Sampling-based multilingual alignment

Adrien Lardilleux Yves Lepage
GREYC, University of Caen Basse-Normandie, France
Firstname.Lastname@info.unicaen.fr

Abstract

We present a sub-sentential alignment method that extracts high quality multi-word alignments from sentence-aligned multilingual parallel corpora. Unlike other methods, it exploits low frequency terms, which makes it highly scalable. As it relies on alingual concepts, it can process any number of languages at once. Experiments have shown that it is competitive with state-of-the-art methods.

Keywords

Sub-sentential alignment, low frequency term, hapax, sampling.

1 Motivation

Sub-sentential alignment from parallel corpora covers a variety of applications, such as the constitution of lexical resources or machine translation.

The widely used IBM models [2] and their extensions, implemented in the open source tool Giza++ [14], constitute the standard. Many alternatives or improvements have been proposed in the past years. Most of them are based on statistics, *e.g.* [6, 12, 13, 17], other ones are non-statistical methods, *e.g.* [1, 5]. All of them mainly address the issue of *quality* of alignments, *i.e.*, getting as close to human judgment as possible, or making machine translation as efficient as possible. Yet quality is only one aspect of alignment. Other issues still deserve to be explored:

- Some applications require alignments in more than two languages. This is particularly true for multilingual lexicography. As sub-sentential alignment was introduced as a *bilingual* problem since its early stages, obtaining truly multilingual alignments (in at least three languages) always required pair-by-pair processing of languages [16]. But the quality of alignments is hindered when relying on "pivot" languages.
- 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.
- These models are generally complex. This makes them difficult to integrate in actual applications, unless some free tool is available.

We propose a different approach to sub-sentential alignment that solely relies on *low frequency terms*. While often neglected, they actually provide an elegant solution to the above-mentioned issues.

This paper is organized as follows. Section 2 gives an overview of the concepts of the proposed multilingual alignment technique. Section 3 describes the technique in more details. Section 4 addresses the issue of multilingual alignment scoring. Section 5 compares the method with state-of-the-art tools.

2 Rationale

2.1 From high to low frequencies

Intuitively, one naturally trusts high frequency words, because of their statistical significance. As a result, low frequency words are often neglected and discarded, *e.g.* by removing all words which frequency is below a given threshold.

A practical answer has been long known: increase the amount of input data. Doing so naturally increases all word frequencies, turning low frequency words into high frequency ones. However, new words are always introduced meanwhile, that bear low frequencies. This is a vicious circle!

If one could safely align low frequency terms instead of focusing on high frequency ones, one would not need to indefinitely increase the amount of input data. Instead, *removing* input data would do the job, by turning high frequency words into low frequency ones. This would inherently lead to less processing, less resources required, and simpler mechanisms.

2.2 *Hapax legomena*

Amongst low frequency terms, *hapax legomena* (hapaxes for short), *i.e.*, words that appear only once in a corpus, are certainly those that show the greatest potential. While usually discarded, we have shown that they can be safely aligned [10]. Indeed, given a sentence-aligned parallel corpus in multiple languages, *sequences of hapaxes* contained in a particular sentence in all languages can be safely assumed to be lexical equivalences. Note that any number of languages can be processed simultaneously with this principle.

It is worth reminding that hapaxes typically represent 50% of the total vocabulary of a text [10]. As they are massively present, they can serve as the basis for the design of a sub-sentential alignment method.

Another advantage of hapaxes is that they are *unambiguous* in their corpus. Because they occur only once, they only have one possible meaning within this corpus. In other words, high frequency words can be naturally disambiguated — temporarily — by the simple means of removing data.

2.3 Bringing together low and high frequencies

Starting from the previous remarks, one could design a sub-sentential alignment method that consists in removing input data until some term to be aligned become a corpus hapax. By filtering input sentences so that this term be the only hapax in a particular sentence, hapaxes of the corresponding sentences in other languages would be expected to be its translations.

While some experiments have shown that this principle already delivers promising results, it simply lacks the ability to align very high frequency terms like periods, which appear in almost all sentences of a corpus. The only way to make a period become a hapax is to cut the corpus down to one sentence only. However, all words on this sentence would become hapaxes as well, which prevents them from being aligned separately.

This problem can easily be tackled by noticing that alignments of hapaxes are just a particular case of what we shall refer to as "perfect alignments," *i.e.*, sequences of words that strictly appear in the same sentences. An example is shown in Fig. 1. Most of these alignments are alignments of hapaxes [10], but they also include high frequency terms. Again, this is not restricted to language pairs: any number of languages can be processed simultaneously.

