Improving Sampling-based Alignment by Investigating the Distribution of N-grams in Phrase Translation Tables ⋆

Juan Luo[a], Adrien Lardilleux[b], and Yves Lepage[a]

[a] Graduate School of Information, Production and Systems, Waseda University,
2-7 Hibikino, Wakamatsu-ku, Kitakyushu, Fukuoka 808-0135, Japan
{juanluoonly@suou, yves.lepage@aoni}.waseda.jp
[b] TLP Group, LIMSI-CNRS,
BP 133, Orsay Cedex 91403, France
adrien.lardilleux@limsi.fr

Abstract. This paper describes 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. The quality of phrase translation tables output by this approach and that of MGIZA++ is compared in statistical machine translation tasks. Significant improvements for this approach are reported. In addition, merging translation tables is shown to outperform state-of-the-art techniques.

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

1 Introduction

Phrase translation tables play an important role in the process of building machine translation systems. The quality of translation table, which identifies the relations between words or phrases in the source language and those in the target language, is crucial for the quality of the output of most machine translation systems. Currently, the most widely used state-of-the-art tool to generate phrase translation tables is GIZA++ (Och and Ney, 2003), which trains the ubiquitous IBM models (Brown *et al.*, 1993) and the HMM introduced by (Vogel *et al.*, 1996), in combination with the Moses toolkit (Koehn *et al.*, 2007). MGIZA++, a multi-threaded word aligner based on GIZA++, is proposed by (Gao and Vogel, 2008).

In this paper, we investigate a different approach to the production of phrase translation tables: the sampling-based approach (Lardilleux and Lepage, 2009b). This approach is implemented in a free open-source tool called Anymalign.[1] Being in line with the associative alignment trend illustrated by (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 illustrated by (Brown *et al.*, 1991; Och and Ney, 2003; Liang *et al.*, 2006). In addition, it is capable of aligning multiple languages simultaneously; but we will not use this feature here as we will restrain ourselves to bilingual experiments in this paper.

In sampling-based alignment, only those sequences of words sharing the exact same distribution (i.e., they appear exactly in the same sentences of the corpus) are considered for alignment.

⋆ 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).

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

The key idea is to make more words share the same distribution by artificially reducing their frequency in multiple random subcorpora obtained by sampling. Indeed, the smaller a subcorpus, the less frequent its words, and the more likely they are to share the same distribution; hence the higher the proportion of words aligned in this subcorpus. In practice, the majority of these words turn out to be *hapaxes*, that is, words that occur only once in the input corpus. Hapaxes have been shown to safely align across languages (Lardilleux and Lepage, 2009a).

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

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

where k denotes the size (number of sentences) of a subcorpus and n the size of the complete input corpus. Note that this function is very close to $1/k^2$: it gives much more credit to small subcorpora, which happen to be the most productive (Lardilleux and Lepage, 2009b). 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 implemented with an *anytime* algorithm: 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, after a very short amount of time, *quality* is no more 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, as they are calculated on the basis of the number of time each alignment was obtained. This is possible because high frequency alignments are quickly output with a fairly good estimation of their translation probabilities. As time goes, their estimation is refined, while less frequent alignments are output in addition.

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 describes a preliminary experiment on the sampling-based alignment approach implemented in Anymalign baseline and provides the experimental results from which the problem is defined. In Section 3, we propose a variant in order to improve its performance on statistical machine translation tasks. Section 4 introduces standard normal distribution of time to bias the distribution of n-grams in phrase translation tables. Section 5 describes the effects of pruning on the translation quality. Section 6 presents the merge of two aligners' phrase translation tables. Finally, in Section 7, conclusions and possible directions for future work are presented.

[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. The training corpus 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 (Koehn *et al.*, 2007), MERT (Minimum Error Rate Training) to tune the parameters of translation tables (Och, 2003), and the SRI Language Modeling toolkit (Stolcke, 2002) to build the target language model.

As for the evaluation of translations, four standard automatic evaluation metrics were used: mWER (Nießen *et al.*, 2000), BLEU (Papineni *et al.*, 2002), NIST (Doddington, 2002), and TER (Snover *et al.*, 2006).

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

		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

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.

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

	mWER	BLEU	NIST	TER
MGIZA++	**0.5714**	**0.2742**	**6.6747**	**0.6170**
Anymalign	0.6186	0.2285	6.0764	0.6634

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 7 (a) and Table 7 (b). 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 × m n-grams), the number of alignments in Anymalign table is more than 10 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 (Lardilleux *et al.*, 2009).

