Lexicons or phrase tables? An investigation in sampling-based multilingual alignment

Adrien Lardilleux* Jonathan Chevelu*† Yves Lepage* Ghislain Putois† Julien Gosme*

* GREYC, Université de Caen Basse-Normandie
Caen, France
firstname.lastname@info.unicaen.fr

† Orange Labs
Lannion, France
firstname.lastname@orange-ftgroup.com

Abstract

Sampling-based multilingual alignment is an example-based approach to sub-sentential alignment that has proven to be able to outperform ubiquitous statistical models on some tasks. As for machine translation however, it still typically does not provide as good results as was first expected. In this paper, we propose two experiments to determine the nature of alignments produced by this method, and what they would still lack of. We then deduce what possible improvements will make the method perform better on machine translation tasks.

1 Introduction

Sub-sentential alignment from parallel corpora is the starting point of most data-driven machine translation systems. In [10], we proposed an alternative to the ubiquitous statistical models (*e.g.*, [11, 13]) to deal with the sub-sentential alignment task. This method only relied on straightforward example-based techniques.

The main goals of this method were to deal with the following issues:

multilinguality: sub-sentential alignment was introduced as a *bilingual* problem since its early stages. As a result, obtaining truly multilingual alignments (in three or more languages) always required pair-by-pair processing of languages [15]. Yet truly multilingual alignments could prove to be useful for many tasks, such as multilingual lexicography or the so-called "multi-source" approach to machine translation [3, 7];

scalability: traditional statistical methods may not scale up, nor even scale *down* [1]. Despite the growing availability of resources for numerous languages, some will probably never reach a coverage that could make them usable in real applications. On the other hand, huge amounts of input, while known to produce better results, quickly turn out to be a plague in processing time;

(true) simplicity: although efforts have already been undertaken to overcome this issue (*e.g.*, [11, 12]), there still is much room for simplification.

Our first evaluations focused on machine translation tasks, *i.e.*, evaluating the quality of the output of an SMT decoder with its default phrase tables (obtained from statistical models) and ours (example-based techniques). While we could surpass the former on some tasks [10], on average we still typically score 2-3 BLEU points behind, depending on the corpora used.

Yet some experiments (described hereafter) focusing on multilingual lexicon induction show that this approach can significantly outperform statistical models. These results lead us to the following question: does this method actually build phrase tables, or multilingual lexicons? The goal of this paper is to determine the nature of the alignments produced by this method, and to deduce what possible improvements will make it perform better on machine translation tasks.

This paper is organized as follows. Section 2 briefly describes the sampling-based approach to sub-sentential alignment. Section 3 gives some typical results on a machine translation task and a bilingual lexicon induction task. Section 4 investigates the nature of the alignment produced by the method, and discusses some possible improvements.

Mikel L. Forcada, Andy Way (eds.)
Proceedings of the 3rd Workshop on Example-Based Machine Translation, p. 45–52
Dublin, Ireland, November 2009

2 The method

We briefly describe the sampling-based approach to sub-sentential alignment, which allows alignments to be extracted from parallel corpora in multiple languages simultaneously. A complete description is available in [10].

Basically, the method consists in outputting sequences of words that share exactly the same distribution in a multilingual corpus, whatever the language they come from. Since there are more chances for words to have the same distribution on small corpora, we apply this technique on small random subcorpora obtained by sampling from a large corpus. The same is repeated for numerous random subcorpora, and the result is a list of alignments along with the number of times they were obtained. Translation probabilities and lexical weights can be computed so that the list of alignments become a full-fledged phrase table (or not, see next sections).

This method solely relies on distribution similarities to detect multilingual alignments, and on differences between sequences of words to extract these alignments. It thus positions itself in the example-based paradigm. It can be used as a component of various corpus-based NLP tasks, including lexical substitution for various MT architectures.

3 Two evaluations

In this section, we perform two experiments to evaluate the quality of the output of the above described alignment method. The analysis of the results will serve as a starting point to determine possible flaws in the method. In the first experiment, the list of alignments is assimilated to a phrase table to be used by a standard phrase-based SMT system (Moses [8]). In the second one, alignments are compared to a reference bilingual lexicon. We use the freely available implementation named Anymalign.[1]

Both experiments are performed using two distinct training parallel corpora:

1. 40,000 pairs of Japanese-English sentences from the BTEC [17] (sentence length in words: avg $=$ 10 $\pm$ std.dev. $=$ 5);

2. 200,000 pairs of French-English sentences from the Europarl corpus [6] (sentence length in words: avg $=$ 31 $\pm$ std.dev. $=$ 18).

