis 3	his ||| 0.13	NULL ||| 0.09	NULL ||| 0.07	by developing the economy ||| 0.11
...	...	...	...	...

Table 1: Translation hypotheses in the baseline SMT system

is able to produce a perfect partial hypothesis ”eager in its heart”. Unfortunately, in most state-of-the-art SMT models, there is no specific mechanism to generate such important target function words explicitly and appropriately. All the target function words can only be produced either from the translation of source function words directly or as the consequence of content word translation.

According to our analysis, the incompleteness of target function word generation is mainly caused by the noisy translation knowledge automatically learnt based on word alignment.Table 2 gives the alignment statistics on top *N* function words with high frequencies in Chinese-English parallel training corpus. For example, when considering the top eight function words, about 63.9% of Chinese function word occurrences are not aligned to any English words and about 74.5% of Chinese sentences contain at least one unaligned Chinese function word. On the English side, about 36.5% of English function word occurrences are not aligned to any Chinese words and about 88.8% of English sentences contain at least one unaligned English function word. Besides, we also investigated the alignment quality of those aligned function words coming from 200 randomly selected bilingual sentence pairs. We found that still lots of function words aligned incorrectly as shown in Table 3. The survey results illustrated that the translation knowledge of function words are not reliable enough for SMT systems. Therefore, beyond the standard SMT models, extra efficient models are also needed to process function words separately for better performance.

# Function words	**Chinese**		**English**	
	Words	**Sents**	**Words**	**Sents**
$N = 8$	63.9%	74.5%	36.5%	88.8%
$N = 16$	53.3%	76.9%	36.4%	90.5%
$N = 32$	48.2%	79.1%	36.0%	91.3%
$N = 64$	41.6%	81.2%	36.0%	91.9%

Table 2: Many function words in each language get no links after word alignment.

# Function words	**Chinese**		**English**	
	Words	**Sents**	**Words**	**Sents**
$N = 8$	24.4%	35.5%	19.8%	71.0%

Table 3: The alignment error ratio of those aligned function words from 200 bilingual sentences on the word-level and sentence-level statistics.

In this paper, we try to explore the research on the processing of function words. A novel method is proposed to generate target function words in SMT outputs. It works as on-the-fly generation rather than as post-generation over SMT outputs. There are two steps in our method to handle target function words. In the first step, the target function words are removed before conducting word alignment and translation modeling. In the second step, those removed target function words are carefully inserted back into the partial hypotheses during SMT decoding. The purpose of the first step is to maintain the translation adequacy while alleviating the noisy impacts of function words as much as possible. Meanwhile, the second step is aim to recover those function words to make translation results more fluent. Intuitively, it is expected that the correct insertion of target function words can lead to better reordering and better partial hypotheses ranking in SMT decoding. The experimental results show that our method brings significant BLEU improvements over the baseline system.

2 Related Work

Some previous work has been done to improve SMT performance by leveraging function words. (Chang et al., 2009) studied the special source function words translation, such as the Chinese function word "的". (Setiawan et al., 2007; Setiawan et al., 2009) use function words to get better reordering in both phrasal SMT and hierarchical SMT systems. In addition, (Hermjakob, 2009) proposed to improve word alignment by separately considering function words and content words. These previous work only used function words as lexical anchors, which is different from our work in this paper.

Moreover, some other work focused on function words insertion and deletion. (Li et al., 2008) proposed three models to address spurious source words deletion during SMT decoding. The method brings significant improvements on Chinese-English translation task. (Zhang et al., 2008) tried to generate Chinese measure words in the target-side of English-Chinese translation task. They proposed a statistical model to calculate the probability of measure word generation by utilizing lexical and syntactic knowledge. The method works as a post-generation step over the decoder's output. High precision and recall for measure word generation can be achieved in their experimental results. As function words are more flexible than measure words, the generation of function words faces more challenges. In addition, (Menezes and Quirk, 2008) introduced an extension approach to the syntactic-based SMT system that allows structural word insertion and deletion. The effectiveness of these methods motivates us to address the generation of target function words in the phrase-based SMT which is a popular system in both academic and industrial areas.

