Production of Phrase Tables in 11 European Languages using an Improved Sub-sentential Aligner

Juan Luo, Yves Lepage

IPS, Waseda University
2-7 Hibikino, Wakamatsu-ku, Kitakyushu, Fukuoka 808-0135, Japan
juan.luo@suou.waseda.jp, yves.lepage@waseda.jp

Abstract

This paper is a partial report of an on-going Kakenhi project which aims to improve sub-sentential alignment and release multilingual syntactic patterns for statistical and example-based machine translation. Here we focus on improving a sub-sentential aligner which is an instance of the association approach. Phrase table is not only an essential component in the machine translation systems but also an important resource for research and usage in other domains. As part of this project, all phrase tables produced in the experiments will also be made freely available.

Keywords: phrase table, sub-sentential alignment, machine translation

1. Introduction

Sub-sentential alignment is an important step in the process of building machine translation systems. Given a parallel corpus, sub-sentential 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 both statistical and example-based machine translation systems. 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.

Many researchers have investigated the issue of sub-sentential alignment. One of the earliest and widest used alignment approaches is the *estimation approach*. It employs statistical models and the parameters are estimated through maximization process. It is based on IBM models (Brown et al., 1993). Many studies are carried out in this trend (Vogel et al., 1996; Och and Ney, 2000; Liang et al., 2006; Neubig et al., 2011; Dyer et al., 2013). Another approach to alignment is the *association approach*. It utilize different similarity measures and association tests, for example, mutual information (Gale and Church, 1991), t-scores (Ahrenberg et al., 1998), and log-likelihood-ratio association measure (Moore, 2005).

In this paper, we focus on improving phrase distribution in phrase table produced by a sub-sentential aligner: Anymalign[1] (Lardilleux and Lepage, 2009). It is an instance of the association approach and implements the sampling-based alignment method. The sampling-based alignment method, 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 (i.e., those words sharing the same distribution over a set of source-target sentence pairs) are considered for alignment.

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

Sub-sentential alignment produces phrase tables. As a stand-alone application, sub-sentential alignment and phrase table are used in other domains, for instance, bilingual terminology extraction (Itagaki et al., 2007; Morishita et al., 2008; Ideue et al., 2011), and creation of lexicon entries (Lardilleux et al., 2010; Thurmair and Aleksić, 2012). Itagaki et al. (2007) proposed a method to extract bilingual terminologies and validate their quality. They firstly extract term pairs from phrase table. The quality of extracted term pairs are then validated by using a Gaussian mixture model classifier. Morishita et al. (2008) proposed a semi-automatic method of acquiring technical terms from parallel patent documents by combining a phrase table and a bilingual lexicon. Support vector machines is then applied to validate phrase pairs in the phrase table. In (Ideue et al., 2011), three statistical measures for extracting bilingual terminologies from a phrase table are compared. They showed that a combination of these three measures ranks valid bilingual terms highly. Lardilleux et al. (2010) presented a protocol to evaluate three word aligners. They then select the most appropriate one to produce bilingual lexicons. In (Thurmair and Aleksić, 2012), a tool was described for extracting terms and lexicons from phrase tables. The term candidates in phrase tables are filtered on several levels to identify "good" terms.

Phrase table is not only a vital component in the machine translation systems, but also an important resource for research and usage in other domains. Therefore, we will release phrase tables produced with different approaches in various experimental settings presented in this paper.

The objectives of this part of the work are:

- improving phrase distribution in phrase tables produced by Anymalign;
- producing and releasing phrase tables.

The remainder of this paper is organized as follows. Section 2. details the proposed method. In Section 3., merging and pruning phrase tables produced by two aligners are presented. Section 4. describes phrase table resources. We conclude in Section 5. with future works.

2. Alignment of n-grams

It has been shown in (Lardilleux et al., 2009) that Anymalign (the sampling-based alignment method) excels in bilingual lexicon induction, i.e., one-to-one alignment. However, it does not align enough long n-grams, which makes it less competitive in phrase-based machine translation tasks. One important feature of this 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. A detailed description of the sampling-based alignment algorithm is given in (Lardilleux et al., 2009). The translation probabilities in the sampling-based alignment method are calculated as proposed in (Koehn et al., 2003), however, there is a slight difference in the counts of the phrase pairs. In this method, the counts for the phrase pair are collected from all sampled subcorpora. Therefore, in some cases, the frequency count would be larger than the one collected from the whole corpus.

