Higher coverage in word-to-word associations for the sampling-based alignment method

Meng Kong Yves Lepage
Graduate School of Information, Production and Systems, Waseda University
{koumou9130@ruri.,Yves.Lepage@}waseda.jp

1 Introduction

The most used method to compute word-to-word associations for statistical machine translation is a probabilistic method, based on expectation-maximization, implemented in the tool MGIZA++[1] [1]. Run in two directions and followed by the grow-diag-final heuristic[2], it allows to extract aligned phrases from a bilingual parallel corpus. The sampling-based alignment method [5], implemented in Anymalign[3], is an associative method that was proposed to easily compute word-to-word associations and extract aligned phrases. It has several advantages: it can be interrupted at any time; it is bidirectional, even multilingual by design. It is known that using it in combination with standard statistical machine translation tools results in lower translation accuracy as measured by BLEU, but it has also been shown that it delivers better results than GIZA for lexicon induction [7].

The coverage of a word association method is defined as the percentage of words aligned over the total number of words contained in the training corpus. It has been observed that not all words are necessarily aligned by Anymalign and that coverage increases with larger time-outs.

In this paper, we propose a new sampling method to increase the coverage of Anymalign in the word-to-word association task[4], to increase the number of words aligned. We also show how to keep reasonable processing times by monitoring the sizes of samples.

2 The original sampling-based alignment method

2.1 Sampling of sub-corpora

The core idea in Anymalign is to continuously randomly select several aligned sentence pairs from the input parallel corpus and process them to extract word correspondences. Such samples are called sub-corpora. The probability of drawing a sub-corpus of a given size k is proportional to a function defined as [7]:

$$P(k) \propto \frac{-1}{k \times log(1 - k/n)} \quad (1)$$

Here, n is the size of the input parallel corpus. This ensures that ultimately all line pairs of the input parallel corpus will be drawn. Since the value of $P(k)$ is close to $1/k^2$, sub-corpora of small sizes are given more probability to be drawn. Previous work ([6], [8]) has shown that smaller sub-corpora produce more numerous alignments in a shorter time. Our experiments reported in Section 4.2.2 confirm this.

Words which appear exactly on the same lines in a sub-corpus are stored as word correspondences with the number of lines they appear in. At the end of the process, when the time-out is elapsed or the process manually interrupted by the user, the same word correspondences are grouped together, and the number of lines summed up. Conditional translation probabilities can be obtained by considering each word and each of its corresponding words with their relative number of lines.

2.2 Source of incomplete coverage

Since aligned sentence pairs are selected from the input training corpus randomly, some aligned sentence pairs may not be sampled at all. If a word occurs only in such sentence pairs, it will not be aligned. Hapaxes have a higher probability to be such missed words. This reduces the coverage of the method.

It may also happen that all the words occurring in a drawn sub-corpus of aligned sentence pairs all become hapaxes in this drawn sub-corpus. In such a configuration, Anymalign is unable to separate words to align them separately, and, for the word-to-word alignement task, the result on such a sub-corpus is empty. Processing such a sub-corpus is a waste of time for the word-to-word association task. This may also reduce the number of words aligned and reduce the coverage.

[1] http://www.cs.cmu.edu/~qing/giza/
[2] http://www.statmt.org/moses/?n=FactoredTraining.AlignWords
[3] https://anymalign.limsi.fr/
[4] I.e., used with the options -n 1 -N 1 to output only word-to-word correspondences.

To increase coverage, we propose to identify particular types of words. For each of these particular types of words, we will show how to build additional specific sub-corpora. These specific sub-corpora will be processed by Anymalign in the standard way but their design will allow to easily extract the associations for the particular types of words the specific sub-corpora were created for.

3 Data and tools used in experiments

For our experiments, we use part of the French–English parallel Europarl Corpus v3 [3]. The tuning set and the test set are selected randomly, the remaining becoming the training set. The number of sentences in each set is given below:

	# of lines	avg. # of words / line	
		French	English
Training	347,614	31.52	28.61
Tuning	500	32.31	29.23
Test	5,000	32.31	29.23

For the machine translation experiments reported in Section 5, we first compute word-to-word association using the original version of Anymalign or the improved versions described in Sections 4.1 and 4.2, and we rely on these results to further build a statistical machine translation system in the standard way using tools provided in Moses [4], i.e., we just replace the use of MGIZA++ by the use of Anymalign. We use the KenLM toolkit [2] to build a target language model. We use MERT to tune the parameters of the models.

Coverage in each language will be measured as the number of different words appearing in the word-to-word association results divided by the total number of different words in the input parallel corpus.

Translation accuracy is measured using BLEU [9].

4 Building sub-corpora for different types of words

In preparatory experiments, we investigated the coverage of words per frequency in the word-to-word associations output by Anymalign. The results show that some low frequency words are never aligned, especially hapaxes. Hapaxes usually represent an important part of the vocabulary of a corpus, e.g., 30% in most languages of the Europarl Corpus [3].

Consequently, in order to increase the coverage of word-to-word associations, we distinguish between two types of words: hapaxes and non-hapax words.

