Faster Development of Statistical Machine Translation Systems with Sampling-based Alignment and Hierarchical Sub-sentential Alignment

Baosong Yang, Hao Wang and Yves Lepage
Graduate School of IPS, Waseda University
{yangbaosong@fuji., oko_ips@ruri., yves.lepage@}waseda.jp

Abstract

Bilingual hierarchical sub-sentential alignment used in conjunction with sampling-based multilingual alignment has been shown to achieve high quality of sub-sentential alignment, despite its simplicity. We accelerated this method. In this paper, we evaluate the resulting alignments on several machine translation tasks. We demonstrate that this method leads to state-of-the-art results. The training phrase is much faster than the standard baseline (GIZA++/Moses). Also, decoding times are reduced because smaller phrase tables are obtained.

1 Introduction

Phrase-based Statistical Machine Translation (PB-SMT), which is built upon the word alignment output, has matured greatly over the past 10 years and is the most prominent approach nowadays. Still, the construction of high-performance translation systems requires a significant processing time in spite of using optimized computing resources and parallel programming techniques. As one of the most time-consuming processing step, word alignment affects the quality of translation systems. A more effective word alignment leads to decrease in time cost because decoding with a smaller phrase table produced by the training process, is faster.

Sub-sentential alignment consists in identifying translation units from a sentence-aligned parallel corpus. Numerous methods have been proposed to compute sub-sentential alignments from parallel sentences. These methods are divided into two main categories, the *probabilistic* and the *associative*[1] approaches.

The *probabilistic* approach attempts at determining the best set of alignment links between source and target words or phrases in parallel sentences. IBM models [2] are the most widely used representatives in this category, due to their tight integration within the PB-SMT framework (e.g. GIZA++[12] for Moses Toolkit[2]). A global optimization process simultaneously considers all possible associations in the entire corpus and estimates the parameters of the parallel corpus.

However, this global process is time-consuming especially when working on large parallel corpora. It results in all words or phrases in the parallel corpus having to be aligned synchronously. Unfortunately, not all aligned parallel sentence pairs are used to translate a given input text. In addition, to supplement new parallel sentences into an original corpus is a squander of resource when the system runs the alignment process for the whole corpus while only a few of new sentences are appended.

The *associative* approaches, introduced in [4], do not rely on an alignment model, but on independence statistical measures. The Dice coefficient, mutual information [4], and likelihood ratio [3] are representative cases of this approach. The *associative* approaches use a local maximization process in which each sentence is processed independently.

Anymalign, introduced in [11], is an associative sub-sentential alignment method which has been shown to obtain better results than state-of-the-art methods on bilingual lexicon induction tasks if the evaluation is performed by comparing word associations with reference dictionaries. This method samples large numbers of sub-corpora randomly to obtain source and target word occurrence distributions. The more often two words have the same occurrence distribution over particular sub-corpora, the better the association score between them. However, phrase tables directly produced by Anymalign fail to perform on par with state-of-the-art methods.

Cutnalign, introduced in [10], is a bilingual hierarchical sub-sentential alignment method. The method obtains better sub-sentential alignments at the sentence level, due to a recursive binary segmentation of the alignment matrix used to process the parallel sentences [10]. The improvement is due to its simplicity. It yields a comparable performance in comparison with other methods because processing large numbers of randomly sampled sub-corpora with Anymalign delivers accurate association measures and a good coverage for the whole corpus. In this work, we extend the work in [10], especially in decreasing time costs by bilingual hierarchical sub-sentential alignment so that the resulting alignments can

[1] Also called "heuristic" in [12].
[2] http://www.statmt.org

be used to obtain state-of-the-art results with even only half of the processing time required by the usual baseline (GIZA++/Moses). The rest of this paper is organized as follows: Section 2 describes this associative alignment approach and our improvement. Section 3 presents an evaluation on several, complementary machine translation experiments, and some analysis of the results. Section 4 concludes this work and discusses future research directions.