In this section, building on the strengths of this alignment method and making use of the *anytime* feature and the possibility of allotting time freely, we propose a method to force the sampling-based alignment method to align more of the n-grams of the longer kind. It is presented in the following sections.

2.1. Enforcing alignment of 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 1 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.

2.2. Phrase translation subtables

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 an *unigramized n-m corpus*. These corpora are then used as input to Anymalign to produce phrase translation subtables, as shown in Table 2. Practically, we call *Anymalign1-N* the process

N-grams	Sentences
1-gram	resumption of the session
2-gram	resumption_of of_the the_session
3-gram	resumption_of_the of_the_session
...	...

Table 1: Transforming n-grams into unigrams by inserting underscores and reduplicating words for one part of the input parallel corpus. The same procedure is applied for the other part of the parallel corpus.

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 $\times$ N times. All phrase translation subtables are finally merged together into one large phrase table, where translation probabilities are re-estimated given the complete set of alignments.

		Target				
		1-grams	2-grams	3-grams	$\cdots$	N-grams
Source	1-grams	$\text{PT}_{1\times1}$	$\text{PT}_{1\times2}$	$\text{PT}_{1\times3}$	$\cdots$	$\text{PT}_{1\times N}$
	2-grams	$\text{PT}_{2\times1}$	$\text{PT}_{2\times2}$	$\text{PT}_{2\times3}$	$\cdots$	$\text{PT}_{2\times N}$
	3-grams	$\text{PT}_{3\times1}$	$\text{PT}_{3\times2}$	$\text{PT}_{3\times3}$	$\cdots$	$\text{PT}_{3\times N}$
	...	...	...	...	...	...
	N-grams	$\text{PT}_{N\times1}$	$\text{PT}_{N\times2}$	$\text{PT}_{N\times3}$	$\cdots$	$\text{PT}_{N\times N}$

Table 2: List of n-gram phrase translation subtables (PT) generated from the training corpus. These subtables will then be merged together into a single phrase table.

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 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.

2.3. Standard normal time distribution

By examing the distribution matrix of MGIZA++'s phrase table, one would observe that the majority of the alignments is along the diagonal. Therefore, in order to increase the number of phrase pairs along the diagonal of the Anymalign's phrase table distribution matrix and decrease this number outside the diagonal (Table 2), we distribute the total alignment time among translation subtables proportionally to the standard normal distribution:

$$\text{time}(\text{PT}_{n\times m}) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(n-m)^2} \quad (1)$$

This distribution attempts to distribute more phrase pairs along the diagonal and less when departing from it. Table 3 shows an example of alignment times allotted to each subtable up to 4-grams, for a total processing time of 12 hours. The number of phrase pairs produced will depend upon the amount of time.

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

		Target			
		1-grams	2-grams	3-grams	4-grams
Source	1-grams	5267	3194	713	59
	2-grams	3194	5267	3194	713
	3-grams	713	3194	5267	3194
	4-grams	59	713	3194	5267

Table 3: Alignment time in seconds allotted to each unigramized parallel corpus of Anymalign1-4. The sum of the figures in all cells amounts to twelve hours (12 hrs = 43,200 seconds).

2.4. Experiments

Standard statistical machine translation systems were built by using the Moses toolkit (Koehn et al., 2007), Minimum Error Rate Training (Och, 2003), and the SRI Language Modeling toolkit (Stolcke, 2002). We built systems for 11 European languages[3]. For each language pair, the training corpus is made of 347,614 sentences from the Europarl parallel corpus release v3 (Koehn, 2005). The development set contains 500 sentences, and 38,123 sentences were used for testing. Here we used the common part of the Europarl corpus, so that all sentences are translations of one another across 11 languages.

