Improving Function Word Alignment with Frequency and Syntactic Information*

Jingyi Zhang[1,2] **Hai Zhao**[1,2†]
(1) MOE-Microsoft Key Laboratory of Intelligent Computing and Intelligent System
(2) Department of Computer Science and Engineering, Shanghai Jiao Tong University
#800 Dongchuan Road, Shanghai, China, 200240
zhangjingyizz@gmail.com, zhaohai@cs.sjtu.edu.cn

Abstract

In statistical word alignment for machine translation, function words usually cause poor aligning performance because they do not have clear correspondence between different languages. This paper proposes a novel approach to improve word alignment by pruning alignments of function words from an existing alignment model with high precision and recall. Based on monolingual and bilingual frequency characteristics, a language-independent function word recognition algorithm is first proposed. Then a group of carefully defined syntactic structures combined with content word alignments are used for further function word alignment pruning. The experimental results show that the proposed approach improves both the quality of word alignment and the performance of statistical machine translation on Chinese-to-English, German-to-English and French-to-English language pairs.

1 Introduction

Word alignment is defined as identifying word-level correspondence in sentence-aligned parallel corpus. Its quality is an important factor for the performance of statistical machine translation (SMT). One of the earliest and widest used alignment approaches is based on IBM models [Brown *et al.*, 1993]. IBM model 1 is based on words' co-occurrence and the main parameter of model 1 is the translation probability of a source word f given a target word e.

Grammatically, two categories of words can be identified according to their aligning characteristics during machine translation, content words and function words. Content words are those that have a stable lexical meaning, such as noun, verb or adjective. While function words are those that have little lexical meaning, but instead indicate syntactic functions in sentences.

*This work was partially supported by the National Natural Science Foundation of China (Grant No.60903119, Grant No.61170114, and Grant No.61272248), and the National Basic Research Program of China (Grant No.2009CB320901 and Grant No.2013CB329401).

†corresponding author

Figure 1 shows two typical translation probability distributions of different target words in Chinese-to-English language pair. In fact, the size of source vocabulary is 287848. We cut out the translation probabilities of words whose frequency is lower than a proper threshold for each distribution since those translation probabilities are nearly zero, so that the characteristics of distributions will be presented clearly. The upper picture is about content word "*objectives*" while the lower one is about function word "*of*". It is clear that IBM model 1 gives better alignment precision for content word alignment since a content word usually only has evidently high translation probability with its corresponding word in an aligned sentence pair, e.g. if German word "*wasser*" and English word "*water*" occur in an aligned sentence pair then they are quite possible to be aligned. But the translation probabilities between function words in an aligned sentence pair are probably all pretty high and easily fail to determine the right function word alignment. In practice, there is a high frequency that the corresponding word of a function word is not the word which has the highest translation probability with it. For example, English word "*of*" may correspond different French words "*de*", "*en*" or "*pour*" in aligned sentence pairs. Since function words usually have quite high frequency, when they all occur in one aligned sentence pair, which French word "*of*" should be aligned to does not rely much on which one has the highest translation probability with "*of*". Sometimes a function word even does not have a corresponding word in the aligned sentence pair, e.g. "*of*" is usually omitted during English-to-Chinese translation, which causes "*of*" more easily to be incorrectly aligned. Function word alignment actually relies more on positional information according to a great deal of empirical observation.

Most existing alignment models only consider simple statistic facts for modeling. IBM Models 2-4 [Brown *et al.*, 1993] incorporate positional information and word classes trained by using a maximum-likelihood criterion. They can somehow be considered as word-based deformations of reordering model in phrase-based SMT. The Hidden Markov alignment model [Vogel *et al.*, 1996] which is usually trained along with IBM models always prefers consecutive alignments. LEAF [Fraser and Marcu, 2007] is a generative word alignment model, and it considers word dependency relationships in monolingual sentence through the concept of non-head word. All the above approaches do not directly consider