2 Hierarchical sub-sentential alignment

[10] proposed a sub-sentential alignment algorithm based on a recursive binary segmentation process of the alignment matrix between a source sentence and its translation. This method mainly has three steps. Firstly, the strength of the translation link between any source and target pair of words (s,t) is computed as the product of the two translation probabilities $p(s|t)$ and $p(t|s)$.

$$w(s,t) = p(s|t) \times p(t|s) \quad (1)$$

The work of [15] on document clustering inspired the method for the segmentation of sentences into subparts. [10] adapted it to the search of the best alignment between words of a source sentence and those of a target sentence by recursively segmenting and aligning subparts of sentences. Figure 1 shows an example of the alignment result using the method.

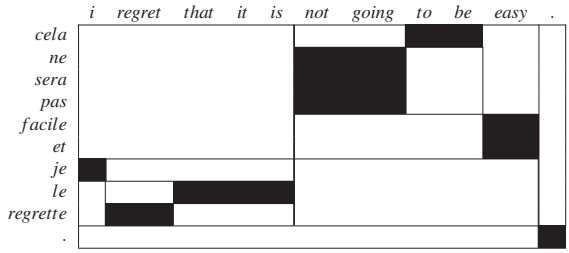

Figure 1: An example of alignment result using this method. A black block represents a word alignment.

In comparison with the implementation used in [10], based on an analysis of all particular cases and their separate implementation, our new implementation is 50 times faster than the one used in [10]. In order to visualize and compare sub-sentential alignments output by MGIZA++ and Cutnalign, we also implemented a visualization tool of alignment matrices. Figure 2 and Figure 3 show examples that illustrate different alignments obtained on the same sentence pair using this method, and the commonly used method implemented in Moses.

3 Experiments

3.1 Experiment settings

Here, we basically reproduce experiments in [10] with similar results, but show large reductions in overall processing time. Our alignment method is evaluated within a PB-SMT system built by using the Moses toolkit [9], the Ken Language Modeling toolkit [6] and a lexicalized reordering model [8]. We built systems for three languages pairs involving 5 European languages[3]: fr-en, fi-en (agglutinating language-isolating language), and es-pt (very close languages). For each language pair, the training corpus is made of 347,614 sentences from the Europarl parallel corpus v3 [7] (10 million word tokens in English). The tuning set contains 500 sentences, and 38,123 sentences were used for testing. Each group of experiments was run on machines with the same configuration. Translations were evaluated using BLEU, WER, and NIST. We compared two settings: MGIZA++ [5] and Anymalign-4 + Cutnalign, as described below.

MGIZA++ MGIZA++ implements the IBM models and the HMM [14]. It is run with default settings: 5 iterations of IBM1, HMM, IBM3, and IBM4 in two directions of source to target and target to source. The alignments are produced by MGIZA++ and a phrase table is extracted from the alignments using the grow-diag-final-and heuristic [1] integrated in the Moses toolkit. This training process is described in Figure 4.

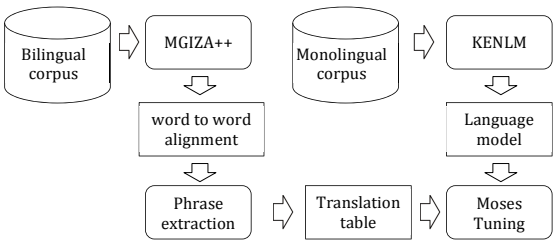

Figure 4: Main training modules and models of MGIZA++ with Moses Toolkits.

Anymalign-4 + Cutnalign Anymalign has been designed to directly build phrase tables. As it can be stopped at any time, and in order to reduce the total time of alignment, its running time is set so that it runs for half the time used by MGIZA++. In addition, we set the length of output phrases to 4 (hence our notation Anymalign-4). Cutnalign implements the algorithm described in Section 2. We passed the alignments output by Cutnalign, whose input is the phrase table produced by Anymalign-4, to the grow-dial-final-and heuristic of the Moses Toolkit to build phrase tables. Figure 5 describes the training process in this setting.

[3] English (en), French (fr), Spanish (es), Portuguese (pt), Finnish (fi).