We compared two settings: MGIZA++ (Gao and Vogel, 2008) and the proposed method. The evaluation results are given in Table 6 and Table 7. On the whole, MGIZA++ outperforms Anymalign1-4. We also investigated the number of entries in phrase tables. It is shown in Figure 1. From the graph it can be seen that there are more longer n-grams in phrase table of MGIZA++ while the majority of phrases (more than 80%) in Anymalign baseline are unigrams. In the phrase table of Anymalign1-4, we can see a significant increase in the number of longer n-grams by comparing with Anymalign baseline.

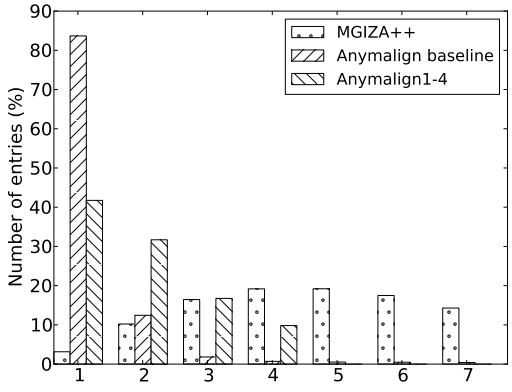

Figure 1: Distribution of n-grams in phrase tables (French-English).

3. Merging and pruning phrase tables

In order to check exactly how different the phrase tables of MGIZA++ and those of Anymalign are, we performed experiments in which MGIZA++'s phrase tables is simply merged with those of Anymalign baseline. Here we used the union of the two phrase tables. As for the feature scores (i.e., translation probabilites and lexical weights) in the phrase tables for the intersection part of both aligners, i.e., phrase pairs in both phrase tables, we adopted the parameters computed by MGIZA++ for evaluation.

In addition, we applied the technique of pruning presented in (Johnson et al., 2007). In this work, they showed that a substantial number of phrase pairs can be eliminated without sacrificing the translation quality. We investigated the impact of pruning[4] on merged phrase tables in terms of final translation quality. On average, 49.73% of phrase pairs in the phrase tables were discarded.

The evaluation results of merging and pruning phrase tables are shown in Table 8 and Table 9, respectively. The phrase table size reduction by pruning brings gains in BLEU scores. We analyzed how much overlap there was between phrase tables of Anymalign and those of MGIZA++. This is shown in Table 4 and Table 5. From the tables it can be seen that the two alignment approaches produce different phrases. For the overlap portion, Figure 2 shows that there is not much difference in the translation probabilities produced by Anymalign and MIGIZA++.

	Entries	Overlap
MGIZA++	13,214,402	244,224
Anymalign	3,137,641	244,224

Table 4: Overlap between phrase table of MGIZA++ and that of Anymalign (French-English).

		Target						
		1-g	2-g	3-g	4-g	5-g	6-g	7-g
Source	1-g	61.16	16.00	3.55	0.84	0.49	0.24	0.00
	2-g	12.76	3.65	0.64	0.28	0.32	0.64	1.16
	3-g	2.89	0.46	0.38	0.20	0.23	0.37	0.63
	4-g	1.05	0.22	0.13	0.20	0.25	0.36	0.61
	5-g	0.50	0.17	0.13	0.15	0.22	0.37	0.60
	6-g	0.45	0.23	0.15	0.19	0.23	0.30	0.42
	7-g	0.82	0.43	0.24	0.24	0.26	0.31	0.39

Table 5: Overlap to MGIZA++ phrase table in percentage (cell-by-cell) (French-English).

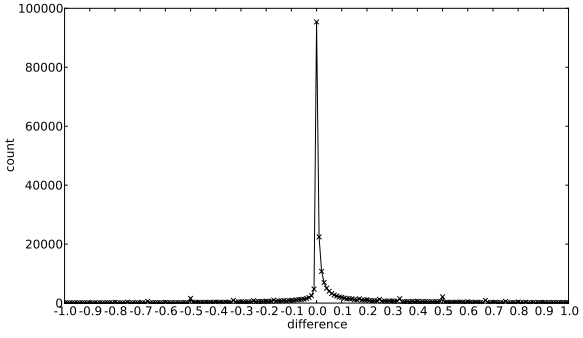

Figure 2: Difference of translation probability $p(en|fr)$, avg.±stddev.: 0.037±0.239 (French-English).