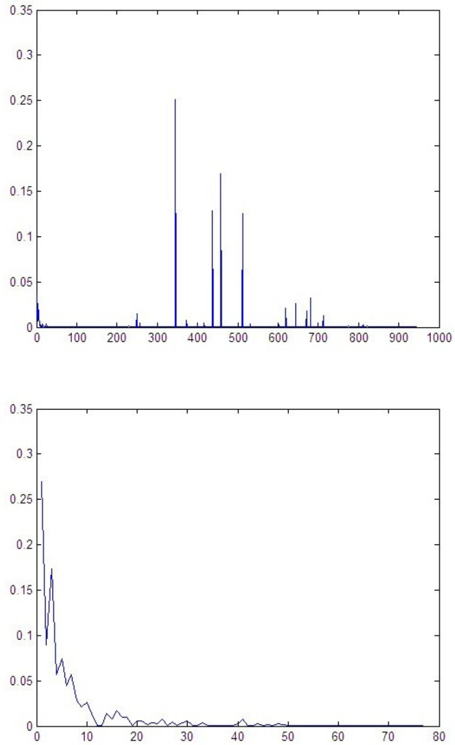

Figure 1: Translation probability distributions. The horizontal axis represents different source words ranked by decreasing frequency.

syntactic information in their models.

Syntactic parse tree is defined to characterize deep structure of language. Based on frequency and syntactic information, we propose a novel approach in this paper to prune away improbable function word alignments from an existing alignment model. We first propose a new language-independent function word recognition algorithm based on various monolingual and bilingual frequency characteristics. Then content word alignment is treated as reliable alignment while function word alignment as unreliable alignment. We define the concepts of *related word sets* of a word based on the constituency-based parse tree and then an unreliable alignment will be identified to be correct and preserved if and only if a reliable alignment exists in the corresponding *related word sets*.

A lot of work has been done to use syntactic information to improve word aligning. Some used syntactic dependency relationships to improve word alignment [Hermjakob, 2009; Ma *et al.*, 2009]. The *related word sets* of a given word w defined based on the constituency-based parse tree in our approach can be considered as sets of words that may have different syntactic relationships with w since parsing performance is not currently satisfactory for SMT and using syntactic dependency relationships will make the performance of aligning more sensitive to parsing errors. Ulf Hermjakob [2009] identified function words by a function word list. In fact, there are no clear boundaries between function words and content words, so it is troublesome and time-consuming to make a function word list. Instead our function word recognition algorithm exploits the fact that function words have no strong co-occurrence relationships in frequency and enjoys the merit of language-independent. Compared to the existing methods of identifying reliable alignments by meeting the minimal threshold restrictions [Tufis *et al.*, 2006], using intersection of bidirectional IBM models [Ma *et al.*, 2009] or additional linguistic resources such as translation lexicons [Tufis *et al.*, 2006], our approach can recognize much more reliable alignments. In addition, there is other existing work that considers general alignment pruning problem. The link deletion algorithm [Fossum *et al.*, 2008] based on various syntactic, structural, lexical and history features and the alignment refinement method [Crego and Habash, 2008] by discarding alignments that fall out of the projections of chunks did not treat content and function words differently.

2 Function Word Recognition

A function word recognition algorithm will be presented in this section. Before presenting the algorithm, we introduce necessary preprocessing over the corpus.

Given a sentence-aligned parallel corpus (F, E), a null word is added at the first position for each sentence to get (F', E') as what IBM models [Brown *et al.*, 1993] do except that null words are added for both source and target sentences not just for target sentences. Then all repeated words of each sentence in (F', E') are removed to obtain (F'', E''). Table 1 shows an example of processing a sentence.

(F, E)	the traffic stretched the patience of many thousands of people in the EU to the limit .
(F', E')	null-word the traffic stretched the patience of many thousands of people in the EU to the limit .
(F'', E'')	null-word the traffic stretched patience of many thousands people in EU to limit .

Table 1: A processing example.

(F'', E'') is used to calculate co-occurrence for each word pair $[f, e]$ as:

$$C(f, e) = \frac{N^2(f, e)}{N(f)\,N(e)}, \tag{1}$$

where $N(f)$ is number of f as source word, $N(e)$ is number of e as target word, and $N(f, e)$ is number of times when both f and e occur in an aligned sentence pair.

