A Comparison of Association and Estimation Approaches to Alignment in Word-to-Word Translation

Juan Luo **Yves Lepage**
Graduate School of Information, Production and Systems
Waseda University
Fukuoka 808-0135, Japan
juan.luo@suou.waseda.jp yves.lepage@waseda.jp

Abstract

Word alignment is the very first step in the process of building statistical machine translation systems. In this paper, we investigate one-to-one alignments output by the sampling-based alignment method, which is an instance of the associative method. The contribution of one-to-one alignments contained in phrase tables to translation quality is examined on 10 European language pairs. We compare the sampling-based alignment method with the state-of-the-art estimative method. It is shown that the one-to-one alignments produced by sampling-based alignment method can achieve competitive results in a lesser amount of time.

Keywords: Word alignment, phrase table, statistical machine translation

1 Introduction

Given a parallel corpus, word alignment identifies the correspondences between words in the source language and those in the target language. It is mainly used to constitute the *phrase table*, which is a fundamental component in the context of a statistical machine translation system. A phrase table is a list of phrase pairs that are translations of each other with feature scores. It is usually constructed in two steps: firstly, generating source-to-target and target-to-source word alignments; secondly, extracting bilingual phrase pairs from these alignments through heuristic combination of both directions. As a stand-alone application, word alignment is also used in bilingual terminology extraction [1] and creation of lexicon entries [2].

Dominant approaches to word alignment can be categorized into two groups: the *estimation approach* and the *association approach*.

The estimation approach employs statistical models and the parameters are estimated through maximization process. It originates from the IBM models [3] and is augmented by an HMM-based model [4]. This constitutes the standard estimative method. Many studies are carried out in this trend [5; 6]. The association approach tries to utilize different similarity measures and association tests, for instance, mutual information [7], or log-likelihood-ratio association measure [8].

Analyzing the distribution of phrases used in the decoding process, we found that, even in the standard phrase-based statistical machine translation setting, the majority of phrases used during translation is one-to-one alignments. This motivated us to perform a comparison of one-to-one alignments between an instance of the association approach, sampling-based alignment method, and the standard estimative method.

As for comparison, we consider two criteria:

- Translation quality as a final criterion.
- Speed as a practical criterion.

While the standard estimative method is generally computationally expensive, we show that, based on these criteria, the sampling-based alignment method not only produces one-to-one alignments that yield better translation quality, but also reduces drastically the processing time needed.

The remainder of this paper is organized as follows. In Section 2, we present an analysis of phrase lengths used during the translation in the standard pipeline. In Section 3, we briefly introduce the sampling-based alignment method. Section 4 reports experiments of one-to-one alignments between two approaches and evaluation results are analyzed. We conclude in Section 5 with future works.

Table 1: Statistics on the parallel corpus.

		da	de	el	es	fi	fr	it	nl	pt	sv	en
Train	sentences						347,614					
	word tokens	9,503,830	9,545,086	10,064,464	10,515,842	7,210,239	11,615,955	10,216,790	10,067,563	10,350,101	9,019,636	10,014,963
	word types	144,404	158,606	124,035	89,378	289,054	72,042	86,946	116,165	88,872	148,274	57,728
Dev.	sentences						500					
	word tokens	14,032	14,062	14,672	15,440	10,580	17,132	15,106	14,710	15,348	13,271	14,697
	word types	3,375	3,664	4,067	3,489	4,576	3,395	3,558	3,291	3,595	3,497	2,929
Test	sentences						1,000					
	word tokens	27,959	28,073	29,906	31,220	21,473	34,271	30,217	29,888	30,634	26,497	29,521
	word types	5,369	5,888	6,507	5,385	7,688	5,157	5,410	5,110	5,486	5,533	4,381

2 Analysis of Phrases Used in Decoding

In this section, we report a preliminary examination that shows the importance of one-to-one alignments in the standard phrase-based statistical machine translation pipeline.

2.1 Distribution

We begin with an investigation of the distribution of phrase lengths that are actually used during the decoding process. The experiments were carried out using the Europarl parallel corpus [9] in 10 language pairs.[1] For each pair, we used a training set of 347,614 sentence pairs. The development set was made up of 500 sentence pairs, and test set contained 1,000 sentence pairs. A detailed description of the data sets is given in Table 1.

