Exploring N-grams Distribution for Sampling-based Alignment

Juan Luo*, **Adrien Lardilleux****, **Yves Lepage***

* IPS, Waseda University, 808-0135 Kitakyushu, Japan
** TLP Group, LIMSI-CNRS, 91403 Orsay, France
juanluoonly@suou.waseda.jp, adrien.lardilleux@limsi.fr, yves.lepage@aoni.waseda.jp

Abstract

We describe an approach to improve the performance of sampling-based multilingual alignment on translation tasks by investigating the distribution of n-grams in the translation tables. This approach consists in enforcing the alignment of n-grams. We compare the quality of phrase translation tables output by this approach and that of MGIZA++ in statistical machine translation tasks. We report significant improvements for this approach and show that merging translation tables outperforms state-of-the-art techniques.

Keywords: phrase translation table, alignment, statistical machine translation task.

1. Introduction

Phrase translation tables play an important role in the process of building machine translation systems. This quality is crucial for the quality of translation. The most widely used state-of-the-art tool to generate phrase translation tables is MGIZA++ (Gao and Vogel, 2008), which trains the IBM models (Brown et al., 1993) and the HMM introduced in (Vogel et al., 1996) in combination with the Moses toolkit (Koehn et al., 2007).

In this paper, we investigate a different approach to the production of phrase translation tables: the sampling-based approach (Lardilleux and Lepage, 2009), available as a free open-source tool called Anymalign.[1] Being in line with the associative alignment trend (see e.g. (Gale and Church, 1991; Melamed, 2000; Moore, 2005)), it is much simpler than the models implemented in MGIZA++, which are in line with the estimating trend (e.g. (Brown et al., 1991; Och and Ney, 2003; Liang et al., 2006)).

In sampling-based alignment, only those sequences of words that appear exactly in the same sentences of the corpus are considered for alignment. The key idea is to produce more candidate words by artificially reducing the size of the input corpus, i.e., many subcorpora of small sizes are obtained by sampling and processed one after another. Indeed, the smaller a subcorpus, the less frequent its words, and the more likely they are to share the same distribution.

The subcorpus selection process is guided by a probability distribution that ensures a proper coverage of the input parallel corpus:

$$\mathrm{p}(k) = \frac{-1}{k \log(1-k/n)} \quad \text{(to be normalized)}$$

where k denotes the size (number of sentences) of a subcorpus and n the size of the complete input corpus. This function is very close to $1/k^2$ and gives more credit to small subcorpora, which happen to be the most productive (Lardilleux and Lepage, 2009). Once the size of a subcorpus has been chosen according to this distribution, its sentences are randomly selected from the complete input corpus according to a uniform distribution. Then, from each subcorpus, sequences of words that share the same distribution are extracted to constitute alignments along with the number of times they were aligned.[2]

Eventually, the list of alignments is turned into a full-fledged translation table by calculating various features for each alignment. In the following, we use two translation probabilities and two lexical weights as proposed by (Koehn et al., 2003), as well as the commonly used phrase penalty, for a total of five features.

One important feature of the sampling-based alignment method is that it is *anytime* in essence: the number of random subcorpora to be processed is not set in advance, so the alignment process can be interrupted at any moment. Contrary to many approaches, *quality* is not a matter of time, however *quantity* is: the longer the aligner runs (i.e. the more subcorpora processed), the more alignments produced, and the more reliable their associated translation probabilities.

Intuitively, since the sampling-based alignment process can be interrupted without sacrificing the quality of alignments, it should be possible to allot more processing time for n-grams of similar lengths in both languages and less time to very different lengths. For instance, a source bigram is much less likely to be aligned with a target 9-gram than with a bigram or a trigram. The experiments reported in this paper make use of the anytime feature of Anymalign and of the possibility of allotting time freely.

This paper is organized as follows: Section 2 defines the problem. Section 3 proposes a variant in order to improve the translation performance. Section 4 describes the merge of two aligners' phrase translation tables. Section 5 provides the conclusion.