3 Anymalign1-N

3.1 Enforcing Alignment of N-grams

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.

Similar works on the idea of enlarging n-grams have been reported in (Ma *et al.*, 2007), in which "word packing" is used to obtain 1-to-*n* alignments based on co-occurrence frequencies, and (Henríquez Q. *et al.*, 2010), in which collocation segmentation is performed on bilingual corpus to extract *n*-to-*m* alignments.

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.

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_.

3.2 Phrase Translation Subtables

It it 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 an *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, Anymalign is thus run N × N times. 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.

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

		Target				
		unigrams	bigrams	trigrams	···	N-grams
Source	unigrams	TT 1 × 1	TT 1 × 2	TT 1 × 3	···	TT 1 × N
	bigrams	TT 2 × 1	TT 2 × 2	TT 2 × 3	···	TT 2 × N
	trigrams	TT 3 × 1	TT 3 × 2	TT 3 × 3	···	TT 3 × N
	···	···	···	···	···	···
	N-grams	TT N × 1	TT N × 2	TT N × 3	···	TT N × N

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 alignment 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.

3.3 Evaluation Results with 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, which is, uniformly distributed time among subtables. For a fair comparison, the same amount of time was given: seven hours in total. The results are shown in Table 6. On the whole, MGIZA++ significantly outperforms Anymalign, by more than 4 BLEU points. The proposed approach, Anymalign1-N, produces better results than Anymalign in its basic version, with the best increase with Anymalign1-3 or Anymalign1-4 (+1.3 BP).

The comparison of Table 7 (c) (see last page) and Table 7 (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.

4 Time Distribution among Subtables

In order to increase the number of phrase pairs along the diagonal of the translation table matrix and decrease this number outside the diagonal (Table 4), we distribute the total alignment time among translation subtables according to the standard normal distribution:

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

The alignment time allotted to the subtable between source n-grams and target m-grams will thus be proportional to $\phi(n,m)$. Table 5 shows an example of alignment times allotted to each subtable up to 4-grams, for a total processing time of 7 hours.

4.1 Evaluation Results with Standard Normal Time Distribution

We performed a third evaluation using the standard normal distribution of time, as in previous experiments, again with a total amount of processing time (7 hours).

The comparison between MGIZA++, Anymalign in its standard use, and Anymalign1-N with standard normal time distribution is shown in Table 6. Anymalign1-4 shows the best performance

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

Table 5: Alignment time in seconds allotted to each unigramized parallel corpus of Anymalign1-4. The sum of the figures in all cells amounts to seven hours (25,200 seconds).

		Target			
		unigrams	bigrams	trigrams	4-grams
Source	unigrams	3,072	1,863	416	34
	bigrams	1,863	3,072	1,863	416
	trigrams	416	1,863	3,072	1,863
	4-grams	34	416	1,863	3,072

in terms of mWER and BLEU scores, while Anymalign1-3 gets the best results for the two other evaluation metrics. There is an increase in BLEU scores for almost all Anymalign1-N, from Anymalign1-3 to Anymalign1-10, when compared with the translation qualities of Anymalign1-N with equal time distribution. The greatest increase in BLEU is obtained for Anymalign1-10 (almost +2 BP). Anymalign1-4 shows the best translation qualities 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 7 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.

5 Translation Table Pruning

Until now, we were concerned with the shape of phrase translation tables in standard configurations. However, (Johnson *et al.*, 2007) have shown that substantially pruning the phrase translation tables can lead to slight but consistent improvements in translation quality.

They use Fisher's exact significance test to eliminate a substantial number of phrase pairs. The significance of the association between a (source, target) phrase pair is evaluated and their probability of co-occurrence in the corpus is calculated. The hypergeometric distribution is used to compute the observed probability of joint occurrence $C(\tilde{s},\tilde{t})$, with $\tilde{s}$ a source phrase and $\tilde{t}$ a target phrase:

$$p_h(C(\tilde{s},\tilde{t})) = \frac{\binom{C(\tilde{s})}{C(\tilde{s},\tilde{t})}\binom{N-C(\tilde{s})}{C(\tilde{t})-C(\tilde{s},\tilde{t})}}{\binom{N}{C(\tilde{t})}} \tag{3}$$

Here, N is the number of sentences in the input parallel corpus. The p-value is calculated as:

$$\text{p-value}(C(\tilde{s},\tilde{t})) = \sum_{k=C(\tilde{s},\tilde{t})}^{\infty} p_h(k) \tag{4}$$