3 Our Method

Our method focuses on the processing of target function words, including the deletion and the insertion. The deletion takes place during the model training, where the target function words are removed from the training data before conducting translation modeling. The insertion is performed during SMT decoding, where the target function words that have been deleted are inserted into appropriate positions in the partial hypotheses. To predict where and which target function words are inserted, a statistical model is proposed and trained with rich contextual information. It is integrated into the log-linear model of SMT framework, which is expected to potentially provide useful information for both better reordering and partial hypotheses ranking.

3.1 Function word deletion

Function words have high frequencies in both source and target languages. As investigated in Section 1, the translations of source and target function words are noisy. In our work, we only delete target function words rather than both of them during model training. The reason lies in that source function words are well organized to express the structure of a sentence. Moreover, they can provide context information for predicting word reordering in SMT decoding.

In standard SMT models, target function words can only be introduced with the translation of source words because no explicit mechanism introduce them by {NULL} in the source-side (i.e., some source words must be consumed to generate the target function words). However, the portion of target function words introduced by source words is limited due to data sparseness. Even worse, sometimes source words even introduce incorrect target function words which may cause translation errors. For example, in Table 1, the occurrence of the target function word "by" introduced by the source phrase "发展 经济" may lead to bad translation results.

To alleviate the noises caused by target function words, it had better remove them before translation modeling. To choose which ones should be deleted, we collect the candidate target function words that are frequently unaligned in word alignment thus more prone to be spurious words (Li et al., 2008). In particular, only a portion of whole training data is leveraged to conduct word alignment for the collection.

Once the set of target function words to be deleted is determined, they will be removed from whole training data before re-conducting word alignment and model training. Especially, given any two contiguous words, if both of them belong to the determined target function words set, none of them will be deleted from the training corpus.

3.2 Function word insertion

Those target function words removed during model training will be recovered in SMT decoding. Thus, there are three questions arising with regard to the insertion of target function words:

1. Where are the appropriate positions in the hypotheses to insert target function words?

2. Given the position, which target function word should be inserted?

3. How the insertion is performed to be seamlessly integrated into the SMT framework?

In the following sections, we answer these questions. Potentially, appropriate insertions of target function words may lead to better reordering and prevent pruning promising partial hypotheses. Figure 2 illustrates an example of appropriate target function words insertion, where two function words ”in” and ”to” are inserted in different decoding stages. The insertion of the word ”in” leads to a reasonable reordering between the phrases ”its heart” and ”eager”. And the word ”to” helps to produce a promising hypothesis ”eager in its heart to develop the economy”. Such insertions make sure that the promising hypotheses are not pruned based on the ranking score demonstrated in Table 1.

3.3 Insertion position

In our framework, target function words can be inserted at any possible positions between two contiguous target words in SMT decoding. To be efficient, we only consider some promising positions whose surrounding words co-occur in the training corpus. The context information of the deleted function words is recorded during model training, which is leveraged to guide the insertion of target function words later. In practice, we keep the leftmost and rightmost words surrounding the target function words that have been deleted. For example, suppose $w_1w_2w_3$ are three consecutive target words occurring in the training corpus and w_2 is the function word to be deleted, then a key-value pair $\langle w_1w_3, w_2\rangle$, called ***Insertion Index***, will be maintained as a clue for inserting w_2 between w_1 and w_3 once the word pair w_1w_3 is encountered in the hypotheses. Especially, w_1 and w_3 may represent the sentence boundary words $\langle s\rangle$ and $\langle /s\rangle$ when the deleted target function words locate in the boundaries of the sentences. The possibility of inserting w_2 depends on the prediction model presented in the next section.

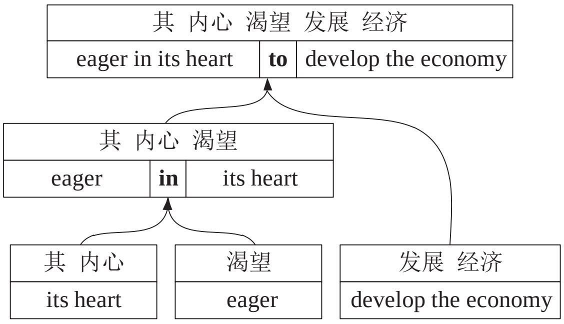


Figure 2: An example of function word insertion.