We compare Anymalign's phrase table to Moses', *i.e.*, symmetric alignments obtained from IBM model 4, using MGIZA++ [5].[2] The default heuristic for symmetric alignment is used, as it has shown to produce the best results in our experiments. Anymalign's bilingual alignments are symmetric already, so this step is not necessary. Although both tools are suited for parallel processing, for a fair evaluation we only use them on a single processor.

3.1 Evaluation on a machine translation task

In a first experiment, we compare the quality of the output of the Moses decoder using its default phrase tables and Anymalign's. Since the latter can be stopped at any time, we start by producing Moses's default phrase tables, measure the elapsed CPU time, and run Anymalign for the same amount of time. In the case of the Japanese-English task, we used the data provided at the IWSLT07 campaign [4] (roughly 500 Japanese sentences for testing and 6 reference translations). As for the French-English task, 500 random pairs of sentences from the Europarl corpus were used for tuning and 500 other random pairs were used for testing. Translations are evaluated using the BLEU [14] and TER [16] metrics.

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

[2] `http://www.cs.cmu.edu/~qing/`

Table 1: Comparison of Anymalign's alignments against symmetric alignments obtained from IBM model 4 on a typical machine translation task. Quality is comparable when translating short Japanese sentences, while there is a significant gap in favor of IBM model 4 when translating long French sentences.

	BTEC: ja-en			Europarl: fr-en		
	# entries in phrase table	BLEU	TER	# entries in phrase table	BLEU	TER
Anymalign	**377,753**	**0.39**	**0.45**	3,528,674	0.25	0.60
IBM4	141,238	0.38	**0.45**	**6,788,046**	**0.29**	**0.56**

Results are shown in Table 1. The quality of translations is similar on the BTEC Japanese-English task. However, symmetric alignments from IBM model 4 give much better results on the Europarl French-English task. The phrase tables are also very different in terms of size: Anymalign's phrase table is much larger on the BTEC task, while it is much smaller on the Europarl task. The origin of these differences will be scrutinized in the next sections.

3.2 Evaluation on a bilingual lexicon induction task

In a second experiment, we compare the alignments against a reference bilingual lexicon. We use the EDICT Japanese-English dictionary[2] (about 120,000 entries) and a French-English dictionary from the Freelang project[3] (about 60,000 entries). We filter these dictionaries so that the actual reference contains only entries that can actually be extracted from the training corpora. Practically, an entry in a dictionary is kept as a reference if it is a subsequence of a pair of sentences from the corresponding training corpus. The final reference dictionaries contain 11,583 translation pairs for Japanese-English and 20,036 pairs for French-English.

Then, we compute the following three standard scores:

precision: sum of the source-to-target translation probabilities for those alignments from a phrase table that match an entry in the reference dictionary, divided by the number of unique source entries covered by these alignments;

recall: same as for precision except that we divide by the number of unique source entries covered by the reference dictionary;

f-measure: harmonic mean of precision and recall.

Results are presented in Figure 1. Since Anymalign can be stopped at any time, we perform the same evaluation for numerous processing times. As for MGIZA++, we vary the number of iterations of each model (IBM1, HMM, IBM3, and IBM4) and measure the elapsed CPU time. MGIZA++ needs much more time than Anymalign to reach the same level of quality when extracting alignments from long French-English sentences, and is clearly outperformed on short Japanese-English sentences.

3.3 Discussion

These two experiments clearly show that Anymalign performs much better at bilingual lexicon induction than phrase table production. It typically gives better results than statistical models on bilingual lexicons induction tasks, while phrase tables produced from statistical models are much more appropriate for MT

