To Filter Discontinuous Word Alignment for Statistical Machine Translation

Chenchen Ding and Mikio Yamamoto
Department of Computer Science
University of Tsukuba
1-1-1 Tennodai, Tsukuba, 305-8573, Japan
tei@mibel.cs.tsukuba.ac.jp, myama@cs.tsukuba.ac.jp

***Abstract*—We propose a language-independent approach to clean up word alignment errors in an aligned parallel corpus, which are caused by the unsupervised word-align process. In such an aligned corpus, we evaluate the alignment patterns of one-to-many discontinuous words by statistical measures of collocation. The alignment of discontinuous words without strong collocation tendencies will be taken as errors and deleted. We conduct experiments on two-directional Japanese-English and German-English translation tasks. The experiment results show the state-of-the-art word alignment filtered by the proposed approach can lead to a better translation performance.**

Keywords—statistical machine translation; discontinuous word alignment

I. Introduction

To align words of sentence pairs in a parallel corpus is the foundation of almost all the statistic machine translation (SMT) systems nowadays. The word alignment task is first motivated by the word-based statistic translation approaches [1]. Afterwards, in the phrase-based SMT system [2] and the hierarchical phrase-based SMT system (HIERO) [3], word alignment is utilized in translation template (rule) extraction to construct their translation models. The word alignment is also utilized in syntax-driven template extraction approaches [4], [5].

As a parallel corpus used for training an SMT system usually contains millions of sentence pairs, where to align words for all the sentence pairs manually is impossible, a word aligner is needed. Presently, the most widely used word aligner is **GIZA++**[1] [6], which can generate one-to-many word alignment on a parallel corpus in an unsupervised way. In practical application, **GIZA++** is used to generate two-directional one-to-many word alignments of the source and target languages separately; and then symmetrization heuristic rules [2] are applied to combine them to generate a symmetrized many-to-many word alignment.

Because **GIZA++** is word-based and unaware of sentence structures, it often makes global and syntactic alignment errors even with the symmetrization heuristic rules, which are also local. On the other hand, there have been word-align approaches taking the sentence structure into consideration. The earliest attempt is the stochastic inversion transduction grammar (ITG) model, proposed by [7]. Based on the ITG formulation, some structure-sensitive word-align approaches have been proposed [8], [9]. However, the ITG formulation is not general enough to represent many corresponding structural patterns [7], [10], [11]. More linguistically-motivated, [12] introduces syntactic parsing tree into word-align process.

Despite global and syntactic errors, the word-align approach of **GIZA++** with the symmetrization heuristics is without the loss of generality as ITG-based approaches; and it needs no extra parser. Based on this, we propose an approach to handle the global and syntactic alignment errors in the result of **GIZA++** with the symmetrization heuristics. Specifically, we focus on the alignment patterns of one-to-many discontinuous words, and delete all of those without strong collocation tendencies, according to statistical measures. So, our approach is simple and needs no linguistic information. We conduct experiments on two-directional translations of Japanese-English and German-English. Because both Japanese and German have different sentence structures of English, **GIZA++** tends to make mistakes in word alignment in these language pairs. The experiment results show our approach can improve the translation performances.

II. Related Work

Generally, word-based and ITG-based approaches are two main genres of the word alignment task.

GIZA++ with symmetrization heuristics [2] is a typical and widely-used word-based approach. Due to the unawareness of sentence structure, in this kind of approach, syntactic errors are often made. In Fig. 1, we show a word alignment example of Japanese and English. We can see the structure-sensitive function words are easily mis-aligned, i.e. the genitive case-marker "の" and the topic-marker "は" of Japanese are wrongly aligned to the English article *the*, which will prevent the translation rule extraction.

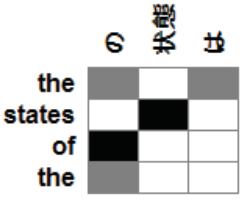

Fig. 1. Word alignment of Japanese and English, generated by **GIZA++** with symmetrization heuristics. The colored boxes stand for the aligned words but the gray ones are errors.