3.4 Insertion model

A statistical target function word insertion model, denoted by ***TFWIM***, is proposed to predict the probability of the target function words insertion. Given the contextual information *C*, TFWIM is computed by the maximum entropy (ME) approach as follows:

$$P(w|C) = \frac{\exp[\sum_i \lambda_i h_i(w, C)]}{\sum_{w' \in W \bigcup \{NULL\}} \exp[\sum_i \lambda_i h_i(w', C)]} \tag{1}$$

where h_i is a feature function, λ_i is the weight of h_i, W is the set of candidate target function word to be inserted. In addition, a special word $NULL$ is also included in the candidate set which stands for the cases where no function words are inserted. Explicitly, TFWIM can be easily integrated into the log-linear model for SMT.

If each distinct function word is regarded as a single class, TFWIM can be considered as a multi-class classification problem. Especially, when $|W| = 1$ in Equation (1), the TFWIM is reduced to a binary classification problem.

3.5 Model training

A training instance for TFWIM consists of a label and corresponding contextual information associated with the target function word. Let $w_1w_2w_3$ be three contiguous target words in the original training corpus and w_2 be a target function word deleted during model training, and $C(w)$ be the context information of w. To train TFWIM, $\langle w_2, C(w_2)\rangle$ is constructed as a training instance for inserting w_2 between w_1 and w_3. Meanwhile, for each occurrence of w_1w_3 in the original training corpus, its contextual information is used to construct a instance $\langle NULL, C(w_1w_3)\rangle$ as it indicates that the function word w_2 should not occur between w_1 and w_3. Also, w_1 and w_3 may stand for the sentence boundary words of $\langle s\rangle$ and $\langle /s\rangle$ respectively when the deleted target function words locate in the boundaries of training sentences. Exceptionally, for the occurrences that contain two contiguous function words, their context information cannot be used to construct the training instances for TFWIM because neither of them is deleted during model training.

In our work, all the training instances of TFWIM are automatically constructed from the target-side of bilingual training data. Only target context information is leveraged to predict function words insertion during SMT decoding, while source context information is excluded due to the deficient alignment results during decoding. Naturally, ME approach is able to leverage a variety of features together to predict the probability of each class. We consider lexical and part-of-speech (POS) features rather than other complicated syntactic features, which bring minimal overhead to the SMT system and are illustrated in Table 4.

In general, the lexical features are necessary to predict the function words insertion as they are sensitive enough to the changes of contexts. Meanwhile, the POS features are good at capturing the positions and type of function words. Furthermore, they can well distinguish the same words with different semantic meanings.

Type	Name	Description
Lexical Features	W_{w-2}	The second word to the left of w
	W_{w-1}	The word to the left of w
	W_{w+1}	The word to the right of w
	W_{w+2}	The second word to the right of w
POS Features	P_{w-2}	POS tag of W_{w-2}
	P_{w-1}	POS tag of W_{w-1}
	P_{w+1}	POS tag of W_{w+1}
	P_{w+2}	POS tag of W_{w+2}

Table 4: Feature template (w is the target function word to be deleted during the translation modeling.).

3.6 Integration into SMT decoder

Suppose the hypothesis of span w_i^j is generated by the partial hypotheses of two consecutive sub-spans w_i^k and w_{k+1}^j during SMT decoding, where w_i^k and w_{k+1}^j can construct larger hypothesis w_i^j in the order of either monotone or inverse. For the monotone order, the function word might be inserted between the words w_k and w_{k+1}, while for the inverse order, the function word might be inserted between the words w_j and w_i. As mentioned in Section 3.3, based on the Insertion Index, w_kw_{k+1} and w_jw_i are two keys to decide whether it is possible to insert function words. Once the keys exist in the Insertion Index, TFWIM will be used to calculate the probability of function words insertion. Otherwise, no function words will be inserted at all. For each kind of function word to be inserted, there is a corresponding Insertion Index to identify the possible insertion positions.