Figure 2 shows $C(f, e)$ distributions of English word "*of*" calculated on German-to-English corpus before (upper picture) and after (lower picture) removing repeated words. Those two distributions are pruned as the same way for Figure 1. The horizontal axis represents different German words ranked by decreasing frequency except that null word is always ranked at the first place. It can be seen that

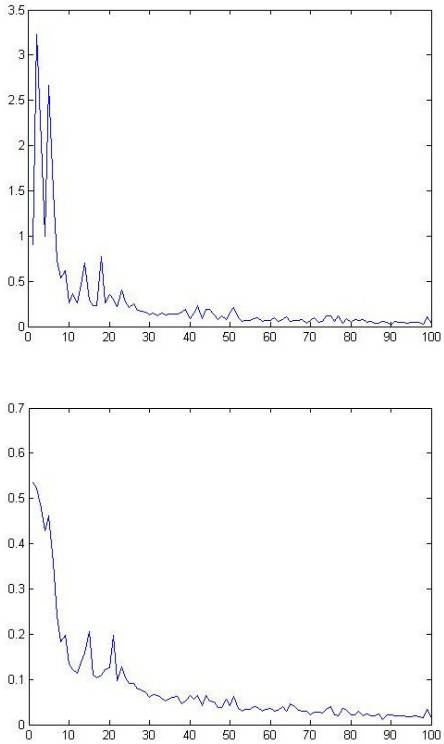

Figure 2: $C\,(f, e)$ distributions of English word "*of*".

$C\,(null - word, of)$ increases significantly and reaches the maximum value in $C\,(f, e)$ distribution of "*of*" after removing repeated words, which becomes a useful recognition characteristic for function words.

The main idea of the proposed function word recognition algorithm is if a word w is aligned to a function word according to co-occurrence then it is also a function word with null word being initialized as function word. Additionally our algorithm requires that the frequency of a recognized function word has to be higher than a predefined threshold because some content words also have a larger $C\,(f, e)$ with null word than that with its correct corresponding word due to data sparseness. For example, German word "*unterbrochene*" and English word "*adjourned*" are corresponding content words which should have high co-occurrence. But they may occur rarely together in some corpus, which causes that they have larger $C\,(f, e)$ with null word than with each other. The definition of $C\,(f, e)$ is motivated by the fact that the majority of content word pairs do not occur more than once in one aligned sentence pair.

For each sentence pair in (F', E'), function words are recognized according to Algorithm 1.

We will describe how to choose frequency threshold T later. After Algorithm 1 recognizes function words, all the remaining words will be automatically treated as content words.

Algorithm 1 The procedure of recognizing function words.

Input: An aligned sentence pair f_0^m and e_0^l from (F', E').
Output: Two arrays rf and re with lengths $m+1$ and $l+1$ respectively. $rf[j](re[i])= 1$ represents $f_j(e_i)$ is recognized as function word.

Initialize two arrays fs and es with lengths $m+1$ and $l+1$. Each element of these two arrays is a pointer to a set of integers.
Initialize $rf[0] := 1, re[0] := 1, threshold := T,$ $rf[j] := 0 (1 \le j \le m), re[i] := 0 (1 \le i \le l),$ $fs[j] := \emptyset (0 \le j \le m), es[i] := \emptyset (0 \le i \le l).$
for $i := 1$ **to** l **do**
    $i' := \underset{j \in (0 \ldots m)}{\arg\max}\, C\,(f_j, e_i)$
    add i to $fs[i']$
end for
for $j := 1$ **to** m **do**
    $j' := \underset{i \in (0 \ldots l)}{\arg\max}\, C\,(f_j, e_i)$
    add j to $es[j']$
end for
$e\_recognize(es, fs, re, rf, 0)$
$f\_recognize(fs, es, rf, re, 0)$

procedure : $e\_recognize(es, fs, re, rf, index)$
for each $element$ in $es[index]$ **do**
    if $C(f_{element}, e_0) > threshold$ **then**
        $rf[element] := 1$
        $f\_recognize(fs, es, rf, re, element)$
    end if
end for

procedure : $f\_recognize(fs, es, rf, re, index)$
for each $element$ in $fs[index]$ **do**
    if $C(f_0, e_{element}) > threshold$ **then**
        $re[element] := 1$
        $e\_recognize(es, fs, re, rf, element)$
    end if
end for

3 Unreliable Alignment Identification

For a sentence pair f_1^m and e_1^l from (F, E) with an alignment list $[f_j, e_i]$ from an existing alignment model, we propose two strategies to identify unreliable alignments as the following.