[3] `http://www.freelang.org/`

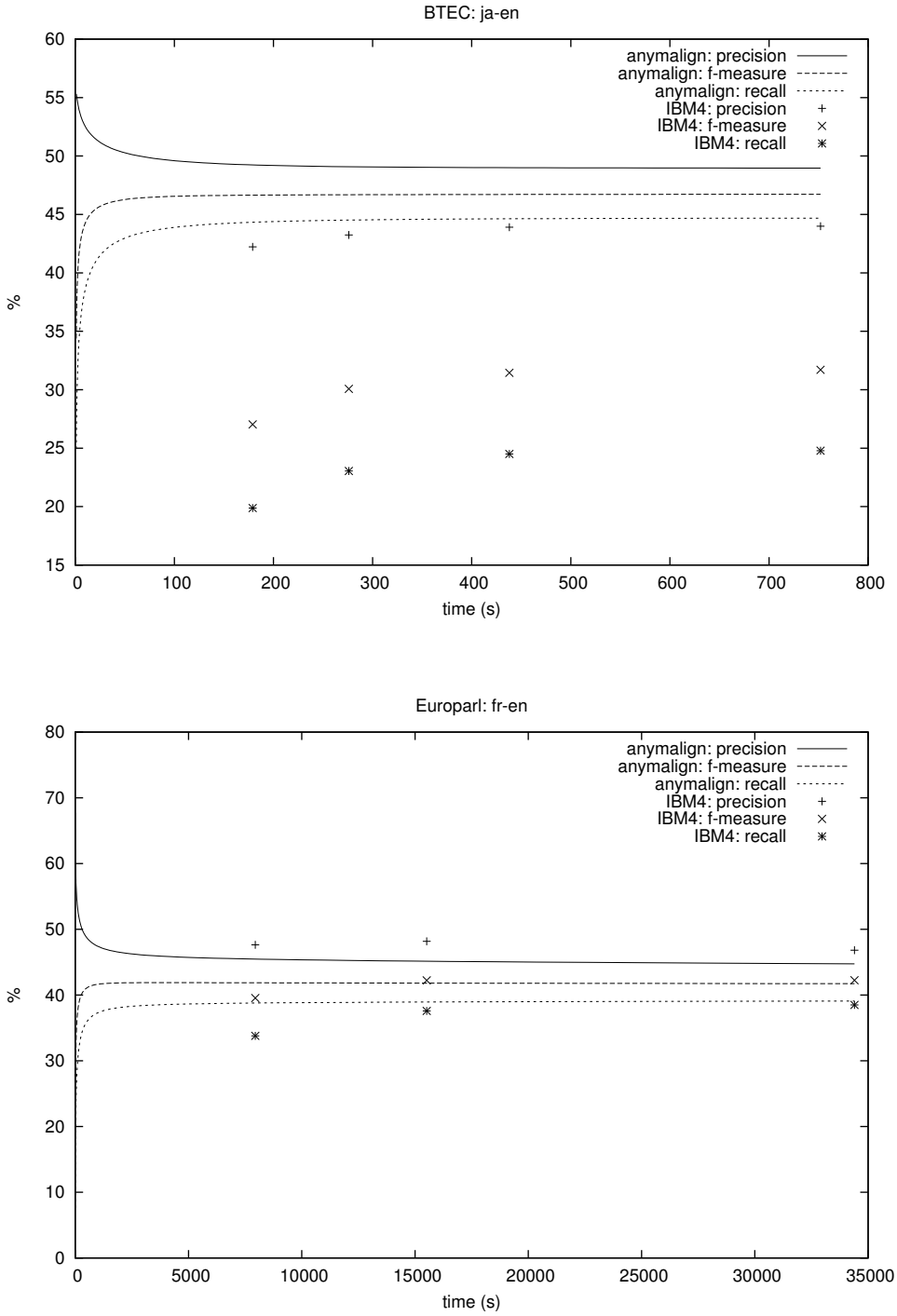

Figure 1: Comparison of Anymalign's alignments against symmetric alignments obtained from IBM model 4 on a bilingual lexical induction task. Anymalign generally yields equal or better results than MGIZA++, much faster. MGIZA++ eventually reaches Anymalign when extracting alignments from long French-English sentences, but it is far behind on short Japanese-English sentences.