TFWIM can be easily integrated into the standard log-linear model for SMT. In Equation (2), two new features are added into the log-linear model:

$$\begin{aligned}\hat{e}_1^I &= \arg\max_{e_1^I}\{Pr(e_1^I|f_1^J)\}\\ &= \arg\max_{e_1^I}\{\sum_{m=1}^{M}\lambda_m h_m(e_1^I, f_1^J)\}\\ &= \arg\max_{e_1^I}\{\sum_{m=1}^{N}\lambda_m h_m(e_1^I, f_1^J)\\ &+ \lambda_{N+1}h_{N+1}(e_1^I, f_1^J) + \lambda_{N+2}h_{N+2}(e_1^I, f_1^J)\}\end{aligned} \quad (2)$$

where the first N features come from the baseline

SMT model, h_{N+1} is the logarithm of TFWIM score and h_{N+2} is the number of function words that have been inserted, λ_{N+1} and λ_{N+2} are the corresponding feature weights.

4 Experiments

4.1 Experiment settings

We conducted our experiments on Chinese-English machine translation task. The implementation of our CKY-style phrasal decoder is based on Bracketing Transduction Grammar (BTG) (Wu, 1997) with a lexicalized reordering model (Xiong et al., 2006) under ME framework. SMT decoding is performed with cube-pruning (Chiang, 2007) and the beam size is set to 20 for efficiency. The proposed TFWIM is trained by the MaxEnt toolkit (Zhang, 2006). The bilingual corpus for SMT training contains 498K sentence pairs from LDC. The training data of the lexicalized reordering model comes from LDC2003E14, which contains 128K sentence pairs. A 5-gram language model (LM) is trained over the English portion of parallel data with the Xinhua portion of LDC English Gigaword, where no target function words are deleted. The LM is integrated into the SMT decoder rather than used in a post reranking step. In addition, a CRF-based POS tagger is trained over Penn Treebank to label the target portion of bilingual data. The development data is NIST 2003 data set and the test data comes from NIST 2005 and NIST 2006 evaluation data set. The case-insensitive BLEU4 (Papineni et al., 2002) is used as the evaluation metric, where statistical significance test is performed using the bootstrap re-sampling method proposed by (Koehn, 2004).

4.2 Accuracy of TFWIM

Following (Setiawan et al., 2009), the selection of the target function words is based on their frequencies in the training corpus. Although our method can be applied to the generation of any number of distinct function words, in our experiments we mainly focus on five typical target function words contained in the set W={”the”, ”of”, ”to”, ”in”, ”for”}. They are the most frequent target function words that are unaligned (i.e., aligned to NULL) in word alignment. Table 5 shows their statistical information.

Words	**Frequency**	**# Unaligned**	**Unaligned Ratio**
of	419,070	164,027	39.1%
to	312,962	118,014	37.7%
in	268,606	97,355	36.2%
for	110,164	40,952	37.2%
the	853,752	298,228	34.9%

Table 5: Statistical information of function words.

Settings	**# Training instances**	**Accuracy**
TFWIM$_{of}$	1,164,938	97.7%
TFWIM$_{to}$	1,262,976	97.7%
TFWIM$_{in}$	1,603,213	96.7%
TFWIM$_{for}$	1,105,268	98.0%
TFWIM$_{the}$	2,218,428	89.8%
TFWIM$_{of,to}$	2,426,701	96.5%
TFWIM$_{of,to,in}$	3,552,432	93.8%
TFWIM$_{of,to,in,for}$	4,409,409	93.1%
TFWIM$_{of,to,in,for,the}$	4,143,441	90.3%

Table 6: The accuracies of TFWIMs.

For convenience, each setting of TFWIM is denoted by TFWIM$_{fw}$, where $fw \subseteq W$. Table 6 lists the number of training instances and model accuracies with ME training which is reported based on 10-fold cross validation.

According to Table 6, all TFWIMs get high accuracies, among which the model involving the word ”the” get relatively low accuracy. The reason is that the position of ”the” is very ambiguous in English sentences. Therefore, the classifier might be confused on deciding whether to insert the word ”the” given the limited context information. In general, the high accuracies of TFWIMs indicate that they are very effective to predict target function words insertion.