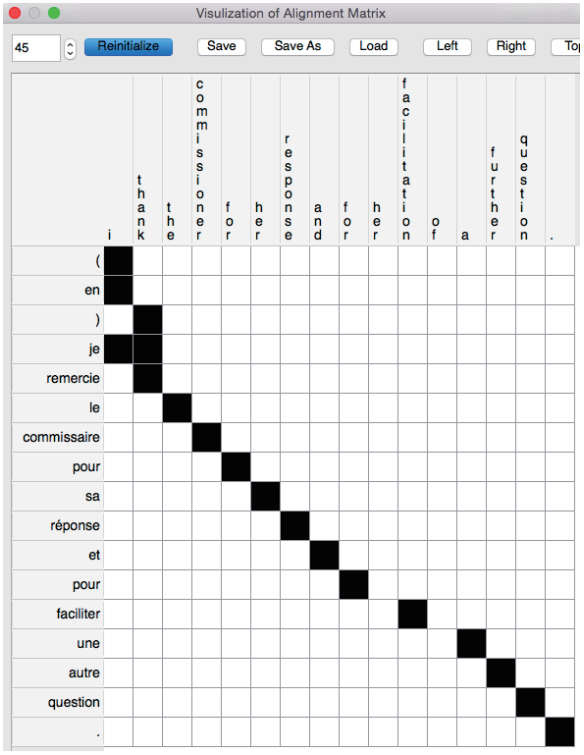

Figure 2: Sub-sentential alignment output by MGIZA++

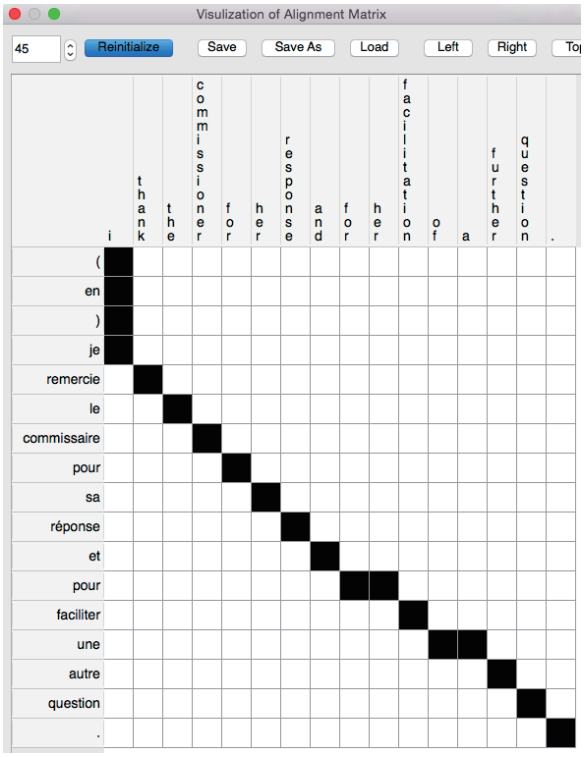

Figure 3: Sub-sentential alignment output by Cutnalign on the same sentences as in Figure 2. This alignment is better.

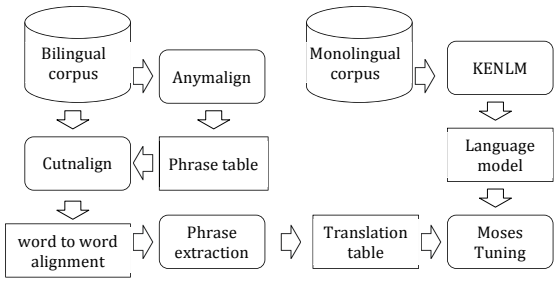

Figure 5: Main training modules and models of Anymalign-4 + Cutnalign with Moses Toolkits.

3.2 Results

The results of the experiments are presented in Table 1. Anymalign-4 + Cutnalign produces better alignments in French-English and Spanish-Portuguese relatively to MGIZA++, and gets slightly better scores for this reason. In Finnish-English, however, the score in BLEU decreases. The phrase tables extracted from our method are much smaller than those obtained with MGIZA++: they contain twice less entries in average. The average lengths of entries are almost equal to those in MGIZA++'s phrase tables.

Because we set the running time of Anymalign to half that of MGIZA++, the total time required by our alignment method, which includes two parts Anymalign-4 and Cutnalign, is much shorter than that of MGIZA++. In addition, as smaller phrase tables are extracted by our method, lower times for tuning and for decoding are observed. Obviously, our method yields large improvements in processing speed. Moreover, alignments produced by our method still lead to state-of-the-art scores in two out of three of the representative language pair experiments.