3 The method

We now describe the process by which alignments can be extracted from parallel corpora in multiple languages simultaneously. A free implementation is available at:
`http://users.info.unicaen.fr/~alardill/anymalign/`

3.1 Introducing *alingual* corpora

As stated previously, one of the main advantage of the method is that it can align any number of languages simultaneously. Fig. 1 shows examples in three languages. More languages could be added with absolutely no change. More surprising, the principle still holds with a *monolingual* corpus. Indeed, the simple process of searching words that strictly appear on the same lines (assuming one sentence per line) can be applied to a single language. What we obtain then is just some particular case of collocations. Doing so in multiple languages simultaneously is thus tantamount to extract "multilingual collocations."

Therefore, the whole alignment process can heavily be simplified by assimilating a multilingual input corpus to a monolingual one. This is done by discriminating all surface forms according to the language they come from: words with identical surface forms from different languages are considered to be different. Boundaries between languages are removed, and recovered after the alignment process, based on the origin of words.

Such a corpus is a view over multiple languages, and does not involve any language-dependent concept. We thus refer to it as an *alingual* corpus. It is the entry point of all subsequent processing. An example of alingual corpus is shown in Fig. 2.

3.2 Sampling input data

The core of the method consists in removing data from the input to decrease word counts. This process makes new "perfect alignments" appear, most of them being hapaxes. More precisely, numerous subcorpora are forged from which alignments are extracted.

We set on a sampling-based approach. In addition to be straightforward, this approach appears to be the most accurate because the natural distribution of words in the alingual corpus is left untouched. Because of its randomness, the complete coverage of input data cannot be ensured. This issue is easily tackled by extracting alignments from numerous random subcorpora of various sizes. Handling a large number of subcorpora is no problem since processing a subcorpus is fast. In addition, since all subcorpora are independent, parallel processing is possible.

Biasing the sampling

We note x the number of subcorpora of size k to be processed. We define it as follows: it must ensure that the probability that none of the sentences from a subcorpus of length k is ever chosen is below a certain threshold t, an indicator of the coverage of the input corpus. The lower t is, the better the coverage.

With n the size of the (alingual) input corpus ($1 \leq k \leq n$):

- the probability that a particular sentence is chosen is k/n;
- the probability that this sentence is not chosen is $1 - k/n$;
- the probability that none of the k sentences is chosen is $(1 - k/n)^k$;
- the probability that none of these k sentences is ever chosen is $(1 - k/n)^{kx}$.

Hence, the number of random subcorpora of size k to forge by sampling is defined by $(1 - k/n)^{kx} \leq t$, which yields:

$$x \geq \frac{\log t}{k \log (1 - k/n)}$$

Processing at least x random subcorpora of size k will thus ensure a proper coverage of the input corpus.

However, rather than setting in advance some particular degree of coverage (hence imposing a fixed number of subcorpora to process), we deduce from the above result a probability distribution to randomly draw the sizes of the subcorpora to process:

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

The numerator ($\log t$) was substituted for -1 because t is a constant: $t \leq 1 \Rightarrow \log t \leq 0$. This distribution highly favors small subcorpora. Experiments have shown that they provide more accurate and more numerous alignments than large subcorpora, in addition to be much faster to process [11].

Input corpus:

	English		French		German
1	One coffee , please .	↔	Un café , s'il vous plaît .	↔	Einen Kaffee , bitte .
2	This coffee is not bad .	↔	Ce café est correct .	↔	Dieser Kaffee ist nicht schlecht .
3	One strong tea .	↔	Un thé fort .	↔	Einen starken Tee .

⇓

"Perfect alignments:"

The words:					appear on lines:
One	↔	Un	↔	Einen	1 3
coffee	↔	café	↔	Kaffee	1 2
, please	↔	, s'il vous plaît	↔	, bitte	1
.	↔	.	↔	.	1 2 3
This is not bad	↔	Ce est correct	↔	Dieser ist nicht schlecht	2
strong tea	↔	thé fort	↔	starken Tee	3

Fig. 1: *Extracting "perfect alignments" from a toy parallel corpus in English, French, and German. Each line in the input corpus is a triple of aligned sentences. Sequences of words that strictly appear on the same lines are expected to be translations of each other.*