4.3 Comparison results with related work

We compared several SMT methods, which are introduced in the following:

Baseline: The baseline is an implementation of BTG-based phrasal SMT system. The phrase table is trained on the corpus where no target function words are deleted. The standard features (including LM) are employed and the performance is state-of-the-art.

Post-generation: We implement the post-generation approach in (Zhang et al., 2008) with the same five target function words, which is performed over the N-best (N = 10) list from baseline.

Our method (TFWIM): Different from (Zhang et al., 2008), our method works as on-the-fly generation rather than post-generation over SMT outputs. The phrase table is trained on the same corpus but target function words are deleted.

LM prediction: As language model (LM) can be used to predict target function words insertion independently, we will also compare the performance of our method (TFWIM) to LM prediction with the same phrase table and decoding procedure.

In (Menezes and Quirk, 2008)'s work, the approach is based on treelet SMT system. The function words are inserted or deleted by syntactic rules. We did not compare this work due to the difference between treelet system and our BTG-based system.

As shown in Table 7, all approaches outperform the baseline due to the extra model which is introduced to guide target function words insertion. Beyond that, our method outperforms post-generation significantly because it can lead to better reordering and promising partial hypotheses. Furthermore, TFWIM feature is indispensable because POS information provides additional benefits in deciding which target function word is more appropriate.

Settings	NIST 2005	NIST 2006
Baseline	0.3670	0.3282
Post-generation	0.3698(+0.28%)	0.3316(+0.34%)
LM prediction	0.3764(+0.94%)	0.3355(+0.73%)
Our method	**0.3798(+1.28%)**	**0.3401(+1.19%)**

Table 7: Comparison results, our method is significantly better than the baseline, the post-generation and LM prediction. ($\boldsymbol{p} < 0.05$).

4.4 Effect of different settings

We compare the results of the baseline system to that of the proposed method where different target function words are inserted during the decoding. The experimental results are shown in Table 8.

Settings	NIST 2005	NIST 2006
Baseline	0.3670	0.3282
TFWIM$_{of}$	0.3735	0.3350
TFWIM$_{to}$	0.3715	0.3343
TFWIM$_{in}$	0.3713	0.3314
TFWIM$_{for}$	0.3716	0.3309
TFWIM$_{the}$	0.3714	0.3330
TFWIM$_{of,to}$	0.3749	0.3357
TFWIM$_{of,to,in}$	0.3767	0.3374
TFWIM$_{of,to,in,for}$	0.3777	0.3358
TFWIM$_{of,to,in,for,the}$	**0.3798**	**0.3401**

Table 8: Comparison results for different TFWIMs.

According to Table 8, all settings of TFWIM outperform the baseline system, which shows the proposed method can improve the performance consistently. These benefits come from two aspects: 1) Function words deletion reduces the noises contained in the translation model. 2) Function words insertion leads to more promising hypotheses during SMT decoding. Among the settings of $|fw| = 1$, the model TFWIM$_{of}$ obtains the largest BLEU gains, which improves 0.65% and 0.68% BLEU points on NIST 2005 and NIST 2006 evaluation data sets respectively compared to the results of the baseline. The reason is that the correct insertion of function word ”of” can lead to better reordering in the translations, which improves the SMT performance.

In addition, as more and more distinct function words are involved, the performance is improved. As shown in Table 8, TFWIM$_{of,to,in,for,the}$ performs the best among all settings, which gives 1.28% and 1.19% BLEU points improvements on NIST 2005 and NIST 2006 data sets. This indicates that each kind of function word generation can bring some gains. However, such gains do not linearly increase with the contribution of each function word involved. The reason might be that the risk of false insertions is increasing as well with the number of distinct function words to be inserted.