- **Strategy 1** If either of f_j and e_i is function word, then $[f_j, e_i]$ will be identified as unreliable alignment, otherwise as reliable alignment.
- **Strategy 2** If both f_j and e_i are function words, then $[f_j, e_i]$ will be identified as unreliable alignment, otherwise as reliable alignment.

We will empirically determine how to choose an appropriate strategy for a specific translation task.

4 Alignment Pruning

We first put words into three categories according to their positions in the parse tree: *left*, *middle* and *right*. For a word w in a constituency-based parse tree, we denote its nearest ancestor node whose parent node has more than one child node

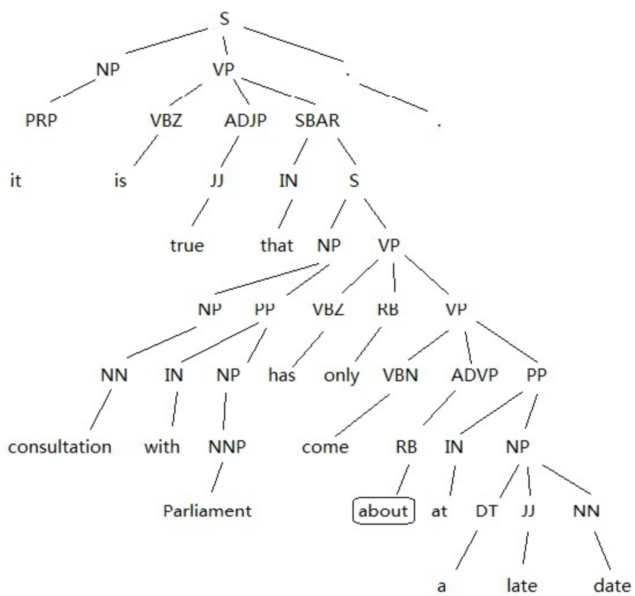

Figure 3: An example of constituency-based parse trees.

and contains at least one content word as N_{firstp} and then identify w's category according to following rules.

- If N_{firstp} is the leftmost child node of its parent node, then w will be regarded as *left* word.
- If N_{firstp} is the rightmost child node of its parent node, then w will be regarded as *right* word.
- When N_{firstp} is neither the leftmost nor the rightmost child node of its parent node, if all the words that are on the left side of w and contained in the parent node of N_{firstp} are function words, then w will be regarded as *left* word; if all the words that are on the right side of w and contained in the parent node of N_{firstp} are function words, then w will be regarded as *right* word; otherwise w will be regarded as *middle* word.

As shown in Figure 3, N_{firstp} of word "*about*" is"*ADVP*", and it is neither the leftmost nor the rightmost child node of its parent node, so "*about*" is a *middle* word.

For three types of words, we define two kinds of *related word sets* in Definition 1: *inside set* and *outside set* for *left* word, *right* word, and *middle* word, respectively. *Middle* words do not have *outside sets*.

Three examples of *related word sets* for words in Figure 3 are given as follows:
$S_i(with) = \{with, Parliament\}$
$S_o(with) = \{consultation, with\}$
$S_i(about) = \{come, about, at, a, late, date\}$

We give the concept about how to determine two *related word sets* are aligned in Definition 2.

For an unreliable alignment $[f_j, e_i]$, we check the aligning status of the corresponding *related word sets* to perform the alignment pruning according to rules below.

- If neither f_j nor e_i is *middle* word, then $[f_j, e_i]$ will be identified as correct alignment if and only if $align(S_i(f_j), S_i(e_i)) = true$ or $align(S_o(f_j), S_o(e_i)) = true.$
- Otherwise, $[f_j, e_i]$ will be identified as correct alignment if and only if $align(S_i(f_j), S_i(e_i)) = true.$

All identified as incorrect alignments will be pruned away from the existing alignment model.