[1] http://code.google.com/p/giza-pp/

On the other hand, word-based approaches can deal with any kinds of complex structural correspondences, some of which may go beyond the ability of ITG based-approaches [8], [9]. The most frequently discussed problem is the *inside-outside pattern*, which is not an uncommon phenomenon but cannot be deduced by the ITG formulation [7], [10], [11]. In Fig, 2 and Fig. 3, we show the examples of the inside-outside pattern (composed of the four double-lined boxes) in German-English and in Japanese-English word alignment.

As to our approach, we take the word alignment generated by **GIZA++** with symmetrization heuristics as the baseline, and further detect and clean up the alignment errors in it. For the example shown in Fig. 1, our approach will determine the one-to-many discontinuous word alignments: "*the* ↔ の ... は" and "の ↔ *the* ... *of the*" are whether real linguistic phenomena or just occasional errors, and then clean up them (take the involved words unaligned) if they are errors.

Many attempts are made to improve the quality of word-based approaches. There are methods utilizing linguistically-oriented heuristics. A typical one proposed by [13] takes the advantage of function words. There are also correction approaches as the same idea of our approach, such as [14], which, however, still needs selected features and a heuristic phrase dictionary for their unsupervised adaption. Compared to these previous approaches, the proposed approach is without any linguistic heuristics or dictionaries, while it only introduces a light weight statistical interface as a post-process for the rigid symmetrization heuristics. The proposed approach thus offers a complementary process to handle those nonlocal, discontinuous aligned words, which affect the quality of extracted translation models.

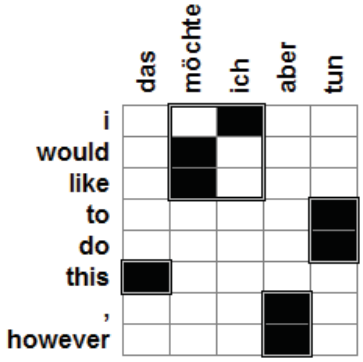

Fig. 2. An inside-outside pattern in German-English word alignment.

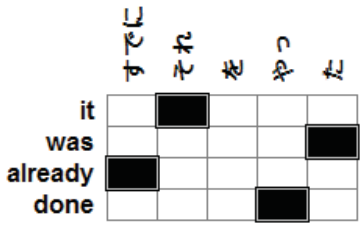

Fig. 3. An inside-outside pattern in Japanese-English word alignment.

III. Proposed Approach

We focus on the alignment pattern of one-to-many discontinuous words because it is often related to sentence structures or particular expression ways in different languages, and thus tends to be mis-aligned. Specifically, in the word-based baseline word alignment of a language pair A and B[2], we consider the following pattern[3]

$$a \leftrightarrow \beta_1 ... \beta_2 ... \cdots ... \beta_k ... \cdots ... \beta_n \tag{1}$$

where a is a word in language A; β_k is a continuous sequence of one or more words in language B; and the "$\leftrightarrow$" means the left and right sides are aligned by the baseline output. We use "..." to denote the discontinuous part, i.e., for all the k, between the word sequence β_k and β_{k+1}, there is at least one word not aligned to a.

Then, we treat the problem to be the identification of the collocation of all word sequence β_k on the right-hand given the left-hand word a. We assume each continuous word alignment $a \leftrightarrow \beta_k$ is an independent event and determine whether the event of Exp. (1) is occasional or not.

To the task of word collocation detection, there have been many proposed and discussed measures. [15] summarizes 84 different measures and [16] gives a systematic comparison. Typically, there are measures like mutual information (MI), t-score, log-likelihood ratio, etc. We use MI in our approach for it is the basic measure of the mutual dependence of random events.