Any phrase pair with a p-value greater than a given threshold will thus be filtered out. In practice, this mainly removes phrase pairs with different frequencies. A special case happens when a source phrase and a target phrase, hence the resulting phrase pair as well, occur only once in the corpus (called a 1-1-1 phrase pair in (Johnson *et al.*, 2007)). By considering a p-value of

$\alpha = \log(N)$, $\alpha + \varepsilon$ (where ε is very small) is the smallest threshold that results in none of the 1-1-1 phrase pairs being included, while $\alpha - \varepsilon$ is the largest threshold that results in those pairs being included.

We investigate the impact of pruning on Anymalign's translation tables in terms of n-gram distribution and final translation quality.

5.1 Evaluation Results with Pruning

In a fourth set of experiments, we thus compare the phrase translation tables of MGIZA++, and Anymalign1-N (standard normal time distribution), after applying this pruning. The $\alpha - \varepsilon$ filter was used.

Evaluation results on machine translation tasks with pruned translation tables are given in Table 6. The phrase table size reduction brings gains in BLEU scores. Among all Anymalign1-N, Anymalign1-4 once again gets the highest BLEU score of 0.2511 and shows the best performance in all evaluation metrics.

As an example, the number of entries in Anymalign1-4's translation table, after pruning, is shown in Table 7 (e). The largest difference when compared with the non-pruned translation table (Table 7 (d)) is visible in the cell corresponding to 1-1 entries: a substantial decrease of almost 200,000 entries is observed, which corresponds to a reduction of 76%. As a consequence, the most numerous entries are now 2-2 phrase pairs, which account for 19% of the total number of phrase pairs. On the whole, 54% of entries were filtered out from Anymalign1-4's translation table.

6 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 and we used the union of the two translation tables. As for the feature scores in the translation tables for the intersection part of both aligners, i.e., entries in two translation tables share the same phrase pairs but with different feature scores, we adopted parameters computed either by MGIZA++ or by Anymalign for evaluation.

Evaluation results on machine translation tasks with merged translation tables are given in Table 6. 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.

7 Conclusions and Future Work

We have presented a method to improve the translation quality of the sampling-based subsentential alignment approach for statistical machine translation tasks. Our approach is based on adapting the number of n-grams by investigating their distribution in phrase translation tables. Furthermore, we inspected the influence of pruning the translation tables, a technique described in (Johnson *et al.*, 2007), and merging the translation tables from two aligners (i.e., Anymalign and MGIZA++). Adapting the number of n-grams leads to significantly better evaluation results than the original approach. Merging two translation tables outperforms MGIZA++ alone. As for future work, we plan to modify the computation of the feature scores in Anymalign's phrase translation tables to make them closer to those of MGIZA++.

Table 6: Evaluation results.

	mWER	BLEU	NIST	TER
MGIZA++	**0.5714**	**0.2742**	**6.6747**	**0.6170**
Anymalign	0.6186	0.2285	6.0764	0.6634

Anymalign1-N	equal time distribution				std.norm.distribution				pruning			
	mWER	BLEU	NIST	TER	mWER	BLEU	NIST	TER	mWER	BLEU	NIST	TER
Anymalign1-1	0.6818	0.1984	5.6353	0.7188	0.6818	0.1984	5.6353	0.7188	0.6871	0.1953	5.6042	0.7258
Anymalign1-2	0.6121	0.2406	6.2789	0.6536	0.6121	0.2404	6.2674	0.6535	0.6102	0.2425	6.3093	0.6515
Anymalign1-3	**0.6075**	0.2403	**6.3009**	**0.6507**	0.6079	0.2441	**6.2928**	**0.6517**	0.6117	0.2413	6.2501	0.6561
Anymalign1-4	0.6142	**0.2423**	6.2087	0.6583	**0.6071**	**0.2442**	6.2844	0.6526	**0.5978**	**0.2511**	**6.3985**	**0.6435**
Anymalign1-5	0.6099	0.2376	6.2331	0.6551	0.6134	0.2436	6.2426	0.6548	0.6076	0.2457	6.3120	0.6504
Anymalign1-6	0.6193	0.2349	6.1574	0.6634	0.6165	0.2403	6.1595	0.6589	0.6104	0.2459	6.2687	0.6545
Anymalign1-7	0.6157	0.2371	6.2107	0.6559	0.6136	0.2405	6.2124	0.6564	0.6079	0.2419	6.2569	0.6516
Anymalign1-8	0.6353	0.2253	5.9777	0.6794	0.6151	0.2366	6.1639	0.6597	0.6060	0.2446	6.2986	0.6496
Anymalign1-9	0.6279	0.2296	6.0261	0.6722	0.6136	0.2402	6.1928	0.6564	0.6078	0.2461	6.2974	0.6493
Anymalign1-10	0.6475	0.2182	5.8534	0.6886	0.6192	0.2361	6.1803	0.6587	0.6076	0.2459	6.3079	0.6490