Definition 1
A related word set for a word w_k in the sentence w_0^K is a set of consecutive words extracted from w_0^K:

Type of w_k	*related word set*
	inside set $S_i(w_k)$
$left$	$\{w_k, ..., w_{k_2}\}$
$right$	$\{w_{k_1}, ..., w_k\}$
$middle$	$\{w_{k_1}, ..., w_{k_2}\}$
	outside set $S_o(w_k)$
$left$	$\{w_{k_1}, ..., w_k\}$
$right$	$\{w_k, ..., w_{k_2}\}$

where w_{k_1} (w_{k_2}) is the nearest word to w_k that satisfies specific conditions:

	Conditions
w_{k_1}	w_{k_1} *is on the left side of* w_k w_{k_1} *is a* $left$ *word* $w_{k_1}^k$ *contains at least one content word*
w_{k_2}	w_{k_2} *is on the right side of* w_k w_{k_2} *is a* $right$ *word* $w_k^{k_2}$ *contains at least one content word*

Definition 2 *We say that two $related\ word\ sets$, S_1 and S_2 are aligned, i.e. $align(S_1, S_2)$=true, if and only if there is a reliable alignment $[f_j, e_i]$ ($f_j \in S_1$ and $e_i \in S_2$).*

Our approach exploits language-independent syntactic properties. And the three parse tree based categories of a word w actually carry the positional information of w in the smallest sub-tree that contains w. The size of S_i approximately reflects the size of the smallest sub-tree that contains w while the size of S_o reflects that of the second smallest one. For example, the word "*the*" in English and words after it in a sentence probably constitute a noun phrase so its *inside set* is very small. But the smallest sub-tree that contains English word "*in*" probably has more complicated structure, so the *inside set* of "*in*" is usually larger. We do not use words contained in the smallest sub-tree that contains w as *inside set* of w since only one alignment link is used to align *related word sets*, in that way the size of *inside set* might be too large and more incorrect alignment will be preserved.

5 Experiment

5.1 Experiment settings

To verify the effectiveness of the proposed approach, we perform a group of machine translation experiments on three different language pairs. For Chinese-to-English (CE) translation, we use datasets officially provided for Patent Machine Translation Task at NTCIR-9 [Goto *et al.*, 2011]. For German-to-English (GE) and French-to-English (FE) translation, we use the standard datasets provided for the Sixth Workshop on Statistical Machine Translation[1] mainly taken from version 6 of the *Europarl corpus*. Statistics for these data sets are shown in Table 2.

			SOURCE	TARGET
CE	TRAINING	SENTS	953958	
		WORDS	37176191	40417993
		VOCAB	287848	503742
	DEV	SENTS	2000	
	TEST	SENTS	2000	
DE	TRAINING	SENTS	1728211	
		WORDS	45599891	48020558
		VOCAB	375539	120667
	DEV	SENTS	2525	
	TEST	SENTS	2489	
FE	TRAINING	SENTS	1820291	
		WORDS	56193409	50598643
		VOCAB	143990	123477
	DEV	SENTS	2525	
	TEST	SENTS	2489	

Table 2: Data sets.

We train standard phrase-based SMT systems with a 5-gram language model (LM) as baseline using IRST LM Toolkit[2] and Moses [Koehn *et al.*, 2007]. GIZA++ [Och and Ney, 2003] and the *grow-diag-final-and* heuristic [Koehn *et al.*, 2003] are used to obtain symmetric word alignment model in baseline. The Stanford Parser[3] and pre-trained parsing models along with the source code are adopted to produce constituency-based parse trees for three training sets including Chinese [Levy and Manning, 2003], German [Rafferty and Manning, 2008], French [Green *et al.*, 2011] and English [Klein and Manning, 2002]. We do experiments with different combinations of thresholds and strategies for each language pair and select the one with the best BLEU score on development set. The selected thresholds and strategies are shown in Table 3.

5.2 Results

Details of alignment pruning on different language pairs are given in Table 4. The last two rows are from manual judgement of 100 sentence pairs randomly extracted from each training corpus.

[1]http://www.statmt.org/wmt11/

[2]http://hlt.fbk.eu/en/irstlm

[3]http://nlp.stanford.edu/software/lex-parser.shtml

Language pair	Threshold	Strategy
CE	0.1	Strategy 2
DE	0.01	Strategy 1
FE	0.02	Strategy 2

Table 3: Parameter settings for translation tasks on different language pairs.