5 Discussion

To reduce the noises caused by function words, we delete them from the target side of training data during model training. However, it is not a trivial task and the brute-force deletion can lead to *over-deletion problem* where function words in some cases should not be deleted at all. One over-deletion problem takes place in the idioms and collocations. For example, the collocation ”a piece of cake” is commonly used and the word ”of” should not be deleted. Another over-deletion problem is introduced by the correspondences between source function words and target function words. For instance, the Chinese word “为了” corresponds to the English function word ”to” (used with a verb to indicate the intention). If the word ”to” is completely deleted from the English sentences, the translation knowledge of the word “为了” will be deficient since it cannot be automatically leant from the training corpus. Besides, there also exist *over-insertion problem* where redundant or incorrect target function words are inserted. The over-insertion problem may degrade the fluency of the translation results.

Although our method faces the challenges of both over-deletion and over-insertion problems, it still gets promising results in our experiments, which indicates that it is worthwhile to pay more attention to the processing of function words in SMT systems.

6 Conclusion and Future Work

In this paper, a novel method is designed to separate the generation of target function words from target content words in SMT systems. We have demonstrated that the specific TFWIM can indeed benefit SMT performance. Experimental results illustrate that our approach brings significant improvement over the baseline.

In the future, we plan to continue our work over a larger set of target function words, as well as extending it on other SMT systems where more useful information is available to help target function words generation. In addition, we will test our method on the translation tasks over other language pairs to confirm its effectiveness.

References

Adam L. Berger, Vincent J. Della Pietra, and Stephen A. Della Pietra. 1996. *A Maximum Entropy Approach to Natural Language Processing*. *Computational Linguistics*, 22(1): pages 39-72.

Pi-Chuan Chang, Daniel Jurafsky, and Christopher D. Manning. 2009. *Disambiguating ”DE” for Chinese-English Machine Translation*. In *Proc. Fourth Workshop on SMT*, pages 215-223.

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

Ulf Hermjakob. 2009. *Improved Word Alignment with Statistics and Linguistic Heuristics*. In *Proc. EMNLP*, 33(2): pages 229-237.

Philipp Koehn. 2004. *Statistical Significance Tests for Machine Translation Evaluation*. In *Proc. EMNLP*.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. *Statistical Phrase-Based Translation*. In *Proc. HLT-NAACL*, pages 127-133.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data*. In *Proc. ICML*, pages 282-289.

Chi-Ho Li, Dongdong Zhang, Mu Li, Ming Zhou, and Hailei Zhang. 2008. *An Empirical Study in Source Word Deletion for Phrase-based Statistical Machine Translation*. In *Proc. Third Workshop on SMT*, pages 1-8.

Arul Menezes and Chris Quirk. 2008. *Syntactic Models for Structural Word Insertion and Deletion*. In *Proc. EMNLP*, pages 735-744.

Franz Josef Och. 2003. *Minimum Error Rate Training in Statistical Machine Translation*. In *Proc. ACL*, pages 160-167.

Franz Josef Och and Hermann Ney. 2004. *The Alignment Template Approach to Statistical Machine Translation*. *Computational Linguistics*, 30(4): pages 417-449.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. *Bleu: a Method for Automatic Evaluation of Machine Translation*. In *Proc. ACL*, pages 311-318.

Hendra Setiawan, Min-Yen Kan, and Haizhou Li. 2007. *Ordering Phrases with Function Words*. In *Proc. ACL*, pages 712-719.

Hendra Setiawan, Min-Yen Kan, Haizhou Li, and Philip Resnik. 2009. *Topological Ordering of Function Words in Hierarchical Phrase-based Translation*. In *Proc. ACL*, pages 324-332.

Dekai Wu. 1997. *Stochastic Inversion Transduction Grammars and Bilingual Parsing of Parallel Corpora*. *Computational Linguistics*, 23(3): pages 377-403.

Deyi Xiong, Qun Liu, and Shouxun Lin. 2006. *Maximum Entropy Based Phrase Reordering Model for Statistical Machine Translation*. In *Proc. ACL-COLING*, pages 521-528.

Dongdong Zhang, Mu Li, Nan Duan, Chi-Ho Li, and Ming Zhou. 2008. *Measure Word Generation for English-Chinese SMT Systems*. In *Proc. ACL*, pages 89-96.

Le Zhang. 2006. *Maximum entropy modeling toolkit for python and c++*. available at `http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html`.