The calculation of MI can be represented as $\log \frac{Obs.}{Exp.}$, where the $Obs.$ is the *observed frequency* of an event and the $Exp.$ is the *expected frequency* of the event. Specifically, for the event of Exp. (1), we can calculate the $Obs.$ and $Exp.$ as following

$$Obs. = Count(a \leftrightarrow \beta_1 ... \beta_2 ... \ ... \beta_n) \tag{2}$$

$$Exp. = Count(a) \cdot \prod_{k=1}^{n} P(a \leftrightarrow \beta_k) \tag{3}$$

where $Count(\cdot)$ is the occurrence times of the particular word or aligned pattern in the parallel corpus with the baseline alignment. The $P(a \leftrightarrow \beta_k)$ in Exp. (3) can be calculated as Exp. (4).

$$P(a \leftrightarrow \beta_k) = \frac{Count(a \leftrightarrow \beta_k)}{Count(a)} \tag{4}$$

The value of MI can take any real number. A large positive MI stands for a strong collocation tendency; and a negative MI with large absolute value shows a strong occasionality. MI becomes 0 if $Obs. = Exp.$, where no tendency is shown. We set 0 as the threshold. If an aligned pattern in Exp. (1) is with an MI greater than 0, we remain it; and for the else we take them as errors and delete all word alignments in them, i.e., to make the word a and all words in every β_k unaligned[4].

[2]In this section, we do not explicitly distinguish source language and target language because they are symmetric.

[3]This pattern will not appear in an ITG-based approach.

[4]However, it can be considered that the word a should be actually aligned to not all, but one or some certain β_k. As we cannot determine the specific β_k, we delete the alignment of them all. This may result in a word alignment which is "too loose". We will discuss this in Sec. V.

Because we use 0 as the threshold, the decision depends only on the magnitudes of $Obs.$ and $Exp.$ essentially. This makes the result of using MI identical to using t-score[5], or any other statistical measures based on the magnitudes of observation and expectation frequencies.

IV. Experiment

As Japanese has a different word order of English; and German has different syntax structures of English and has more complex inflections, the errors in word alignment of these language pairs usually appear and become an issue. So, we conduct experiments of our approach on four translation tasks: German-to-English (de–en), English-to-German (en–de), Japanese-to-English (ja–en) and English-to-Japanese (en–ja).

For the corpora, we use the **NTCIR-7** patent corpus [17] of 1.8 million Japanese and English sentence pairs, and the **Europarl**[6] [18] corpus of 1.9 million German and English sentence pairs for model training. The development set and test set in ja–en and en–ja translations are the correspondent sets of **NTCIR-7** evaluation, which contain 915 and $1,381$ sentence pairs respectively. For de–en and en–de translations, we use the official test set of ACL WMT 2007[7] [19] as the development set and the test set on news domain of ACL WMT 2011[7] as the test set, which contain $2,000$ and $3,003$ sentence pairs respectively.

For the models, we train a phrase-based system for each translation task using the state-of-the-art **MOSES** system[8] [20]. For all the translation tasks, the *max-phrase-length* is set to 5 in model extraction and the reordering model is trained with the *msd-bidirectional-fe* option. We use **SRILM**[9] [21] to train a 5-gram interpolated modified Kneser-Ney language model for each language on the single language part of the corresponding parallel corpus[10].

We used the **GIZA++** and the *grow-diag-final-and* heuristics realized by **MOSES** to get the baseline word alignment (no filt.). As to the proposed approach, we conduct filtering both on source-to-target direction (s–t filt.) and target-to-source direction (t–s filt.), and the intersection of the two direction results (all filt.). Translation model and reordering model of each translation task are built on the four kinds of word alignments separately.

In decoding, we set the *ttable-limit* to 10, and stack size to 100 for all the translation tasks. The *distortion-limit* is set to 12 for ja–en and en–ja, and 6 for de–en and en–de translations.

In evaluation, we tune the feature weights by MERT [22] on the development set and use the tuned weights with the same decoding setting to evaluate the test set BLEU [23]. We also conduct significance test of different models by the bootstrap sampling method [24]. The evaluation and significant test are conducted by **bleukit**[11]. We show the experiment results in Table I.