[1] http://users.info.unicaen.fr/~alardill/anymalign/

[2] Contrary to the widely used terminology where it denotes a set of links between the source and target words of a sentence pair, we call "alignment" a (source, target) phrase pair, i.e., it corresponds to an entry in the so-called [phrase] translation tables.

2. Preliminary experiment

In order to measure the performance of the sampling-based alignment approach implemented in Anymalign in statistical machine translation tasks, we conducted a preliminary experiment and compared with the standard alignment setting: symmetric alignments obtained from MGIZA++. Although Anymalign and MGIZA++ are both capable of parallel processing, for fair comparison in time, we run them as single processes in all our experiments.

2.1. Experimental setup

A sample of the French-English parts of the Europarl parallel corpus was used for training, tuning and testing. A detailed description of the data used in the experiments is given in Table 1. To perform the experiments, a standard statistical machine translation system was built for each different alignment setting, using the Moses decoder (Koehn et al., 2007) MERT (Minimum Error Rate Training) (Och, 2003) and the SRILM toolkit (Stolcke, 2002). As for the evaluation of translations, the BLEU metric (Papineni et al., 2002) was used.

		French	English
Train	sentences	100,000	100,000
	words	3,986,438	2,824,579
	words/sentence	38	27
Dev	sentences	500	500
	words	18,120	13,261
	words/sentence	36	26
Test	sentences	1,000	1,000
	words	38,936	27,965
	words/sentence	37	27

Table 1: Statistics on the French-English parallel corpus used for the training, development, and test sets.

2.2. Problem definition

In a first setting, we evaluated the quality of translations output by the Moses decoder using the phrase table obtained by making MGIZA++'s alignments symmetric. In a second setting, this phrase table was simply replaced by that produced by Anymalign. Since Anymalign can be stopped at any time, for a fair comparison it was run for the same amount of time as MGIZA++: seven hours in total. The experimental results are shown in Table 2. In order to investigate the differences between MGIZA++ and Anymalign phrase translation tables, we analyzed the distribution of n-grams of both aligners, The distributions are shown in Table 6(a) and Table 6(b).

	BLEU
MGIZA++	27.42
Anymalign (baseline)	22.85

Table 2: Evaluation results on a statistical machine translation task using phrase tables obtained from MGIZA++ and Anymalign (baseline).

In Anymalign's phrase translation table, the number of alignments is 8 times that of 1×1 n-grams in MGIZA++ translation table, or twice the number of 1×2 n-grams or 2×1 n-grams in MGIZA++ translation table. Along the diagonal (m $\times$ m n-grams) for $m > 2$, the number of alignments in Anymalign table is approximately hundred times less than in MGIZA++ table. This confirms the results given in (Lardilleux et al., 2009) that the sampling-based approach excels in aligning unigrams, which makes it better at multilingual lexicon induction than, e.g., MGIZA++. However, its phrase tables do not reach the performance of symmetric alignments from MGIZA++ on translation tasks. This basically comes from the fact that Anymalign does not align enough long n-grams.

3. Anymalign1-N

3.1. Phrase translation subtables

To solve the above-mentioned problem, we propose a method to force the sampling-based approach to align more n-grams.

Consider that we have a parallel input corpus, i.e., a list of (source, target) sentence pairs, for instance, in French and English. Groups of characters that are separated by spaces in these sentences are considered as words. Single words are referred to as unigrams, and sequences of two and three words are called bigrams and trigrams, respectively.

Theoretically, since the sampling-based alignment method excels at aligning unigrams, we could improve it by making it align bigrams, trigrams, or even longer n-grams as if they were unigrams. We do this by replacing spaces between words by underscore symbols and reduplicating words as many times as needed, which allows to make bigrams, trigrams, and longer n-grams appear as unigrams. Table 3 depicts the way of forcing n-grams into unigrams. The same trick was used in a work by (Henríquez Q. et al., 2010).