4.1 Enforcing word-to-word association of hapaxes

Bearing in mind the core idea of Anymalign (Section 2.1), a natural idea to improve coverage in hapaxes is to design a a specialized sub-corpus to try to align all hapaxes in one-shot. This sub-corpus will be processed by Anymalign in the usual way.

To this end, we first build the sub-corpus made of all the lines which contain at least one hapax. Some words that were not hapaxes in the entire input corpus may appear only once in this sub-corpus, i.e., they become hapaxes inside the sub-corpus. We shall call such words "false hapaxes" by opposition to the real hapaxes. False hapaxes may share the same distribution with some real hapaxes. This will prevent them from being aligned separately. To separate false hapaxes from real hapaxes, for each of them, we draw from the input corpus (without replacement) an additional line that contains this false hapax and add it to the sub-corpus. In this way, as the distribution of any of any real hapax tend to become different from the distribution of any false hapax, real hapaxes will tend to be aligned.

4.2 Enforcing word-to-word association of non-hapax words

4.2.1 Basic principle to enforce word-to-word association of non-hapax words

Let us now consider words which are not hapaxes in the input parallel corpus Let us consider one line of the corpus. A word on this line which is not a hapax can be aligned by using a different line that contains it, assuming that the noise around it will allow to separate it from other words. Thus, for a given line, we can build a sub-corpus by drawing a different line (if possible) for each non-hapax word on this line. This will make a sub-corpus that can be processed by Anymalign in the usual way. Such sub-corpora will be reasonably small if the number of words per line is not so large, and hence can be processed reasonably fast.

4.2.2 Experiments with the basic principle

We performed experiments using the basic principle above to enforce word-to-word association for non-hapax words. We measured average the processing time for different sizes of sub-corpora. The results are shown in the graph on Figure 1.

As is already known (see Section 2.1), the smaller a sub-corpus, the faster it is processed. The time required to process a sub-corpus is approximately linear in its size in sentence pairs. Based on this result, we propose to modify the basic principle and cut sub-corpora into smaller sub-corpora so as to accelerate the overall computation of word-to-word association.

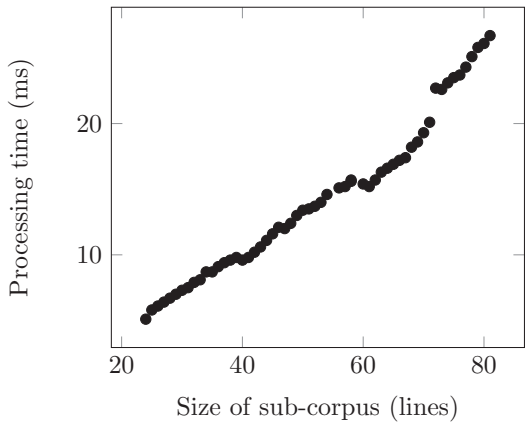

Figure 1: Processing times (in ordinates, in milliseconds) for different sizes of sub-corpora (in abscissae in number of lines)

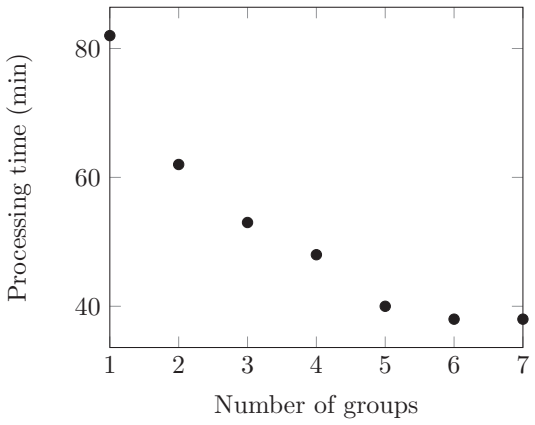

Figure 2: Total processing times (in ordinates, in minutes) for computation of word-to-word associations of non-hapax words when enforcing their alignment and grouping them on a line by groups of equal size (in abscissae)

4.2.3 Grouping non-hapax words on each line into groups of equal size

In the basic principle described above, we build a sub-corpus by basically drawing at most one additional line for each non-hapax word. Hence, the number of lines drawn is the same as or less than the total number of words on the line. So as to make the sub-corpora smaller, and thus reduce processing time, we propose to group non-hapax words on each line into several groups of equal size and to build one sub-corpus for each group. Each sub-corpus will be processed in the standard way by Anymalign.

We investigated the effect of this grouping on the total processing time of word-to-word association, depending on the number of groups used (so not the size itself). The graph in Figure 2 shows that the total processing time decreases as the number of groups increases. It significantly decreases at the beginning: it is divided by two when comparing the use of 5 groups with the use of 1 group (i.e., using the basic principle). Using more than 5 groups does not lead to any significant reduction.

5 Evaluation

Table 1 shows the different values of coverage and translation accuracy obtained using different settings based on the work presented above. For comparison with the original sampling-based method we allot the same time-out as the one obtained with the use of all our proposed improvements (last line in Table 1). Our method yields increased coverage (almost +15%) and significantly better BLEU scores (by almost two confidence intervals).