Standard statistical machine translation systems were built by using the Moses toolkit [10], Minimum Error Rate Training (MERT) [11] to tune the parameters, and the SRI Language Modeling (SRILM) toolkit [12] to build a 5-gram target language model. We use GIZA++ [13] for word alignment and the maximum length of phrase pairs in phrase tables is set to 7 (the default phrase length in Moses).

We are interested in how important the performance of one-to-one alignments is in the standard system. An analysis of the distribution of phrases used during translation is shown in Figure 1. From the graph it can be seen that the majority of phrases used in the decoding process are 1-to-1 translations, i.e., word alignments. Overall, it represents more than half of the phrases. The percentage varies from 48% for sv-en to 72% for nl-en. The average over the 10 language pairs is 61%. This indicates that one-to-one alignments play a decisive role in the standard phrase-based machine translation setting.

[1]Source: Danish (da); German (de); Greek (el); Spanish (es); Finnish (fi); French (fr); Italian (it); Dutch (nl); Portuguese (pt); Swedish (sv). Target: English (en).

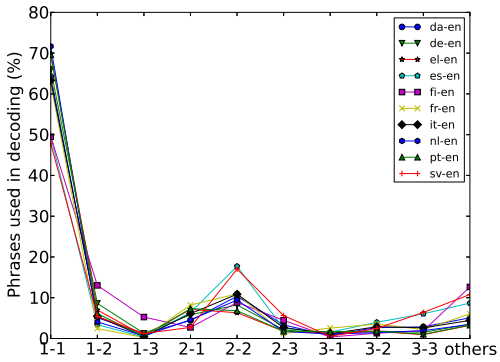

Figure 1: Distribution of phrases used during translation in standard pipeline (standard setting: Moses/GIZA++; default phrase tables: phrase length ≤ 7).

2.2 BLEU scores

We next compare the BLEU scores of the default phrase tables (length ≤ 7) and phrase tables that contain only one-to-one alignments (length = 1). The experiments of one-to-one alignments were conducted by using the same data and system components as described above (Secion 2.1) except that the phrase length in phrase tables is set to one. The parameters were tuned again for all systems. We would like to stress that the motivation of this paper is to investigate and focus on one-to-one alignments, the translation quality scores of default phrase tables are provided here as a reference.

We analyzed the contribution of one-to-one alignments to BLEU scores. This is shown in Table 2. It can be seen that the percentage varies from 46% for fi-en to 78% for nl-en. On average, one-to-one alignments contribute to around 70% of BLEU scores. We also analyzed the number of entries in phrase tables. Overall, one-to-one alignments represent only a small portion of the entries in phrase tables with an average of 0.8%.

The clear conclusion of this investigation is that, even though one-to-one alignments represent a very small number of entries, they contribute to most of the BLEU scores.

Table 2: BLEU scores and number of entries in phrase tables.

	Length ≤ 7		Length = 1		Contri. of 1-to-1	
	BLEU	Entries	BLEU	Entries	BLEU	Entries
da-en	29.39	15,940,557	21.10	192,555	71.79%	1.21%
de-en	25.03	14,322,361	16.87	73,528	67.40%	0.51%
el-en	28.11	13,944,583	20.95	61,853	74.53%	0.44%
es-en	33.53	14,964,779	24.31	99,931	72.50%	0.67%
fi-en	23.49	15,776,626	10.90	177,532	46.40%	1.13%
fr-en	32.17	14,513,041	22.20	136,524	69.01%	0.94%
it-en	30.71	14,980,498	22.38	99,870	72.88%	0.67%
nl-en	29.25	16,087,783	22.72	193,862	77.68%	1.21%
pt-en	26.30	13,699,572	20.06	48,507	76.27%	0.35%
sv-en	33.23	16,348,528	22.03	135,884	66.30%	0.83%

3 Sampling-based Alignment Method

The one-to-one alignments come mainly from the output of GIZA++. Consequently, it seems reasonable to inspect the quality of one-to-one alignments output by an alternative approach to GIZA++, i.e., the association approach. In this section, we give a brief introduction to the sampling-based alignment method [14], which is an instance of the association approach. The method is implemented in a free open-source tool called Anymalign.[2]

The sampling-based alignment model, takes as input a sentence-aligned corpus and outputs pairs of sequences of words similar to those in phrase tables, in a single step. In this method, 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 in the source and target parts, and the more likely they are to share the same distribution. Once the size of a subcorpus has been chosen, 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 table entries along with the number of times they were aligned.