[5] $\frac{Obs.-Exp.}{\sqrt{Obs.}}$

[6] http://www.statmt.org/europarl/

[7] http://matrix.statmt.org/test_sets/list

[8] http://www.statmt.org/moses/

[9] http://www.speech.sri.com/projects/srilm/

[10] In ja–en translation, the English language model is only trained on the English part of the **NTCIR-7** corpus and in de–en translation, only trained on the **Europarl** corpus.

V. Discussion

In Table I, we see that the performance evaluated by BLEU is improved on all the four translation tasks by our filtering approach. We show the size (number of rules) of translation models in Table II. It can be observed, with the two-directional filtering (all filt.), more rules are extracted than baseline (no filt.); and one-directional filtering (en–x filt., x–en filt.) has a medium size between the two-directional filtering and baseline.

However, a model with more rules does not always lead to a better translation performance. On the translation between English and German, we can see that the English-to-German direction filtering works as a matter of fact (t–s filt. of de–en, and s–t filt. of en–de in Table. I). The reason can be explained by the examples in Table III, where we show the most frequent deleted word alignments. On the English-to-German direction filtering (en $\leftrightarrow$ discont.de in Table III), our approach does delete incorrect alignments. But on the German-to-English direction filtering (de $\leftrightarrow$ discont.en in Table III), correct alignments are also deleted, e.g., around the genitive articles *der* and *des* in German. This is because German tends to use inflection but English tends to use periphrases in their expressions. On the translation between English and Japanese, we can see that our approach performs well when the target side is English rather than Japanese. We think this is because Japanese is a language with rich function words but English is not. In Table III, we see that the deleted alignments are mainly around the Japanese function words such as "の", which are hard to control and affect the target side Japanese translation quality largely after filtering.

From Table III, we can see correct word alignments may be deleted excessively, which will result in a too *loose* (fewer aligned words) word alignment. So, we can conclude the properties of our approach from the experiment results. It works well when applied from a poor-inflection language to a

TABLE I. The test set BLEU of each translation task with different filtered word alignments. Compared with the baseline (no filt.), ‡ means the significance of improvement is at $p < .01$ level, and † means at $p < .05$ level.

align.	de–en	en–de	ja–en	en–ja
no filt.	16.66	11.44	28.39	30.15
all filt.	$17.03^{\ddagger}$	$11.92^{\ddagger}$	$28.80^{\dagger}$	**30.42**
s–t filt.	16.67	$\mathbf{11.99}^{\ddagger}$	$\mathbf{28.84}^{\dagger}$	30.32
t–s filt.	$\mathbf{17.06}^{\ddagger}$	11.55	$28.81^{\dagger}$	29.29

TABLE II. Translation model sizes in million (M) rules. One side of the models is English; the other side (x) is German (de) or Japanese (ja). For the *max-phrase-length* is for both source and target languages, swapping the source and target languages will result in models with the same number of rules.

x	no filt.	en–x filt.	x–en filt.	all filt.
de	$49M$	$53M$	$54M$	$58M$
ja	$52M$	$61M$	$62M$	$72M$

[11] http://www.nlp.mibel.cs.tsukuba.ac.jp/bleu_kit/

TABLE III. THE MOST FREQUENT DELETED ONE-TO-MANY DISCONTINUOUS WORD ALIGNMENTS. TOP-5 FOR EACH LANGUAGE PAIR IN CORRESPONDING ALIGNED PARALLEL CORPUS.

en	↔	discont.de	de	↔	discont.en	en	↔	discont.ja	ja	↔	discont.en
to	↔	*zu ... ,*	*der*	↔	*the ... 's*	*of*	↔	の ... を	の	↔	*the ... of*
the	↔	*die ... der*	*,*	↔	*, ... ,*	*of*	↔	の ... の	、	↔	*, ... of*
to	↔	*möchte ... ,*	*der*	↔	*the ... of*	*the*	↔	、 ... の	の	↔	*the ... of the*
to	↔	*zu ... ,*	*zu*	↔	*to ... to*	*to*	↔	に ... さ	、	↔	*the ... the*
to	↔	*der ... der*	*des*	↔	*the ... of*	*to*	↔	を ... に	の	↔	*the ... the*