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

[4]Option `-l a-e -n 30` in the program.

Source language	Target language: da	de	el	en	es	fi	fr	it	nl	pt	sv
da	-	19.97	21.90	31.68	27.61	14.04	23.84	20.72	23.05	25.22	29.71
de	24.06	-	20.80	27.08	25.89	12.02	20.82	18.85	25.33	23.47	21.06
el	24.50	18.85	-	33.39	32.88	13.45	25.89	24.99	22.40	29.96	22.95
en	29.03	19.90	27.47	-	34.13	14.61	29.43	25.62	24.73	30.60	28.60
es	25.52	19.10	27.64	34.70	-	13.07	30.15	28.89	23.35	35.61	24.07
fi	20.62	15.46	18.28	23.74	22.11	-	17.47	16.46	17.80	20.06	19.01
fr	21.04	16.39	23.26	29.36	32.95	10.47	-	25.14	20.95	30.33	19.67
it	22.19	16.97	24.97	30.97	34.44	11.37	28.26	-	21.59	31.10	20.66
nl	24.01	21.44	20.58	28.24	25.88	11.39	20.89	19.70	-	23.62	21.41
pt	24.54	18.57	26.44	33.32	38.12	12.52	30.25	28.46	22.72	-	22.73
sv	32.35	19.88	23.26	34.03	29.11	14.74	23.12	21.83	23.20	26.32	-

Table 6: BLEU points for MGIZA++

Source language	Target language: da	de	el	en	es	fi	fr	it	nl	pt	sv	avg.
da	-	−2.46	−3.76	−3.84	−8.09	−2.64	−4.11	−3.05	−2.32	−4.48	−1.18	−3.59
de	−2.95	-	−3.21	−3.56	−3.53	−1.79	−1.46	−2.19	−2.12	−3.31	−2.55	−2.66
el	−4.39	−2.64	-	−4.63	−3.75	−2.83	−2.16	−2.56	−2.68	−3.70	−3.74	−3.30
en	−3.42	−2.81	−4.02	-	−3.92	−3.23	−4.29	−3.59	−2.73	−4.05	−3.50	−3.55
es	−4.53	−3.17	−3.45	−4.10	-	−2.51	+1.19	−0.71	−2.80	−0.99	−3.99	−2.50
fi	−3.67	−2.94	−3.91	−7.35	−4.69	-	−2.02	−3.01	−3.44	−4.34	−3.48	−3.88
fr	−3.65	−2.67	−4.28	−4.63	−4.52	−1.93	-	−1.44	−3.73	−4.02	−3.54	−3.44
it	−4.20	−3.04	−4.60	−5.59	−4.16	−2.74	−0.15	-	−3.89	−3.01	−3.95	−3.53
nl	−2.80	−2.38	−3.31	−3.44	−3.20	−1.97	−1.77	−2.71	-	−3.35	−2.36	−2.72
pt	−4.39	−2.85	−3.62	−4.42	−1.85	−2.38	−0.52	−0.80	−3.25	-	−4.33	−2.84
sv	−1.39	−2.69	−4.19	−4.31	−5.05	−2.87	−1.55	−3.46	−2.54	−4.71	-	−3.27
avg.	−3.53	−2.76	−3.83	−4.58	−4.27	−2.48	−1.68	−2.35	−2.95	−3.59	−3.26	**−3.52**