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.

It has been shown in [15] that this method requires less memory in comparison with GIZA++. However, phrase tables generated by the model do not reach the performance of the state-of-the-art method on statistical machine translation tasks. In [16], a method was presented to enforce the sampling-based alignment model to align n-grams. [17] presented an alignment algorithm that relies on association scores. It complemented the sampling-based alignment model and led to results comparable to the state-of-the-art method.

4 Comparison of One-to-One Alignments

In this section, we perform a comparison of the quality of one-to-one alignments output by the sampling-based alignment method with those found in phrase tables of the standard setting GIZA++/Moses. As we are only interested in one-to-one alignments, here, the maximum length of phrase pairs in phrase tables of both models are limited to one.

In this setting, we measure the quality according to different times. Machine translation system components were kept the same in all our experiments except for the phrase tables. We compare the translation quality of the output of machine translation systems using phrase tables of Anymalign and that of GIZA++/Moses. Since Anymalign has the *anytime* feature, it can be interrupted at any moment. We chose to start with GIZA++/Moses and measure the elapsed CPU time. Then, Anymalign was run within the time range:

1. the same amount of hours (4 to 5) depending on language pairs (according to GIZA++/Moses training time) (T1);

[2] http://anymalign.limsi.fr/

2. 2 hours only (T2);

3. 10 minutes only (T3).

4.1 Evaluation Results

The experiments were carried out by using the same data (i.e., Europarl parallel corpus in 10 language pairs) and the same system components (i.e., Moses toolkit) as in Section 2. As for evaluation, four standard automatic evaluation metrics were used to assess the output of machine translation systems: BLEU [18], NIST [19], WER [20], and TER [21].

The results of experiments are shown in Table 3. From these results it can be seen that when given the same amount of training time, using phrase tables of Anymalign (T1) achieves results better or approximately equal to that of GIZA++/Moses. For language pairs de-en, el-en, and pt-en, Anymalign outperforms GIZA++/Moses significantly by 2.82, 3.62, and 3.72 BLEU points, respectively. The results for these three language pairs is consistent in terms of three other evaluation metrics. We also investigated the number of entries in the phrase tables of these two alignment models. The number of entries (one-to-one entries) produced by Anymalign is, on average, more than 10 times that of GIZA++/Moses.

By giving almost half of the amount of training time as in the previous setting, the evaluation results of Anymalign (T2) are comparable with those of Anymalign (T1). It is interesting to note that slight improvements can be observed for three language pairs (fi-en, fr-en, it-en) by comparing with Anymalign (T1). It can be seen that Anymalign (T2) achieves better results by comparing with GIZA++/Moses for all language pairs except for nl-en. Significant improvements can be seen for language pairs de-en, el-en, it-en, and pt-en, with an increase of 2.70, 3.62, 1.08, and 3.40 BLEU points, respectively. The number of entries in the phrase tables of Anymalign (T2) decreased compared with (T1) by around 33% to 47%.

In the case of Anymalign (T3) (10 minutes run), the evaluation results show that it is comparable with GIZA++/Moses. In 6 language pairs (da-en, de-en, el-en, es-en, it-en, and pt-en), Anymalign (T3) outperforms GIZA++/Moses, while in the other 4 language pairs, one observes slight decreases in evaluation matrics. Significant increases can be seen for language pairs de-en (+1.95 BLEU), el-en (+2.66 BLEU), and pt-en (+2.85 BLEU). It can be seen that Anymalign requires much less training time than GIZA++/Moses with comparable evaluation results, which is a neat advantage. The number of entries in phrase tables decreases dramatically compared with the previous two settings, however, they are still larger than those of GIZA++/Moses in all language pairs.

5 Conclusion and Future Work

In this paper, we investigated the performance of the sampling-based alignment method (an instance of associative method) on statistical machine translation tasks. We focused on one-to-one alignments output by this method and compared the results with those of the state-of-the-art estimative method in 10 European language pairs. The evaluation results show that this method achieves results better or approximately equal to those of the estimative method. Provided with the same amount of training time, the sampling-based alignment method outperforms in nine tasks over ten in the experiments of one-to-one alignments, which meets the requirement of the *translation quality* criterion. Given less time, the sampling-based model achieves comparable results in 10 minutes, which meets the requirement of the *speed* criterion.