It is thus possible to use various parallel corpora, with different segmentation schemes in the source and target parts. We refer to a parallel corpus where source n-grams and target m-grams are assimilated to unigrams as a *unigramized n-m corpus*. These corpora are then used as input to Anymalign to produce phrase translation subtables, as shown in Table 4. Practically, we call Anymalign1-N the process of running Anymalign with all possible unigramized *n-m* corpora, with *n* and *m* both ranging from 1 to a given N. In total, this corresponds to N $\times$ N runs of Anymalign. All phrase translation subtables are finally merged together into one large translation table, where translation probabilities are re-estimated given the complete set of alignments.

Although Anymalign is capable of directly producing alignments of sequences of words, we use it with a simple filter[3] so that it only produces (typographic) unigrams in output, i.e., n-grams and m-grams assimilated to unigrams in the input corpus. This choice was made because it is useless to produce alignments of sequences of words, since we are only interested in *phrases* in the subsequent

[3]Option `-N 1` in the program.

n	French	English
1	le debat est clos .	the debate is closed .
2	le_debat debat_est est_clos clos_.	the_debate debate_is is_closed closed_.
3	le_debat_est debat_est_clos est_clos_.	the_debate_is debate_is_closed is_closed_.
4	le_debat_est_clos debat_est_clos_.	the_debate_is_closed debate_is_closed_.
5	le_debat_est_clos_.	the_debate_is_closed_.

Table 3: Transforming n-grams into unigrams by inserting underscores and reduplicating words for both the French part and English part of the input parallel corpus.

		Target				
		unigrams	bigrams	trigrams	···	N-grams
Source	unigrams	TT1 × 1	TT1 × 2	TT1 × 3	···	TT1 × N
	bigrams	TT2 × 1	TT2 × 2	TT2 × 3	···	TT2 × N
	trigrams	TT3 × 1	TT3 × 2	TT3 × 3	···	TT3 × N
	···	···	···	···	···	···
	N-grams	TTN × 1	TTN × 2	TTN × 3	···	TTN × N

Table 4: List of n-gram translation subtables (TT) generated from the training corpus. These subtables will then be merged together into a single translation table.

machine translation tasks. Those phrases are already contained in our (typographic) unigrams: all we need to do to get the original segmentation is to remove underscores from the alignments.

3.2. Equal time configuration

The same experimental process (i.e., replacing the translation table) as in the preliminary experiment was carried out on Anymalign1-N with equal time distribution, i.e., uniformly distributed time among subtables. For a fair comparison, the same amount of time was given: seven hours in total. The results are given in Table 5. On the whole, MGIZA++ significantly outperforms Anymalign1-N, by more than 4 BLEU points. However, the proposed approach, Anymalign1-N, produces better results than Anymalign in its basic version, with the best increase with Anymalign1-4 (+1.4 BP).

The comparison of Table 6(c) (see last page) and Table 6(a) shows that Anymalign1-N delivers too many alignments outside of the diagonal ($m \times m$ n-grams) and still not enough along the diagonal. Consequently, this number of alignments should be lowered. A way of doing so is by giving less time for alignments outside of the diagonal.

3.3. Time distribution among subtables

To this end, we distribute the total alignment time among translation subtables according to the standard normal distribution:

$$\phi(n,m) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(n-m)^2}$$

The alignment time allotted to the subtable between source n-grams and target m-grams will thus be proportional to $\phi(n,m)$.

In a third evaluation, we compare this new setting (with a total amount of processing time of 7 hours) with MGIZA++, Anymalign in its standard use, and Anymalign1-N with equal time distribution (Table 5). There is an increase in BLEU scores for almost all Anymalign1-N, from Anymalign1-3 to Anymalign1-10, when compared with equal time distribution. The greatest increase in BLEU is obtained for Anymalign1-10 (almost +2 BP). Anymalign1-4 shows the best translation quality among all other settings, but gets a less significant improvement (+0.2 BP).

Again, we investigated the number of entries in Anymalign1-N run with this normal time distribution. We compare the number of entries in Table 6 in Anymalign1-4 with (c) equal time distribution and (d) standard normal time distribution (see last page). The number of phrase pairs on the diagonal roughly doubled when using standard normal time distribution. We can see a significant increase in the number of phrase pairs of similar lengths, while the number of phrase pairs with different lengths tends to decrease slightly. This means that the standard normal time distribution allowed us to produce much more numerous useful alignments (a priori, phrase pairs with similar lengths), while maintaining the noise (phrase pairs with different lengths) to a low level, which is a neat advantage over the original method.