Table 2: Number of unique source n-grams in Moses' default phrase table and Anymalign's. The "Europarl" column gives the number of distinct n-grams in the French part of the Europarl corpus: this is the number of source n-grams a phrase table should contain to ensure a complete coverage of the original corpus. Anymalign is closer to this reference than MGIZA++/Moses for unigrams, but it is outperformed by far on longer n-grams. The "intersection" column gives the number of source n-grams present in both phrase tables.

n-gram length	Europarl	Moses	Anymalign	Ratio	Intersection
1	95,881	54,961	**85,167**	0.65	52,107
2	1,698,717	**803,832**	196,812	4.08	127,006
3	6,375,026	**2,514,999**	53,732	46.81	37,941
4	11,580,772	**4,012,346**	31,980	125.46	24,567
5	14,622,988	**4,561,992**	34,039	134.02	27,394
6	15,663,784	**4,336,705**	37,200	116.58	30,171
7	15,611,099	**3,487,254**	35,289	98.82	28,781
Total	65,648,267	**19,772,089**	474,219	41.69	327,967

tasks. Since the reference bilingual lexicons we used mainly consist of unigrams, we naturally conclude that Anymalign produces better/more numerous unigram alignments, while it would lack the ability to align long n-grams.

To confirm this, we evaluated again the French-English alignments against the reference bilingual lexicon, after removing any entry in the reference lexicon that contain a n-gram with $n > 1$.[4] We re-evaluate the alignments previously used at the right-most side of Figure 1 (near 35,000s). For both systems, precision increases by 2%. However, recall increases by 5% for Anymalign and by only 2% for Moses' phrase table. The new f-measures are 46% for the former and 44% for the latter, Anymalign thus outperforming statistical models while it was slightly outperformed when including long n-grams in the original evaluation. This confirms that this method is better at unigram extraction, but lacks the ability to properly extract long n-grams.

4 Investigating the contents of alignments

4.1 Failing at aligning n-grams?

This section inspects the reason why the sampling-based approach would fail at aligning n-grams. To this end, we thoroughly investigate the content of the phrase tables obtained with Anymalign. In order to highlight differences with Moses' default phrase table, we now resort to 1,000,000 pairs of French-English sentences from the Europarl parallel corpus. This corpus is chosen because Anymalign has shown to produce the worst results when trained on it in our previous experiments (Table 1).

We are particularly interested in differentiating between unigrams and longer n-grams. Therefore, in a first experiment, we simply count the number of unique source entries in the phrase tables produced by MGIZA++/Moses and Anymalign, for each n-gram of length $1 \leq n \leq 7$. Results are shown in Table 2. Anymalign's phrase table's coverage is much higher on unigrams, which is in conformity with the results reported in [10] (nearly 90% of the source vocabulary is covered). However, it is clearly outperformed on all remaining n-gram lengths. In total, the difference in sizes between Moses' default phrase table

[4] In the original experiment, the average number of words per entry in the reference dictionary was 1.2 in both languages: n-grams up to $n = 7$ were evaluated.