Table 7: Differences in BLEU points for Anymalign1-4 compared with MGIZA++

Source language	Target language: da	de	el	en	es	fi	fr	it	nl	pt	sv	avg.
da	-	−0.15	−0.27	−0.05	−0.36	+0.05	−2.08	−0.30	−0.10	−0.37	+0.08	−0.35
de	−0.56	-	−0.34	−0.32	−0.37	+0.41	−0.34	0.00	−0.15	−0.14	−0.07	−0.18
el	−0.21	−0.13	-	−0.23	−0.07	+0.10	−0.47	−0.12	+0.06	−0.30	+0.19	−0.11
en	−0.20	−0.38	−0.76	-	−0.33	−0.01	−2.75	−0.06	−0.15	−0.21	−0.18	−0.50
es	−0.30	−0.10	+0.10	+0.01	-	+0.11	−0.20	−0.06	−0.39	+0.02	−0.47	−0.12
fi	−0.21	−0.06	−0.03	−0.26	−0.03	-	+0.04	−0.10	−0.21	−0.12	−0.08	−0.10
fr	−0.13	+0.20	−0.20	+0.03	−0.09	+0.21	-	+0.10	−0.08	−0.99	−0.32	−0.12
it	−0.22	−0.28	−0.12	−0.10	−0.19	+0.38	−0.09	-	−0.20	+0.23	−0.29	−0.08
nl	−0.32	+0.08	−0.26	−0.17	−0.08	+0.41	−0.24	−0.32	-	−0.53	−0.26	−0.16
pt	−0.28	−0.12	−0.22	−0.06	−0.08	−0.15	+7.23	+3.64	+0.16	-	−0.52	+0.96
sv	+0.12	−0.28	−0.18	+0.48	−0.25	+0.21	+6.18	+2.58	+0.05	−0.41	-	+0.85
avg.	−0.23	−0.12	−0.22	−0.06	−0.18	+0.17	+0.72	+0.53	−0.10	−0.28	−0.19	**0.00**

Table 8: Differences in BLEU points for merged phrase table compared with MGIZA++

Source language	Target language: da	de	el	en	es	fi	fr	it	nl	pt	sv	avg.
da	-	+0.16	−0.01	+0.19	+0.11	+0.05	−1.65	−0.04	+0.22	+0.07	+0.48	−0.04
de	+0.42	-	+0.17	+0.20	+0.01	+0.42	+6.27	+3.78	+0.34	+0.08	+0.52	+1.22
el	+0.05	+0.19	-	−0.12	+0.13	−0.03	+0.12	+0.09	+0.15	−0.14	+0.37	+0.08
en	+0.20	+0.13	−0.11	-	+0.08	−0.06	−2.47	+0.03	+0.32	+0.27	+0.02	−0.15
es	−0.18	+0.27	+0.03	−0.02	-	+0.24	+0.10	+0.37	−0.14	+0.80	+0.19	+0.16
fi	−0.19	−0.02	+0.03	−0.24	−0.20	-	+0.11	+0.38	−0.01	−0.12	+0.07	−0.01
fr	+0.16	+0.44	−0.09	+0.37	+0.31	+0.34	-	+0.25	+0.02	+0.20	+0.13	+0.21
it	−0.23	+0.40	−0.08	−0.01	+0.29	+0.40	+0.23	-	−0.26	+0.23	+0.16	+0.11
nl	+0.12	+0.28	+0.11	−0.05	+0.02	+0.32	+0.24	−0.10	-	−0.15	+0.30	+0.10
pt	+0.03	+0.33	+0.07	−0.11	+0.15	+0.15	+0.13	+0.03	−0.25	-	+0.31	+0.08
sv	+0.52	−0.02	−0.11	+0.21	−0.12	+0.13	−0.16	−0.05	+0.28	−0.26	-	+0.04
avg.	+0.09	+0.21	0.00	+0.04	+0.07	+0.19	+0.29	+0.47	+0.06	+0.09	+0.25	**+0.18**

Table 9: Differences in BLEU points for merged pruned phrase table compared with MGIZA++

4. Phrase table resources

As it is mentioned in Section 1. phrase tables are not only an essential part of machine translation systems, but also an important resource used in other domains, e.g., bilingual terminology extraction, creation of bilingual lexicon entries. It is interesting to note that phrase pairs in phrase tables produced by Anymalign and those of MGIZA++ are different. This deserves further analysis and research. Phrase tables produced in all experimental settings will be released in the near future. More phrase tables will also be released as the project goes on.

5. Conclusion and future work

In this paper, we have described a method to increase the number of longer n-grams in phrase tables produced by an instance of the associative alignment approach: Anymalign. We also presented merging and pruning of two phrase tables, one from MGIZA++ and the other from Anymalign. An analysis of overlap between phrase tables of two aligners shows that they produce different phrase pairs. Further investigation on how they can complement each other will be conducted. As phrase table is an important resource for research and usage in various fields, and as part of the plan of the project, all phrase tables are made freely available at the URL: `http://133.9.48.109/`.