TABLE IV. AN EXAMPLE OF GERMAN-TO-ENGLISH TRANSLATION.

input	*“ ich habe auch meine kollegen gefragt , aber offensichtlich gibt es so etwas bei* **uns wirklich nicht** *. ”*
reference	*“ i even consulted my colleagues , but it seems that* **we really do not** *have anything like that . ”*
no filt. output	*‘ i have also asked my colleagues , but it is clear that there is something* **we really** *. ’*
t-s filt. output	*‘ i have also asked my colleagues , but it is clear that there is something* **we really not** *’ .*

TABLE V. AN EXAMPLE OF JAPANESE-TO-ENGLISH TRANSLATION.

input	また 、 ＡＤＰＣＭ アナライザ １２ は **中間 データ 等** を 記憶 する メモリ を 内蔵 して いる 。
reference	*moreover , a memory 12a used to store* **additional data or the like** *is embedded in the adpcm analyzer 12 .*
no filt. output	*further ,* **such as** *adpcm analyzer 12 is a memory for storing* **the intermediate data** *built therein .*
s-t filt. output	*further , the adpcm analyzer 12 is a memory for storing* **the intermediate data and the like** *.*

rich-inflection language; as well as applied in the translation task from a language with rich function words to a language with poor ones. We show examples of German-to-English and Japanese-to-English translation in Table IV and Table V respectively. In the German-to-English example, we can see the baseline translation (no filt. output) loses the negation of the input sentence. This is due to that the negative adverb “*nicht*” in German tends to appear at the end of a sentence and thus it is easily dropped in translation. By the proposed approach (t-s filt. output), the “*nicht*” is correctly translated to the English “*not*”. (Actually, the “*uns wirklich nicht*” is translated to “*we really not*” word-by-word.) In the Japanese-to-English example, we can see the noun phrase “中間 データ 等” in the input Japanese sentence is separated in the baseline translation (no filt. output) and the “等” is translated to “*such as*”. This is not a wrong translation for the very word though, the whole output results in a wrong expression. We have a better translation by the proposed approach (s-t filt. output). From the examples, we find the proposed approach does improve the translation performance concerning sentence structures.

VI. CONCLUSION AND FUTURE WORK

In this paper, we have proposed a simple approach to filter the word alignment errors in a word-based word-align approach. The experiment results show our approach can improve the performance of a phrase-based translation system.

There are also other systems to extract more specific rules disregarding the quality of word alignment, which usually leads to too huge a translation model to be tractable. The hierarchical phrase-based system (HIERO) is a typical one, whose model size is discussed by [25]. More extremely, the non-hierarchical approach proposed by [26] extracts all the discontinuous translation patterns. As a result, it must utilize an on-the-fly technique due to the extremely huge model size. In future work, we will conduct experiments of the proposed approach on these SMT systems. We are also ready to take more complex word alignment patterns into consideration and develop more sophisticated measures to improve the filtering process.

REFERENCES

[1] 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*, vol. 19, no. 2, pp. 263–311, 1993.

[2] P. Koehn, F. J. Och, and D. Marcu, “Statistical phrase-based translation,” in *Proceedings of NAACL-HLT 2003*, vol. 1, 2003, pp. 48–54.

[3] D. Chiang, “Hierarchical phrase-based translation,” *Computational Linguistics*, vol. 33, no. 2, pp. 201–228, 2007.

[4] C. Quirk, A. Menezes, and C. Cherry, “Dependency treelet translation: syntactically informed phrasal smt,” in *Proceedings of ACL 2005*, 2005, pp. 271–279.