4 Conclusion

In this paper, we have shown that it is possible to accelerate development of SMT systems following the work of [10] on bilingual hierarchical sub-sentential alignment. The resulting alignments in several machine translation experiments led to superior or similar translation quality result and a significant reduction of processing time. In addition, as stressed in [13], because better alignments extract fewer phrase pairs, the decoding time is also reduced when searching the most possible translation from the phrase tables.

Many future research directions are possible. Firstly, our visualization program shows that, occasionally, some big blocks in alignment matrices cannot be segmented. We will try to solve this problem by improving our implementation of Cutnalign. Secondly, according to experiments in Section 3, the mitigated results in Finnish to English ask for experiments on more language pairs. Finally, thanks to the anytime nature of Anymalign, we want to inquire the relationship between the time alloted

Language pair	Setting	BLEU (%)	WER (%)	NIST	Entries (millions)	Avg lenth of source entries	Avg lenth of target entries	Total time
fr-en	MGIZA++	34.28	45.93	9.0816	23.8	4.42	3.97	14h14min
	Anymalign-4 + Cutnalign	34.33	46.10	9.0623	12.7	4.34	3.95	10h51min
es-pt	MGIZA++	36.49	46.54	9.3545	30.6	4.35	4.30	25h37min
	Anymalign-4 + Cutnalign	36.70	46.26	9.3913	16.8	4.38	4.28	18h03min
fi-en	MGIZA++	24.65	56.68	7.3026	20.3	3.26	4.26	15h13min
	Anymalign-4 + Cutnalign	23.85	59.87	6.9711	7.4	3.31	4.00	7h28min

Table 1: Results: the first three columns (BLEU, WER and NIST) report performance in machine translation. The following three columns display various characteristics of the phrase tables: the number of entries and the average length of source and target phrases in words. The last column shows the total time, i.e., training + tuning + decoding, for each translation system

to Anymalign and the translation quality. Even shorter development times may well lead to state-of-the-art translation quality.

References

[1] Necip Fazil Ayan and Bonnie J Dorr. Going beyond AER: An extensive analysis of word alignments and their impact on mt. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 9–16. Association for Computational Linguistics, 2006.

[2] Peter F Brown, Vincent J Della Pietra, Stephen A Della Pietra, and Robert L Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311, 1993.

[3] Ted Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, 19(1):61–74, 1993.

[4] William A Gale and Kenneth Ward Church. Identifying word correspondences in parallel texts. In *HLT*, volume 91, pages 152–157. Citeseer, 1991.

[5] Qin Gao and Stephan Vogel. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57. Association for Computational Linguistics, 2008.

[6] Kenneth Heafield. Kenlm: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197. Association for Computational Linguistics, 2011.

[7] Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *Machine Translation summit*, volume 5, pages 79–86, 2005.

[8] Philipp Koehn, Amittai Axelrod, Alexandra Birch, Chris Callison-Burch, Miles Osborne, David Talbot, and Michael White. Edinburgh system description for the 2005 iwslt speech translation evaluation. In *IWSLT*, pages 68–75, 2005.

[9] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180. Association for Computational Linguistics, 2007.

[10] Adrien Lardilleux, François Yvon, and Yves Lepage. Hierarchical sub-sentential alignment with anymalign. In *Proceedings of the 16th annual conference of the European Association for Machine Translation (EAMT 2012)*, pages 279–286, 2012.

[11] Adrien Lardilleux, François Yvon, and Yves Lepage. Generalizing sampling-based multilingual alignment. *Machine translation*, 27(1):1–23, 2013.

[12] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51, 2003.

[13] Liang Tian, Derek F Wong, Lidia S Chao, and Francisco Oliveira. A relationship: Word alignment, phrase table, and translation quality. *The Scientific World Journal*, 2014, 2014.

[14] Curtis R Vogel and Mary E Oman. Iterative methods for total variation denoising. *SIAM Journal on Scientific Computing*, 17(1):227–238, 1996.

[15] Hongyuan Zha, Xiaofeng He, Chris Ding, Horst Simon, and Ming Gu. Bipartite graph partitioning and data clustering. In *Proceedings of the tenth international conference on Information and knowledge management*, pages 25–32. ACM, 2001.