4. Merging translation tables

In order to check exactly how different the translation table of MGIZA++ and that of Anymalign are, we performed an additional set of experiments in which MGIZA++'s translation table is merged with that of Anymalign baseline. As for the feature scores in the translation tables for the intersection part of both aligners, we adopted parameters either from MGIZA++ or from Anymalign for evaluation.

Evaluation results on machine translation tasks with merged translation tables are given in Table 5. This setting outperforms MGIZA++ on BLEU scores. The translation table with Anymalign parameters for the intersection part is slightly behind the translation table with MGIZA++ parameters. This may indicate that the feature scores in Anymalign translation table need to be revised.

5. Conclusions and future work

In this paper, by examining the distribution of n-grams in Anymalign phrase translation tables, we presented a method to improve the translation quality of the sampling-based sub-sentential alignment approach implemented in Anymalign: firstly, Anymalign was forced to align n-grams as if they were unigrams; secondly, time was unevenly distributed over subtables; thirdly, merging of the

MGIZA++	Anymalign (baseline)	Anymalign1-N										Merge	
			1-1	1-2	1-3	1-4	1-5	1-6	1-7	1-8	1-9	1-10	MGIZA++ para.	Anymalign para.
27.42	22.85	equal time	19.84	24.06	24.03	**24.23**	23.76	23.49	23.71	22.53	22.96	21.82	**27.54**	**27.47**
		std.norm.	19.84	24.04	24.41	**24.42**	24.36	24.03	24.05	23.66	24.02	23.61		

Table 5: Evaluation results.

two aligners' phrase translation tables was introduced. A baseline statistical machine translation system was built to compare the translation performance of two aligners: MGIZA++ and Anymalign. Anymalign1-N, the method presented here, obtains significantly better results than the original method, the best performance being obtained with Anymalign1-4. Merging Anymalign's translation table with that of MGIZA++ allows to outperform MGIZA++ alone. In the future, we intend to modify the feature scores computed by Anymalign in order to make it better suited to statistical machine translation tasks.

Acknowledgements

Part of the research presented in this paper has been done under a Japanese grant-in-aid (Kakenhi C, A11515600: Improvement of alignments and release of multilingual syntactic patterns for statistical and example-based machine translation).

References

Brown, P., Lai, J. and Mercer, R. (1991). Aligning sentences in parallel corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL'91)*.

Brown, P., Pietra, S. D., Pietra, V. D. and Mercer, R. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics* 19(2):263–311.

Gale, W. and Church, K. (1991). Identifying word correspondences in parallel texts. In *Proceedings of the fourth DARPA workshop on Speech and Natural Language*.

Gao, Q. and Vogel, S. (2008). Parallel implementations of word alignment tool. In Association for Computational Linguistics (ed.), *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*.

Henríquez Q., A. C., Costa-jussà, R. M., Daudaravicius, V., Banchs, E. R. and Mariño, B. J. (2010). Using collocation segmentation to augment the phrase table. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR (WMT 2010)*.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A. and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*.

Koehn, P., Och, F. J. and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*.

Lardilleux, A., Chevelu, J., Lepage, Y., Putois, G. and Gosme, J. (2009). Lexicons or phrase tables? An investigation in sampling-based multilingual alignment. In Mikel Forcada and Andy Way (eds.), *Proceedings of the third workshop on example-based machine translation*.

Lardilleux, A. and Lepage, Y. (2009). Sampling-based multilingual alignment. In *International Conference on Recent Advances in Natural Language Processing (RANLP 2009)*.

Liang, P., Taskar, B. and Klein, D. (2006). Alignment by agreement. In *Proceedings of the Human Language Technology Conference of the NAACL*.

Melamed, D. (2000). Models of translational equivalence among words. *Computational Linguistics* 26(2):221–249.