[5] Y. Liu, Q. Liu, and S. Lin, “Tree-to-string alignment template for statistical machine translation,” in *Proceedings of COLING-ACL 2006*, 2006, pp. 609–616.

[6] F. J. Och and H. Ney, “A systematic comparison of various statistical alignment models,” *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, 2003.

[7] D. Wu, “Stochastic inversion transduction grammars and bilingual parsing of parallel corpora,” *Computational Linguistics*, vol. 23, no. 3, pp. 377–403, 1997.

[8] P. Blunsom, T. Cohn, C. Dyer, and M. Osborne, “A Gibbs sampler for phrasal synchronous grammar induction,” in *Proceedings of ACL-IJCNLP 2009*, vol. 2, 2009, pp. 782–790.

[9] G. Neubig, T. Watanabe, E. Sumita, S. Mori, and T. Kawahara, “An unsupervised model for joint phrase alignment and extraction,” in *Proceedings of ACL-HLT 2011*, vol. 1, 2011, pp. 632–641.

[10] B. Wellington, S. Waxmonsky, and I. D. Melamed, “Empirical lower bounds on the complexity of translational equivalence,” in *Proceedings of COLING-ACL 2006*, 2006, pp. 977–984.

[11] A. Søgaard and D. Wu, “Empirical lower bounds on translation unit error rate for the full class of inversion transduction grammars,” in *Proceedings of IWPT 2009*, 2009, pp. 33–36.

[12] T. Nakazawa and S. Kurohashi, “Statistical phrase alignment model using dependency relation probability,” in *Proceedings of SSST 2009*, 2009, pp. 10–18.

[13] H. Setiawan, C. Dyer, and P. Resnik, "Discriminative word alignment with a function word reordering model," in *Proceedings of EMNLP 2010*, 2010, pp. 534–544.

[14] J. S. McCarley, A. Ittycheriah, S. Roukos, B. Xiang, and J.-m. Xu, "A correction model for word alignments," in *Proceedings of EMNLP 2011*, 2011, pp. 889–898.

[15] P. Pecina, "An extensive empirical study of collocation extraction methods," in *Proceedings of ACL 2005, Student Research Workshop*, 2005, pp. 13–18.

[16] S. Evert, "Corpora and collocations," *Corpus Linguistics. An International Handbook*, vol. 2, 2008.

[17] A. Fujii, M. Utiyama, M. Yamamoto, and T. Utsuro, "Overview of the patent translation task at the ntcir-7 workshop," in *Proceedings of NTCIR-7*, 2008, pp. 389–400.

[18] P. Koehn, "Europarl: a parallel corpus for statistical machine translation," in *Proceedings of MT summit 2005*, vol. 5, 2005.

[19] C. Callison-Burch, C. Fordyce, P. Koehn, C. Monz, and J. Schroeder, "(Meta-) evaluation of machine translation," in *Proceedings of ACL-WMT 2007*, 2007, pp. 136–158.

[20] 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 *Proceedings of ACL 2007*, 2007, pp. 177–180.

[21] A. Stolcke *et al.*, "SRILM–an extensible language modeling toolkit," in *Proceedings of ICSLP 2002*, vol. 2, 2002, pp. 901–904.

[22] F. J. Och, "Minimum error rate training in statistical machine translation," in *Proceedings of ACL 2003*, vol. 1, 2003, pp. 160–167.

[23] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of ACL 2002*, 2002, pp. 311–318.

[24] P. Koehn, "Statistical significance tests for machine translation evaluation," in *Proceedings of EMNLP 2004*, vol. 4, 2004, pp. 388–395.

[25] C. Ding, T. Inui, and M. Yamamoto, "Long-distance hierarchical structure transformation rules utilizing function words," in *Proceedings of IWSLT 2011*, 2011, pp. 159–166.

[26] M. Galley and C. D. Manning, "Accurate non-hierarchical phrase-based translation," in *Proceedings of NAACL-HLT 2010*, 2010, pp. 966–974.