This paper is a partial report of the work on improving the sampling-based alignment method for statistical machine translation. From the experimental results it can be seen that this method can achieve competitive results in one-to-one alignment in a very short period of training time, which is a neat advantage over the standard estimative method. Since the sampling-based model does not produce alignment points between sentences as it is in the traditional model, we plan to output these alignment points and apply heuristic phrase extraction in the standard pipeline.

Acknowledgment

Part of the research presented in this paper has been done under a Japanese grant-in-aid (Kakenhi C, 23500187: Improvement of alignments and release of multilingual syntactic patterns for statistical and example-based machine translation).

Table 3: Summary of evaluation results of **one-to-one alignments**. The column Time displays the amount of time (in seconds) for training. The column Entries presents the number of entries in phrase tables. The column Overlap analyzes how much overlap there was between phrase tables of Anymalign and those of GIZA++/Moses.

	1-1 Alignments	BLEU	NIST	WER	TER	Time (sec.)	Entries	Overlap
da-en	GIZA++/Moses	21.10	6.1894	56.32	61.24	16999	192,555	-
	Anymalign (T1)	**21.92**	**6.5064**	**55.25**	59.94	16999	1,650,032	97,761
	Anymalign (T2)	21.91	6.4829	55.28	**59.87**	7200	1,063,395	83,801
	Anymalign (T3)	21.20	6.3544	55.79	60.57	**600**	260,922	45,420
de-en	GIZA++/Moses	16.87	5.6658	60.82	66.08	17323	73,528	-
	Anymalign (T1)	**19.69**	**6.2657**	**58.12**	**62.58**	17323	1,779,535	49,518
	Anymalign (T2)	19.57	6.2506	58.32	63.08	7200	1,040,155	43,380
	Anymalign (T3)	18.82	6.0837	59.13	63.84	**600**	258,557	27,040
el-en	GIZA++/Moses	20.95	5.7535	59.30	64.01	17809	61,853	-
	Anymalign (T1)	**24.57**	**6.5854**	55.77	60.07	17809	1,900,166	38,298
	Anymalign (T2)	**24.57**	6.5343	**55.31**	**59.91**	7200	1,115,305	33,638
	Anymalign (T3)	23.61	6.3645	55.97	60.65	**600**	261,275	21,380
es-en	GIZA++/Moses	24.31	6.5124	55.47	60.27	18104	99,931	-
	Anymalign (T1)	**25.11**	**6.6134**	**54.81**	**59.35**	18104	1,775,112	65,730
	Anymalign (T2)	24.87	6.5700	55.11	59.72	7200	947,730	55,464
	Anymalign (T3)	24.43	6.5059	55.61	60.33	**600**	246,741	34,050
fi-en	GIZA++/Moses	10.90	3.7178	65.11	68.01	13800	177,532	-
	Anymalign (T1)	11.52	**4.0226**	**64.66**	68.00	13800	1,460,932	78,866
	Anymalign (T2)	**11.70**	3.9647	65.09	**67.99**	**7200**	975,339	68,534
	Anymalign (T3)	10.31	3.6752	66.99	70.12	600	220,870	34,862
fr-en	GIZA++/Moses	22.20	5.9401	62.99	66.89	19947	136,524	-
	Anymalign (T1)	22.15	6.0847	62.66	**66.37**	19947	1,830,491	78,918
	Anymalign (T2)	**22.70**	**6.1612**	**62.42**	65.93	**7200**	1,007,924	65,433
	Anymalign (T3)	22.08	6.0430	62.90	66.59	600	247,414	37,652
it-en	GIZA++/Moses	22.38	6.3151	57.09	61.46	17831	99,870	-
	Anymalign (T1)	23.22	6.4214	57.19	61.45	17831	1,917,444	65,549
	Anymalign (T2)	**23.46**	**6.4446**	**56.58**	**60.79**	7200	1,194,562	57,708
	Anymalign (T3)	22.83	6.3366	57.31	61.77	**600**	267,226	34,083
nl-en	GIZA++/Moses	**22.72**	6.2688	**57.96**	**62.54**	17950	193,862	-
	Anymalign (T1)	22.38	**6.3148**	58.52	62.86	17950	1,840,758	110,502
	Anymalign (T2)	21.84	6.2668	59.05	63.48	7200	1,156,353	94,043
	Anymalign (T3)	21.63	6.1955	59.27	63.73	600	260,567	49,109
pt-en	GIZA++/Moses	20.06	5.6701	60.09	65.13	17894	48,507	-
	Anymalign (T1)	**23.78**	**6.5374**	**55.98**	**60.63**	17894	1,787,247	28,018
	Anymalign (T2)	23.46	6.4890	56.47	61.06	7200	1,141,430	25,099
	Anymalign (T3)	22.91	6.3529	57.07	61.45	**600**	256,751	15,504
sv-en	GIZA++/Moses	22.03	6.3969	53.45	57.87	15884	135,884	-
	Anymalign (T1)	**22.39**	**6.6147**	53.05	**57.09**	15884	1,668,951	81,388
	Anymalign (T2)	22.37	6.5954	**53.00**	57.12	**7200**	1,103,929	71,803
	Anymalign (T3)	21.85	6.4789	53.72	57.78	600	261,452	40,947