1 | One$_1$ coffee$_1$,$_1$ please$_1$.$_1$ Un$_2$ café$_2$,$_2$ s'il$_2$ vous$_2$ plaît$_2$.$_2$ Einen$_3$ Kaffee$_3$,$_3$ bitte$_3$.$_3$
2 | This$_1$ coffee$_1$ is$_1$ not$_1$ bad$_1$.$_1$ Ce$_2$ café$_2$ est$_2$ correct$_2$.$_2$ Dieser$_3$ Kaffee$_3$ ist$_3$ nicht$_3$ schlecht$_3$.$_3$
3 | One$_1$ strong$_1$ tea$_1$.$_1$ Un$_2$ thé$_2$ fort$_2$.$_2$ Einen$_3$ starken$_3$ Tee$_3$.$_3$

Fig. 2: *Assimilating a multilingual corpus to a monolingual one (same corpus as the one presented in Fig. 1, but words have been discriminated with subscripts: 1 for English, 2 for French, and 3 for German).*

3.3 Extracting alignments

To extract "perfect alignments" from all subcorpora obtained by sampling, the same process as depicted in Fig. 1 is applied, except that it runs on alingual sentences (see Fig. 2). In addition, since we can safely assume that "perfect alignments" yield good translations, the remaining parts of the sentences they appear on are likely to be translations of each other as well [3].

In other words, each "perfect alignment" yields up to two multilingual alignments per line:

1. the sequence of words that consists of the "perfect alignment" itself, preserving word order from the sentence;
2. the complementary of this sequence on the line (*i.e.*, its context), ordered as well.

Fig. 3 illustrates the process. Any alignment may be obtained a plurality of times, from different subcorpora and different lines. The result is a list of alignments along with the number of times they have been obtained.

In the general case, the method outputs non-contiguous sequences of words. They can subsequently be filtered according to specific criteria, like word contiguity, number of languages covered, or the number of words in a given language.

4 Scoring alignments

We propose two ways to score multilingual alignments by generalizing two well-known bilingual scoring techniques to the case of multilingual contexts.

4.1 Translation probabilities

Translation probabilities reflect the probability that some monolingual sequence of words of a multilingual alignment translates into the sequences of words in the remaining languages. We use the principle proposed in [9] to compute phrase translation probabilities, except that we generalize it to multilingual contexts: since there is no "source" and "target" languages in our multilingual alignments, each language becomes the "source" in turn, and *all* remaining languages together become a single "target" one.

In other words, assuming an input corpus in L languages, a score is computed for each language i ($1 \leq i \leq L$). It is the probability that the sequence of words s_i generates the rest of the alignment. It is computed by dividing the count of the current multilingual alignment, $C(s_1, \ldots, s_L)$, by the sum of the counts of all alignments in which s_i appears, $C(s_i)$:

$$P(s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_L | s_i) = \frac{C(s_1, \ldots, s_L)}{C(s_i)}$$

Table 1 gives an example of actual data in three languages: each alignment is assigned three scores.

In the case of bilingual alignment, these scores directly correspond to the traditional pair $P(source|target)$ and $P(target|source)$. If the input data is monolingual, the score is always $C(s_1)/C(s_1) = 1$.

4.2 Lexical weights

Lexical weights were proposed in [9] to validate the quality of alignments. Given a bilingual alignment to score, it consists in checking how well each source word translates into the target words it links to. When a source word links to multiple target words, the average of their translation probabilities is used. A source-to-target lexical weight is then the product of all scores. The same holds from target to source, and the result is a pair of lexical weights between 0 and 1. We adapt this technique with three major changes.

First, since there is no source and target languages in multilingual alignments, we use the same principle as previously: each language becomes the source in turn, and the rest of the alignment is assimilated to the target. We end up with as many lexical weights per alignment as there are input languages.

Input corpus: see Fig. 2

⇓

Extract "perfect alignments" and their contexts:

The words:	appear on lines:	from which we extract:
One$_1$ Un$_2$ Einen$_3$	1	One$_1$ Un$_2$ Einen$_3$ coffee$_1$,$_1$ please$_1$.$_1$ café$_2$,$_2$ s'il$_2$ vous$_2$ plaît$_2$.$_2$ Kaffee$_3$,$_3$ bitte$_3$.$_3$
	3	One$_1$ Un$_2$ Einen$_3$ strong$_1$ tea$_1$.$_1$ thé$_2$ fort$_2$.$_2$ starken$_3$ Tee$_3$.$_3$
coffee$_1$ café$_2$ Kaffee$_3$	1	coffee$_1$ café$_2$ Kaffee$_3$ One$_1$ _ ,$_1$ please$_1$.$_1$ Un$_2$ _ ,$_2$ s'il$_2$ vous$_2$ plaît$_2$.$_2$ Einen$_3$ _ ,$_3$ bitte$_3$.$_3$
	2	coffee$_1$ café$_2$ Kaffee$_3$ This$_1$ _ is$_1$ not$_1$ bad$_1$.$_1$ Ce$_2$ _ est$_2$ correct$_2$.$_2$ Dieser$_3$ _ ist$_3$ nicht$_3$ schlecht$_3$.$_3$
⋮	⋮	⋮