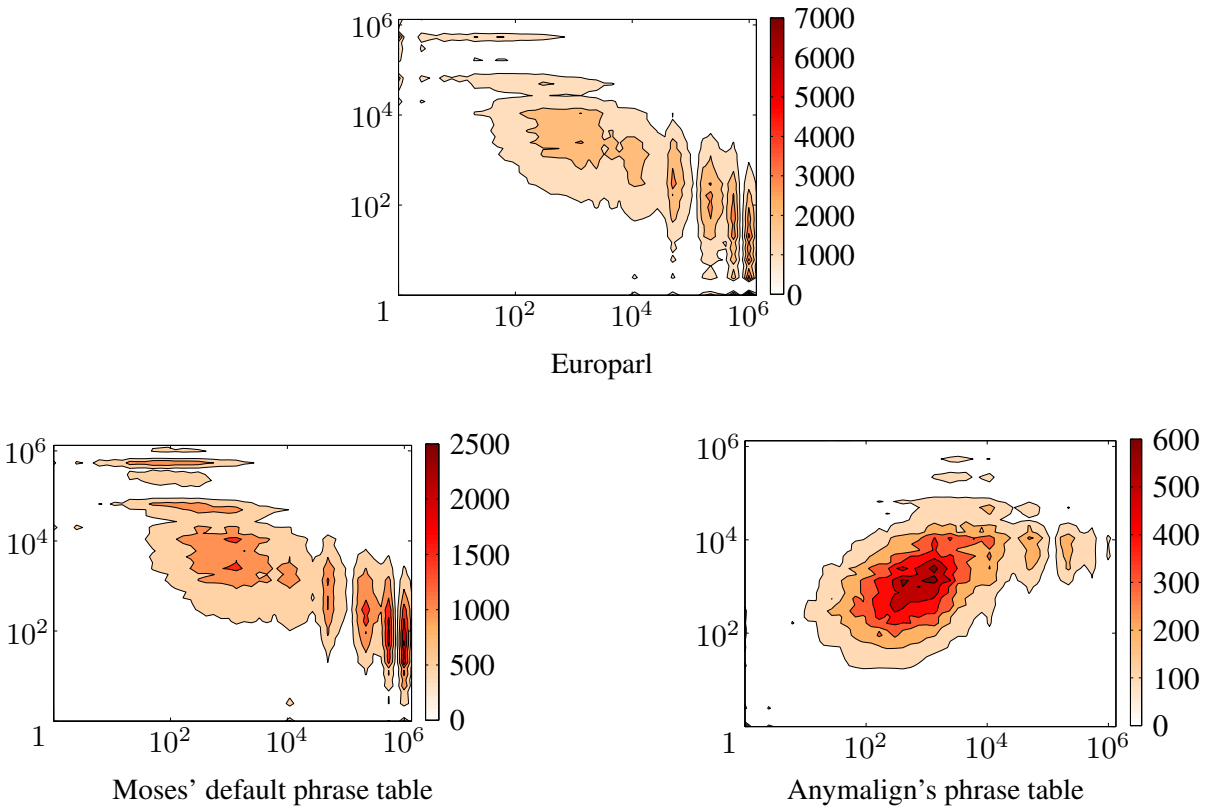

Figure 2: Comparison of distributions of bigrams from the source parts of Moses' default phrase table and Anymalign's. On the x-axis, the number of occurences of the first word of the bigram in the training corpus. On the y-axis, the number of occurences of the second word. The distribution of bigrams in Moses' default phrase table is closer to the natural distribution (Europarl French corpus) than in Anymalign's. Bigrams obtained from Anymalign tend to contain only words with similar frequencies (bottom left to upper right diagonal) while Moses' default's and Europarl's contain numerous words of different frequencies (upper left to bottom right diagonal). Note the difference in scale on the number of bigrams.

and Anymalign's is nearly two orders of magnitude ($\times 42$). This is consistent with the phrase table size difference in Table 1, which was not so large due to the smaller size of the training corpus. These results suggest that the reason why a phrase-based MT system would be at a disadvantage when built upon this method would not be a matter of *quality* of n-grams, but rather of *quantity*: the method simply does not align n-grams for $n \geq 2$ in sufficient number. The "intersection" column in Table 2 also suggests that the phrases produced by the two methods are quite different. Combining the two phrase tables may therefore be worth of interest, but we leave this for future work as it falls out the scope of this paper.

Manual inspection of the actual content of phrase tables suggests that Anymalign would not align sequences of words of different natures, such as a word followed by a punctuation mark, while Moses' default phrase table contains numerous such entries. To confirm or refute this remark in a second experiment, we focus on bigrams. We now count the number of source bigrams in a phrase table according to the frequency of the two words they are made of. We then draw the corresponding distribution, and compare it to the distribution of bigrams in the source part of our French-English Europarl corpus. Results are shown in Figure 2. Moses' default phrase table's bigram distribution is very close to the Europarl

one. The distribution for Anymalign looks very different: no bigram is visible on the upper left and bottom right zones of the figure, while they correspond to high density zones in Moses' and Europarl distributions. This confirms that the sampling-based approach does not align words with different frequencies.

4.2 Failing at aligning words with different frequencies?

Recall the description of the method in Section 2. The method outputs sequences of words that share exactly the same distribution in a subcorpus. The reason why words with different frequencies are not aligned together is that high frequency words (*e.g.*, a fullstop) and low frequency ones (typically a hapax) simply never share the same distribution. The only way for a fullstop and a hapax to share the same distribution would be to reduce a subcorpus down to a single sentence only, which would result in *all* words sharing the same distribution. In such a case, the method can only extract the whole sentence, thus not yielding any new information at all. This can eventually result in alignments as simple as French-English

$$\text{la maison} \leftrightarrow \text{the house}$$

not being extracted, because of the difference of frequency between determiner and noun. On the other hand, the method separately produces the two alignments

$$\text{la} \leftrightarrow \text{the} \qquad\qquad \text{maison} \leftrightarrow \text{house}$$

because the method will typically forge some random subcorpora where the two determiners share the same distribution (same for the two nouns).