Language pair	CE	DE	FE
Before pruning	38.6M	47.7M	54.9M
After pruning	37.6M	46.2M	54.5M
Reduced(%)	2.34	3.14	0.81
Precision	96.7	91.4	87.2
Recall	83.3	60.3	51.5

Table 4: Results of alignment pruning. The above block shows numbers of alignments.

In Table 4, the row right above the last row is the percentage of correctly pruned alignments out of total pruned alignments while the last row is the percentage of correctly pruned alignments out of total incorrect function word alignments. In order to give further analysis why the pruning results on FE pairs are not so good as others, we give some incorrect pruning examples below.

Madame Plooij-van Gorsel , je peux vous dire que cette question est á l' ordre du jour de la réunion des questeurs de mercredi .
Mrs Plooij-van Gorsel , I can tell you that this matter is on the agenda for the Quaestors ' meeting on Wednesday .

Table 5: A sentence pair in FE training corpus.

The correct alignment between "*peux*" and "*can*" in Table 5 is pruned away according to their parsing results: "*...(VN (CL je) (V peux))(VPinf(VN (CL vous) (V dire))...*"; "*...(NP (PRP I))(VP (MD can)(VP (VB tell)(NP (PRP you))...*". The sub-tree parsing results of these two sentences are both correct. Phrases "*je peux vous dire*" and "*I can tell you*" are very similar in language structures but receive different parsing results because of differences in treebank annotation guidelines. "*peux*" and "*can*" are contained in similar language structure and their related content words are aligned correctly, so they should have aligned *related word sets*. But due to the difference of treebank conventions, *inside set* of "*peux*" corresponds *outside set* of "*can*" while *outside set* of "*peux*" corresponds *inside set* of "*can*", which causes incorrectly pruning.

Another example is shown in Table 6. The parsing result of the French sentence is: "*...(NP (D le) (N paragraphe) (A 6)(PP (P du)(NP (N rapport) (N Cunha)))...*". It is correct according to the French parse tree annotation. But in English, similar structure will be annotated as "*...(NP (NP (D le) (N paragraphe) (A 6))(PP (P du)(NP (N rapport) (N Cunha)))...*", i.e. "*le paragraphe 6*" will be additionally annotated as noun phrase. The annotation of English treebank is more appropriate for our pruning approach since French annotation usually produce larger *inside set*, which causes more incorrect alignments preserved. In this example, the incorrect alignment between "*le*" and "*the*" is not pruned away.

le paragraphe 6 du rapport Cunha sur les programmes d' orientation pluriannuels , qui sera soumis au Parlement ce jeudi , propose d' introduire des sanctions applicables aux pays qui ne respectent pas les objectifs annuels de réduction de leur flotte .
the Cunha report on multiannual guidance programmes comes before Parliament on Thursday and contains a proposal in paragraph 6 that a form of quota penalties should be introduced for countries which fail to meet their fleet reduction targets annually .

Table 6: Another sentence pair in FE training corpus.

The difference of treebank annotation between French and English is part of the reason why the alignment pruning performance on FE is not as good as others which can be improved by choosing more proper parsers for both languages or more compatible treebanks for parsing model training.

Our approach prunes function word alignment actually according to content word alignment. Since function words do not have clear co-occurrence relationships in parallel corpus, judging a function word alignment is correct or not mainly relies on the aligning status of content words that source and target function words are related to respectively. So the performance of pruning also depends on the quality of the existing content word alignment. Of course if the function word alignment of baseline has been already pretty high, the improvement given by our approach will be insignificant since there have been few incorrect alignments that our approach can work on.

	Before		After	
Language pair	BLEU	Size	BLEU	Size
CE	32.02	71.1M	32.71	81.4M
DE	18.44	88.6M	18.76	113.3M
FE	22.98	109.2M	23.09	115.6M

Table 7: BLEU score and phrase table size of translation tasks before and after alignment pruning.