6. Acknowledgements

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).

7. References

Ahrenberg, Lars, Merkel, Magnus, and Anderson, Mikael. (1998). A simple hybrid aligner for generating lexical correspondences in parallel texts. In *Proceedings of ACL-COLING*, pages 29–35, Montreal, Canada.

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

Dyer, Chris, Chahuneau, Victor, and Smith, Noah A. (2013). A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the NAACL-HLT*, pages 644–648, Atlanta.

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

Gao, Qin and Vogel, Stephan. (2008). Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57, Columbus, Ohio.

Henríquez Q., Carlos A., Ruiz Costa-jussà, Marta, Daudaravicius, Vidas, Banchs, Rafael E., and Mariño, José B. (2010). Using collocation segmentation to augment the phrase table. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 98–102, Sweden.

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

Itagaki, Masaki, Aikawa, Takako, and He, Xiaodong. (2007). Automatic validation of terminology translation consistency with statistical method. In *Proceedings of MT Summit XI*, pages 269–274, Copenhagen, Denmark.

Johnson, J Howard, Martin, Joel, Foster, George, and Kuhn, Roland. (2007). Improving translation quality by discarding most of the phrasetable. In *Proceedings of EMNLP-CoNLL*, pages 967–975, Prague, Czech Republic.

Koehn, Philipp, Och, Franz J., and Marcu, Daniel. (2003). Statistical phrase-based translation. In *Proceedings of the NAACL*, pages 48–54, Edmonton, Canada.

Koehn, Philipp, Hoang, Hieu, Birch, Alexandra, Callison-Burch, Chris, Federico, Marcello, Bertoldi, Nicola, Cowan, Brooke, Shen, Wade, Moran, Christine, Zens, Richard, Dyer, Chris, Bojar, Ondrej, Constantin, Alexandra, and Herbst, Evan. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th ACL*, pages 177–180, Prague, Czech Republic.

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

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

Lardilleux, Adrien, Chevelu, Jonathan, Lepage, Yves, Putois, Ghislain, and Gosme, Julien. (2009). Lexicons or phrase tables? An investigation in sampling-based multilingual alignment. In *Proceedings of the third workshop on example-based machine translation*, pages 45–52, Dublin, Ireland.

Lardilleux, Adrien, Gosme, Julien, and Lepage, Yves. (2010). Bilingual lexicon induction: Effortless evaluation of word alignment tools and production of resources for improbable language pairs. In *Proceedings of the 7th LREC*, pages 252–256, Valletta, Malta, May.

Liang, Percy, Taskar, Ben, and Klein, Dan. (2006). Alignment by agreement. In *Proceedings of the HLT-NAACL*, pages 104–111, New York.

Ma, Yanjun, Stroppa, Nicolas, and Way, Andy. (2007). Bootstrapping word alignment via word packing. In *Proceedings of the 45th ACL*, pages 304–311, Prague.

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

Morishita, Yohei, Utsuro, Takehito, and Yamamoto, Mikio. (2008). Integrating a phrase-based smt model and a bilingual lexicon for human in semi-automatic acquisition of technical term translation lexicon. In *Proceedings of the 8th AMTA*, pages 153–162.

Neubig, Graham, Watanabe, Taro, Sumita, Eiichiro, Mori,

Shinsuke, and Kawahara, Tatsuya. (2011). An unsupervised model for joint phrase alignment and extraction. In *Proceedings of the ACL-HLT*, pages 632–641, Portland.

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

Och, Franz Josef. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of the 41st ACL*, pages 160–167, Sapporo, Japan.

Stolcke, Andreas. (2002). SRILM-an extensible language modeling toolkit. In *Proceedings of ICSLP*, volume 2, pages 901–904, Denver, Colorado.

Thurmair, Gregor and Aleksić, Vera. (2012). Creating term and lexicon entries from phrase tables. In *Proceedings of the 16th EAMT*, pages 253–260, Trento, Italy.

Vogel, Stephan, Ney, Hermann, and Tillman, Christoph. (1996). Hmm-based word alignment in statistical translation. In *Proceedings of the 16th COLING*, pages 836–841, Copenhagen, Denmark.