Moore, R. (2005). Association-based bilingual word alignment. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics (ACL 2003)*, volume 1.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. In *Computational Linguistics*, volume 29.

Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*.

Stolcke, A. (2002). SRILM-an extensible language modeling toolkit. In *Seventh International Conference on Spoken Language Processing*, volume 2.

Vogel, S., Ney, H. and Tillman, C. (1996). HMM-based word alignment in statistical translation. In *Proceedings of the 16th International Conference on Computational Linguistics (Coling'96)*.

(a) Distribution of phrase pairs in MGIZA++'s translation table.

		Target							
		unigrams	bigrams	trigrams	4-grams	5-grams	6-grams	7-grams	total
Source	unigrams	**89,788**	44,941	10,700	2,388	486	133	52	148,488
	bigrams	61,007	**288,394**	86,978	20,372	5,142	1,163	344	463,400
	trigrams	19,235	149,971	**373,991**	105,449	27,534	7,414	1,857	685,451
	4-grams	5,070	47,848	193,677	**335,837**	106,467	31,011	9,261	729,171
	5-grams	1,209	13,984	73,068	193,260	**270,615**	98,895	32,349	683,380
	6-grams	332	3,856	24,333	87,244	177,554	**214,189**	88,700	596,208
	7-grams	113	1,103	7,768	33,278	91,355	157,653	**171,049**	462,319
	total	176,754	550,097	770,515	777,828	679,153	510,458	303,612	3,768,417

(b) Distribution of phrase pairs in Anymalign's translation table (baseline).

		Target								
		unigrams	bigrams	trigrams	4-grams	5-grams	6-grams	7-grams	⋯	total
Source	unigrams	**791,099**	105,961	9,139	1,125	233	72	37	⋯	1,012,473
	bigrams	104,633	**21,602**	4,035	919	290	100	44	⋯	226,176
	trigrams	10,665	4,361	**2,570**	1,163	553	240	96	⋯	92,268
	4-grams	1,698	1,309	1,492	**1,782**	1,158	573	267	⋯	61,562
	5-grams	378	526	905	1,476	**1,732**	1,206	642	⋯	47,139
	6-grams	110	226	467	958	1,559	**1,694**	1,245	⋯	40,174
	7-grams	40	86	238	536	1,054	1,588	**1,666**	⋯	35,753
	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯
	total	1,022,594	230,400	86,830	55,534	42,891	37,246	34,531	⋯	1,371,865

(c) Anymalign1-4 with equal time for each $n \times m$ n-grams alignments.

		Target							
		unigrams	bigrams	trigrams	4-grams	5-grams	6-grams	7-grams	total
Source	unigrams	**171,077**	118,848	39,253	13,327	0	0	0	342,505
	bigrams	119,953	**142,721**	67,872	24,908	0	0	0	355,454
	trigrams	45,154	75,607	**86,181**	42,748	0	0	0	249,690
	4-grams	15,514	30,146	54,017	**60,101**	0	0	0	159,778
	5-grams	0	0	0	0	0	0	0	0
	6-grams	0	0	0	0	0	0	0	0
	7-grams	0	0	0	0	0	0	0	0
	total	351,698	367,322	247,323	141,084	0	0	0	1,107,427

(d) Anymalign1-4 with standard normal time distribution.

		Target							
		unigrams	bigrams	trigrams	4-grams	5-grams	6-grams	7-grams	total
Source	unigrams	**255,443**	132,779	13,803	469	0	0	0	402,494
	bigrams	134,458	**217,500**	75,441	8,612	0	0	0	436,011
	trigrams	15,025	86,973	**142,091**	48,568	0	0	0	292,657
	4-grams	635	10,516	61,741	**98,961**	0	0	0	171,853
	5-grams	0	0	0	0	0	0	0	0
	6-grams	0	0	0	0	0	0	0	0
	7-grams	0	0	0	0	0	0	0	0
	total	405,561	447,768	293,076	156,610	0	0	0	1,303,015

Table 6: Distribution of phrase pairs in translation tables.