Producing Translation Tables by Separate N-grams Subtables

Juan Luo, Jing Sun, and Yves Lepage
IPS, Waseda University
{juanluoonly@suou,cecily.sun@akane,yves.lepage@aoni}.waseda.jp

Abstract

By investigating the distribution of phrase pairs in translation tables, this paper describes an approach to expand the number of n-gram alignments in translation tables output by the sampling-based alignment method. Translation subtables are produced to increase the number of n-grams. Standard normal time distribution is used to adapt the distribution of n-grams in translation tables and leads to better evaluation results than the original approach. Merging translation table of the sampling-based alignment method and that of MGIZA++ is also examined. An improvement of 1.57 BLEU point is reported by applying the technique of standard normal time distribution and merging translation tables is shown to outperform the state-of-the-art alignment method.

1 Introduction

In the translation process, translation tables play a vital role as their quality has an impact on the translation quality. The most widely used tool to generate translation tables is GIZA++ [7], which trains the IBM models [1] and the HMM introduced in [10] in combination with the Moses toolkit [4]. MGIZA++, a multi-threaded word aligner based on GIZA++, is proposed by [2].

In this paper, we investigate a different approach to the production of phrase translation tables: the sampling-based approach [5], available as a free open-source tool called Anymalign.[1] 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.

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.

2 Problem Statement

In order to measure the performance of 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.

A sample of the French-English parts of the Europarl parallel corpus was used. The training set is made of 100k sentences. The development set contains 500 sentences, and 1,000 sentences were used for testing. To perform the experiments, a standard statistical machine translation system was built for each different alignment setting, using the Moses decoder [4], MERT (Minimum Error Rate Training) [6], and the SRILM toolkit [9]. As for the evaluation of translations, the BLEU metric [8] was used.

In a first setting, we evaluated the quality of translations output by the Moses decoder using the translation table obtained by making MGIZA++'s alignments symmetric. In a second setting, this translation 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 3. In order to investigate the differences between MGIZA++ and Anymalign translation tables, we analyzed the distribution of n-grams of both aligners, The distributions are shown in Table 4(a) and Table 4(b). The number of n-grams ($n \geq 2$) in Anymalign's translation table is much less than in MGIZA++ table.

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

3 Anymalign1-N

3.1 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. The same trick was used in a work by [3].

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 translation subtables, as shown in Table 1. 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 × N runs of Anymalign. All translation subtables are finally merged together into one large translation table, where translation probabilities are re-estimated given the complete set of alignments.

Table 1: List of n-gram translation subtables (TT) generated from the training corpus. These subtables will then be merged together into a single translation table.

		Target				
		1-grams	2-grams	3-grams	⋯	N-grams
Source	1-grams	TT1 × 1	TT1 × 2	TT1 × 3	⋯	TT1 × N
	2-grams	TT2 × 1	TT2 × 2	TT2 × 3	⋯	TT2 × N
	3-grams	TT3 × 1	TT3 × 2	TT3 × 3	⋯	TT3 × N
	⋯	⋯	⋯	⋯	⋯	⋯
	N-grams	TTN × 1	TTN × 2	TTN × 3	⋯	TTN × N

Although Anymalign is capable of directly producing alignments of sequences of words, we use it with a simple filter[2] 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 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.

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

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 3. 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 4(c) (see last page) and Table 4(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 Standard Normal Time Distribution

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 MGIZA++, Anymalign in its standard use, and Anymalign1-N with equal time distribution (Table 3). There is an increase in BLEU scores for almost all Anymalign1-N, from Anymalign1-3 to Anymalign1-10, when compared with equal time distribution. 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 4 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 examine 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 3. 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. An analysis of feature scores of Anymalign (TT1) and MGIZA++ (TT2) is shown in Table 2.

Table 2: Analysis of feature scores.

features	mean +/- stddev.
$\phi(f\|e)$(TT1-TT2)	-0.044755 +/- 0.257766
$lex(f\|e)$(TT1-TT2)	0.122147 +/- 0.389999
$\phi(e\|f)$(TT1-TT2)	0.007148 +/- 0.394855
$lex(e\|f)$(TT1-TT2)	0.151258 +/- 0.263210

5 Conclusions

We have described an approach to improve the translation quality of the sampling-based alignment method. This approach is based on expanding the number of n-grams by investigating the distribution of phrase pairs in translation tables. Translation subtables are used to increase the number of n-grams. Furthermore, standard normal time distribution is applied to adapt the distribution of n-grams in translation tables, which leads to significantly better evaluation results than the original approach (+1.57 BLEU points). Merging translation tables allows to outperform MGIZA++ alone.

References

[1] 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.

[2] 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.

[3] A. Carlos Henríquez Q., R. Marta Costa-jussà, Vidas Daudaravicius, E. Rafael Banchs, and B. José Mariño. Using collocation segmentation to augment the phrase table. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR (WMT 2010)*, pages 98–102, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[4] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 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, 2007.

[5] 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, 2009.

[6] Franz Josef Och. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics (ACL 2003)*, volume 1, pages 160–167, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[7] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. In *Computational Linguistics*, volume 29, pages 19–51, 2003.

[8] 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, 2002.

[9] Andreas Stolcke. SRILM-an extensible language modeling toolkit. In *Seventh International Conference on Spoken Language Processing*, volume 2, pages 901–904, Denver, Colorado, 2002.

[10] Stephan Vogel, Hermann Ney, and Christoph Tillman. HMM-based word alignment in statistical translation. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING 1996)*, pages 836–841, Copenhagen, Denmark, 1996.

Table 3: Evaluation results.

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

Table 4: Distribution of n-grams in phrase translation tables.

(a) Distribution of n-grams in MGIZA++'s translation table.

		Target							
		1-grams	2-grams	3-grams	4-grams	5-grams	6-grams	7-grams	total
Source	1-grams	**89,788**	44,941	10,700	2,388	486	133	52	148,488
	2-grams	61,007	**288,394**	86,978	20,372	5,142	1,163	344	463,400
	3-grams	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 n-grams in Anymalign's translation table (baseline).

		Target								
		1-grams	2-grams	3-grams	4-grams	5-grams	6-grams	7-grams	···	total
Source	1-grams	**791,099**	105,961	9,139	1,125	233	72	37	···	1,012,473
	2-grams	104,633	**21,602**	4,035	919	290	100	44	···	226,176
	3-grams	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							
		1-grams	2-grams	3-grams	4-grams	5-grams	6-grams	7-grams	total
Source	1-grams	**171,077**	118,848	39,253	13,327	0	0	0	342,505
	2-grams	119,953	**142,721**	67,872	24,908	0	0	0	355,454
	3-grams	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							
		1-grams	2-grams	3-grams	4-grams	5-grams	6-grams	7-grams	total
Source	1-grams	**255,443**	132,779	13,803	469	0	0	0	402,494
	2-grams	134,458	**217,500**	75,441	8,612	0	0	0	436,011
	3-grams	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