Merge	mWER	BLEU	NIST	TER
Anymalign param.	**0.5671**	**0.2747**	**6.7101**	**0.6128**
MGIZA++ param.	**0.5685**	**0.2754**	**6.7060**	**0.6142**

References

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

Brown, Peter, Jennifer Lai, and Robert Mercer. 1991. Aligning sentences in parallel corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL'91)*, pp. 169–176, Berkeley (Californie, USA), jun.

Doddington, George. 2002. Automatic evaluation of machine translation quality using N-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, pp. 138–145, San Diego. Morgan Kaufmann Publishers Inc.

Gale, William and Kenneth Church. 1991. Identifying word correspondences in parallel texts. In *Proceedings of the fourth DARPA workshop on Speech and Natural Language*, pp. 152–157, Pacific Grove, feb.

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

Henríquez Q., A. Carlos, R. Marta Costa-jussà, Vidas Daudaravicius, E. Rafael Banchs, and B. José Mariño. 2010. Using collocation segmentation to augment the phrase table. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, WMT '10, pp. 98–102, Stroudsburg, PA, USA. Association for Computational Linguistics.

Johnson, J Howard, Joel Martin, George Foster, and Roland Kuhn. 2007. Improving translation quality by discarding most of the phrasetable. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 967–975, Prague, Czech Republic.

Koehn, Philipp, 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. 2007. Moses: Open source toolkit for statistical

machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, pp. 177–180, Prague, Czech Republic.

Koehn, Philipp, Franz J. Och, and Daniel Marcu. 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*, pp. 48–54, Edmonton.

Lardilleux, Adrien, Jonathan Chevelu, Yves Lepage, Ghislain Putois, and Julien Gosme. 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*, pp. 45–52, Dublin, Ireland.

Lardilleux, Adrien and Yves Lepage. 2009a. Hapax Legomena : their Contribution in Number and Efficiency to Word Alignment. *Lecture notes in computer science*, 5603, 440–450.

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

Liang, Percy, Ben Taskar, and Dan Klein. 2006. Alignment by agreement. In *Proceedings of the Human Language Technology Conference of the NAACL*, pp. 104–111, New York City, jun.

Ma, Yanjun, Nicolas Stroppa, and Andy Way. 2007. Bootstrapping word alignment via word packing. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 304–311, Prague, Czech Republic.

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

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

Nießen, Sonja, Franz Josef Och, Gregor Leusch, and Hermann Ney. 2000. An evaluation tool for machine translation: Fast evaluation for machine translation research. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC)*, pp. 39–45, Athens.

Och, Franz Josef. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pp. 160–167, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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)*, pp. 311–318, Philadelphia.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas (AMTA 2006)*, pp. 223–231, Cambridge, Massachusetts.

Stolcke, A. 2002. SRILM-an extensible language modeling toolkit. In *Seventh International Conference on Spoken Language Processing*, volume 2, pp. 901–904, Denver, Colorado.

Vogel, Stephan, Hermann Ney, and Christoph Tillman. 1996. Hmm-based word alignment in statistical translation. In *Proceedings of the 16th International Conference on Computational Linguistics (Coling'96)*, pp. 836–841, Copenhagen, aug.

Table 7: Distribution of phrase pairs in translation tables.

(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

(e) Distribution of phrase pairs in Anymalign1-4's translation table (after pruning).

		Target							
		unigrams	bigrams	trigrams	4-grams	5-grams	6-grams	7-grams	total
Source	unigrams	**60,297**	59,099	8,819	328	0	0	0	128,543
	bigrams	58,232	**110,415**	51,557	6,954	0	0	0	227,158
	trigrams	9,777	58,604	**69,431**	28,046	0	0	0	165,858
	4-grams	474	8,586	31,209	**31,666**	0	0	0	71,935
	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	128,780	236,704	161,016	66,994	0	0	0	593,494