We first inspected the effect of aligning hapaxes alone. We do not build a translation system, as such a system, based on very low frequency words only, is not meaningful. The original sampling-based alignment method with a time-out of 42 minutes can align 65% of the hapaxes. In comparison the method proposed in Section 4.1 covers 86% of the hapaxes. This represents 52% of all words as indicated in Table 1.

The original version of Anymalign implements an anytime method, i.e., it can be interrupted at any moment by passing a time-out as optional argument or the user interrupting the process. Quality is not so much a matter of time, but coverage is. We thus ran the version of Anymalign that comprises our proposed improvements (enforcing word-to-word associations of hapaxes and non-hapax words, with 5 groups of non-hapax words on each line), using different time-outsand measured the translation accuracy on the same task. The results given in Figure 3 show that, after 8 minutes, our method yields better BLEU scores than the original sampling-based alignment method.

6 Conclusion

In this paper, we showed how to improve the coverage of the sampling-based alignment method, implemented in Anymalign, when used to compute word-to-word association. We proposed to enforce the alignment of hapaxes in one-shot by building a specific sub-corpus for that. We also enforced the alignment of other words by building specific sub-corpora for each line, and showed how to accelerate the overall alignment process by grouping non-hapax words on one line into groups of the same size. Experiments conducted on the French-English part of the parallel corpus Europarl v3, showed that we achieved our goals in increasing the coverage, increasing the translation accuracy, and simultaneously reducing the needed processing time.

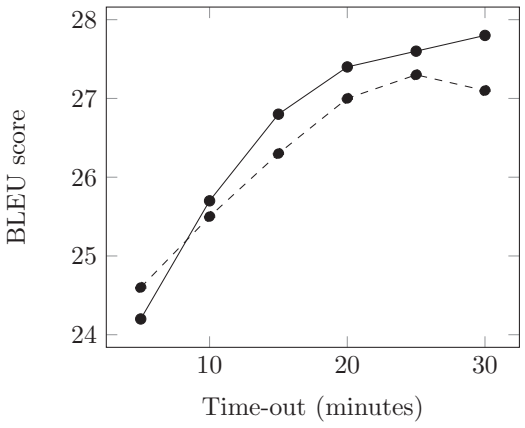

Figure 3: Translation accuracy as measured by BLEU when using different time-outs (in abscissae, in minutes) for the basic usage of Anymalign (dotted line) and for the version of Anymalign incorporating all of our improvements, i.e., enforcing word-to-word associations of hapaxes and non-hapax words with splitting in 5 groups on each line (plain line)

Method	Time (min)	BLEU	Coverage (all voc.)
MGIZA++	297	34.46±0.68	80%
Anymalign	42	27.29±0.67	75%
H	1	not relevant	52%
NH (1 group)	87	28.19±0.68	81%
NH (2 groups)	51	28.36±0.67	83%
NH (5 groups)	41	28.06±0.67	83%
H+NH (1 grp)	88	28.21±0.70	89%
H+NH (2 grps)	52	28.31±0.68	89%
H+NH (5 grps)	42	28.11±0.71	88%

Table 1: Processing times (in minutes) for training, BLEU scores with confidence intervals, and coverage obtained in different settings of Anymalign. H stands for the enforcement of word-to-word associations of hapaxes using one additional sub-corpus (Section 4.1). NH stands for the enforcement of word-to-word associations of non-hapax words (Section 4.2), the number of groups used is given in parentheses

Acknowledgments

This paper is part of the outcome of research performed under a Waseda University Grant for Special Research Project (Project number: 2015A-063).

References

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

[2] Kenneth Heafield. KenLM: faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, United Kingdom, July 2011.

[3] Philipp Koehn. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the tenth Machine Translation Summit (MT Summit X)*, pages 79–86, Phuket, September 2005.

[4] Philipp Koehn, Hieu Hoang, Alexandra Birch, and others. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, pages 177–180, Prague, Czech Republic, June 2007.

[5] Adrien Lardilleux and Yves Lepage. A truly multilingual, high coverage, accurate, yet simple, subsentential alignment method. In *Proceedings of the Xth conference of the Association for Machine Translation in the Americas*, pages 125–132, Waikiki, Hawai'i, oct 2008.

[6] Adrien Lardilleux and Yves Lepage. Sampling-based multilingual alignment. In *International Conference on Recent Advances in Natural Language Processing (RANLP 2009)*, pages 214–218, Borovets, Bulgaria, sept 2009.

[7] Adrien Lardilleux, Yves Lepage, and Julien Gosme. Bilingual lexicon induction: Effortless evaluation of word alignment tools and production of resources for improbable language pairs. In *Proceedings of the 7th international conference on Language Resources and Evaluation (LREC 2010)*, pages 252–256, Valletta, Malta, may 2010. European Language Resources Association (ELRA).

[8] Adrien Lardilleux, Yves Lepage, and François Yvon. The contribution of low frequencies to multilingual sub-sentential alignment: a differential associative approach. *International Journal of Advanced Intelligence*, 3(2):189–217, July 2011.

[9] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, pages 311–318, Philadelphia, July 2002.