Table 7 shows BLEU scores and phrase table sizes of translation tasks before and after alignment pruning. Apparently the performance of phrase-based SMT system improves more significantly as more incorrect function word alignments are pruned. And after removing more incorrect function word alignments, translation system can extract more and longer translation phrases since incorrect alignments might cross boundaries of correct translation phrases. For example, phrase pair "*The United States of America*" and "*Die Vereinigten Staaten von Amerika*" will not be correctly extracted if "*of*" is aligned to some other improper word. In fact, most function word pairs linked by incorrect function word alignments are correct translation lexicon items such as French-to-English word pair $[de, of]$, and they just do not correspond in the aligned sentence pair. If we already have strong reordering and language models, the influence of incorrect function word alignments will be ignorable. But we train reordering model on word-aligned corpus, so poor function word alignment precision will also decrease the quality of reordering model.

6 Conclusion

Since function words do not have clear correspondence between source and target languages and this characteristic makes function words easily aligned incorrectly, usually content words are aligned better than function words. In this paper, we propose a syntax motivated approach to prune function word alignments according to content word alignments from an existing alignment model, i.e. our approach judges a function word alignment is correct following the condition that the related content words are aligned. Our approach exploits various bilingual and monolingual frequency characteristics extracted from parallel corpus to recognize function words and then prunes function word alignments according to the aligning status of corresponding *related word sets* derived from parse tree. Our approach improves word alignment precision and statistical machine translation performance on different language pairs.

References

[Brown *et al.*, 1993] P.F. Brown, V.J.D. Pietra, S.A.D. Pietra, and R.L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311, 1993.

[Crego and Habash, 2008] J.M. Crego and N. Habash. Using shallow syntax information to improve word alignment and reordering for smt. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 53–61. Association for Computational Linguistics, 2008.

[Fossum *et al.*, 2008] V. Fossum, K. Knight, and S. Abney. Using syntax to improve word alignment precision for syntax-based machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 44–52. Association for Computational Linguistics, 2008.

[Fraser and Marcu, 2007] A. Fraser and D. Marcu. Getting the structure right for word alignment: Leaf. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 51–60, 2007.

[Goto *et al.*, 2011] I. Goto, B. Lu, K.P. Chow, E. Sumita, and B.K. Tsou. Overview of the patent machine translation task at the ntcir-9 workshop. In *Proceedings of NTCIR*, volume 9, pages 559–578, 2011.

[Green *et al.*, 2011] S. Green, M.C. de Marneffe, J. Bauer, and C.D. Manning. Multiword expression identification with tree substitution grammars: A parsing tour de force with French. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 725–735, 2011.

[Hermjakob, 2009] U. Hermjakob. Improved word alignment with statistics and linguistic heuristics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 229–237. Association for Computational Linguistics, 2009.

[Klein and Manning, 2002] D. Klein and C.D. Manning. Fast exact inference with a factored model for natural language parsing. *Advances in neural information processing systems*, 15(2003):3–10, 2002.

[Koehn *et al.*, 2003] P. Koehn, F.J. Och, and D. Marcu. 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-Volume 1*, pages 48–54. Association for Computational Linguistics, 2003.

[Koehn *et al.*, 2007] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, et al. Moses: Open source toolkit for statistical machine translation. In *Annual meeting-association for computational linguistics*, volume 45, page 2, 2007.

[Levy and Manning, 2003] R. Levy and C. Manning. Is it harder to parse Chinese, or the Chinese treebank. In *Proceedings of ACL*, volume 3, pages 439–446, 2003.

[Ma *et al.*, 2009] Y. Ma, P. Lambert, and A. Way. Tuning syntactically enhanced word alignment for statistical machine translation. 2009.

[Och and Ney, 2003] F.J. Och and H. Ney. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51, 2003.

[Papineni *et al.*, 2002] K. Papineni, S. Roukos, T. Ward, and W.J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

[Rafferty and Manning, 2008] A.N. Rafferty and C.D. Manning. Parsing three German treebanks: Lexicalized and unlexicalized baselines. In *Proceedings of the Workshop on Parsing German*, pages 40–46. Association for Computational Linguistics, 2008.

[Tufis *et al.*, 2006] D. Tufis, R. Ion, A. Ceausu, and D. Stefanescu. Improved lexical alignment by combining multiple reified alignments. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, 2006.

[Vogel *et al.*, 1996] S. Vogel, H. Ney, and C. Tillmann. Hmm-based word alignment in statistical translation. In *Proceedings of the 16th conference on Computational linguistics-Volume 2*, pages 836–841. Association for Computational Linguistics, 1996.