References

[1] Masamichi Ideue, Kazuhide Yamamoto, Masao Utiyama, and Eiichiro Sumita. A comparison of unsupervised bilingual term extraction methods using phrase-tables. In *Proceedings of MT Summit XIII*, pages 346–351, Xiamen, China, 2011.

[2] Gregor Thurmair and Vera Aleksić. Creating term and lexicon entries from phrase tables. In *Proceedings of EAMT*, pages 253–260, Trento, Italy, 2012.

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

[4] Stephan Vogel, Hermann Ney, and Christoph Tillman. Hmm-based word alignment in statistical translation. In *Proceedings of COLING*, pages 836–841, 1996.

[5] Franz Josef Och and Hermann Ney. Improved statistical alignment models. In *Proceedings of ACL*, pages 440–447, 2000.

[6] Percy Liang, Ben Taskar, and Dan Klein. Alignment by agreement. In *Proceedings of HLT-NAACL*, pages 104–111, 2006.

[7] William A. Gale and Kenneth W. Church. Identifying word correspondences in parallel texts. In *Proceedings of the fourth DARPA workshop on Speech and Natural Language*, pages 152–157, 1991.

[8] Robert Moore. Association-based bilingual word alignment. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 1–8, Ann Arbor, 2005.

[9] Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit X*, pages 79–86, Phuket, 2005.

[10] 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 ACL*, pages 177–180, Prague, Czech Republic, 2007.

[11] Franz Josef Och. Minimum error rate training in statistical machine translation. In *Proceedings of ACL*, pages 160–167, Sapporo, Japan, 2003.

[12] Andreas Stolcke. SRILM-an extensible language modeling toolkit. In *Proceedings of ICSLP*, pages 901–904, Denver, 2002.

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

[14] Adrien Lardilleux and Yves Lepage. Sampling-based multilingual alignment. In *Proceedings of RANLP*, pages 214–218, Borovets, Bulgaria, 2009.

[15] Antonio Toral, Marc Poch, Pavel Pecina, and Gregor Thurmair. Efficiency-based evaluation of aligners for industrial applications. In *Proceedings of EAMT*, pages 57–60, 2012.

[16] Juan Luo, Adrien Lardilleux, and Yves Lepage. Improving sampling-based alignment by investigating the distribution of n-grams in phrase translation tables. In *Proceedings of PACLIC*, pages 150–159, 2011.

[17] Adrien Lardilleux, Francois Yvon, and Yves Lepage. Hierarchical sub-sentential alignment with anymalign. In *Proceedings of EAMT*, pages 279–286, Trento, 2012.

[18] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318, Philadelphia, 2002.

[19] George Doddington. Automatic evaluation of machine translation quality using N-gram co-occurrence statistics. In *Proceedings of HLT*, pages 138–145, 2002.

[20] Sonja Nießen, Franz Josef Och, Gregor Leusch, and Hermann Ney. An evaluation tool for machine translation: Fast evaluation for machine translation research. In *Proceedings of LREC*, pages 39–45, Athens, 2000.

[21] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of AMTA*, pages 223–231, Cambridge, Massachusetts, 2006.