4.3 So what does it miss?

All we have left to do is to recombine the alignments together in order to produce longer alignments. This is nothing more than what was initiated with phrase-based SMT [9], where phrase alignments are obtained by combining word-to-word alignments. In the above example, going through the original training corpus to detect the succession between determiner and noun would suffice to do the job. A re-estimation of the translation probabilities from the new set of alignments may then be necessary. We believe that such newly created phrase tables will close the gap that currently remains when undertaking machine translation tasks such as those based on the Europarl corpus.

5 Conclusion

This paper has investigated the nature of the alignments obtained by the sampling-based approach to sub-sentential alignment. We have shown than this approach excels at unigram extraction, while it lacks the ability to align long n-grams. The resulting alignments thus constitute more multilingual lexicons than actual phrase tables. The missing step would simply consist in recombining the alignments together in order to produce longer ones. Future developments will focus on this operation in order to improve the results of machine translation systems built on top of this alignment method.

References

[1] Tantely Andriamanankasina, Kenji Araki, and Koji Tochinai. Sub-sentential Alignment Method by Analogy. In *Proceedings of the 13th Pacific Asia Conference on Language, Information and Computation (PACLIC 13)*, pages 277–284, Taiwan, February 1999.

[2] Jim Breen. A WWW Japanese Dictionary. In *Language teaching at the crossroads*. Monash University Press, Clayton, 2003. JSC Working paper No. 13.

[3] Josep Maria Crego, Aurélien Max, and François Yvon. Plusieurs langues (bien choisies) valent mieux qu'une : traduction statistique multi-source par renforcement lexical. In *Actes de TALN/RECITAL 2009*, Senlis, France, June 2009.

[4] Cameron Shaw Fordyce. Overview of the IWSLT 2007 Evaluation Campaign. In *Proceedings of the 4th International Workshop on Spoken Language Translation (IWSLT 2007)*, pages 1–12, Trento, Italy, October 2007.

[5] Qin Gao and Stephan Vogel. Parallel Implementations of Word Alignment Tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57, Columbus, Ohio, June 2008. Association for Computational Linguistics.

[6] Philipp Koehn. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the 10th Machine Translation Summit (MT Summit X)*, Phuket, Thailand, September 2005.

[7] Philipp Koehn, Alexandra Birch, and Ralf Steinberger. 462 Machine Translation Systems for Europe. In *Proceedings of the he twelfth Machine Translation Summit (MT Summit XII)*, pages 65–72, Ottawa, Canada, August 2009.

[8] 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)*, Prague, Czech Republic, June 2007.

[9] Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical Phrase-Based Translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL 2003)*, volume 1, pages 48–54, Edmonton, Canada, 2003.

[10] Adrien Lardilleux and Yves Lepage. Sampling-based multilingual alignment. In *Proceedings of Recent Advances in Natural Language Processing (RANLP 2009)*, pages 214–218, Borovets, Bulgaria, September 2009.

[11] Percy Liang, Ben Taskar, and Dan Klein. Alignment by Agreement. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 104–111, New York City, USA, June 2006. Association for Computational Linguistics.

[12] Robert Moore, Wen-tau Yih, and Andreas Bode. Improved Discriminative Bilingual Word Alignment. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 513–520, Sydney, Australia, July 2006.

[13] Franz Och and Hermann Ney. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29:19–51, March 2003.

[14] 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, Pennsylvania, USA, 2002.

[15] Michel Simard. Text-translation Alignment: Three Languages Are Better Than Two. In *Proceedings of the Joint SIGDAT Conference of Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC)*, College Park, Maryland, USA, 1999.

[16] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of Association for Machine Translation in the Americas (AMTA 2006)*, Cambridge, Massachusetts, USA, August 2006.

[17] Toshiyuki Takezawa, Eiichiro Sumita, Fumiaki Sugaya, Hirofumi Yamamoto, and Seiichi Yamamoto. Toward a Broad-coverage Bilingual Corpus for Speech Translation of Travel Conversation in the Real World. In *Proceedings of the third International Conference on Language Resources and Evaluation (LREC 2002)*, pages 147–152, Las Palmas de Gran Canaria, Spain, 2002.