⇓

Collect alignments, count them, and restore boundaries between languages:

English		French		German	Count
One	↔	Un	↔	Einen	2
coffee , please .	↔	café , s'il vous plaît .	↔	Kaffee , bitte .	1
strong tea .	↔	thé fort .	↔	starken Tee .	1
coffee	↔	café	↔	Kaffee$_3$	2
One _ , please .	↔	Un _ , s'il vous plaît .	↔	Einen _ , bitte .	1
This _ is not bad .	↔	Ce _ est correct .	↔	Dieser _ ist nicht schlecht .	1
		⋮			⋮

Fig. 3: *Extracting multilingual alignments from an alingual corpus. Underscores (_) mark discontinuities within one language.*

English (e)		French (f)		German (g)	Count	Translation probabilities			Lexical weights		
						$P(f,g\|e)$	$P(e,g\|f)$	$P(e,f\|g)$	$W(f,g\|e)$	$W(e,g\|f)$	$W(e,f\|g)$
loud applause	↔	vifs applaudissements	↔	lebhafter beifall	122	0.730	0.760	0.826	0.936	0.995	0.990
loud applause	↔	vifs applaudissements	↔	starker beifall	24	0.144	0.143	0.820	0.936	0.995	0.895
loud applause	↔	vifs applaudissements	↔	(lebhafter beifall)	12	0.072	0.092	0.667	0.936	0.995	0.060
loud applause	↔	applaudissements prolongés	↔	lebhafter beifall	8	0.048	0.167	0.048	0.916	0.995	0.990
loud applause	↔		↔	beifall	1	0.006	0.000	0.006	0.836	1.000	0.991

Table 1: *Alignments of the English word sequence "loud applause" obtained from a sample of the Europarl corpus [7], along with their associated scores.*

Second, as we start without any word-to-word alignment, we estimate a simple lexical translation probability distribution D based on relative word frequencies from the input corpus:

$$D(w_j|w_i) = \frac{C(w_i, w_j)}{C(w_i)}$$

where w_i is a word in language i and w_j is a word in language j $(i \neq j)$.

Lastly, the sampling-based approach does not *link* words, as would statistical models do. For example, in the first alignment of Table 1, one would expect English "loud" to link to French "vifs," and "applause" to "applaudissements." Our method does not permit this; instead, the complete phrase "loud applause" is considered to be a translation of the phrase "vifs applaudissements" as a whole. Therefore, where [9] computed the *average* of relative word frequencies for those words that link together, we actually compute the *maximum* of relative word frequencies for *all* possible links, *i.e.*, from all "source" words to all "target" words.

Formally, within an alignment, we look for the best possible translation of a word w_i from sequence s_i (in language i) amongst all words in other languages, according to distribution D, and retain this probability. The lexical weight for language i is the product of all probabilities retained, after determining the best translation for each word in s_i:

$$W(s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_L|s_i) = \prod_{w_i \in s_i} \max_{w_j \in \bigcup_{i \neq j} s_j} D(w_j|w_i)$$

5 Evaluation

We evaluate the method by comparing the output of the Moses statistical machine translation decoder [8] using its default phrase tables (refined alignments from Giza++ [14]), against those produced by our method. We present results on two tasks: the IWSLT07 Japanese to English classical task [4], and a Spanish to French task using the Europarl corpus [7].

For each task, a standard Giza++ training is run using the default set of options, and processing time is

System	BLEU score	Entries in phrase table	Input corpus coverage
Giza++ ($t = 404s$)	0.45	141,338	69%
Our system ($t/2$)	0.42	241,810	89%
+ lexical weights	0.44		
Our system (t)	0.42	324,213	89%
+ lexical weights	0.45		
Our system ($t \times 2$)	0.42	**420,391**	**90%**
+ lexical weights	**0.46**		

Table 2: *Evaluation results on the IWSLT07 Japanese to English machine translation task. The input corpus consists in roughly 40,000 aligned sentences (average sentence length: 10 words).*

System	BLEU score	Entries in phrase table	Input corpus coverage
Giza++ ($t = 27,791s$)	**0.32**	**9,614,327**	67%
Our system ($t/2$)	0.29	1,393,278	85%
+ lexical weights	0.30		
Our system (t)	0.30	1,953,576	85%
+ lexical weights	0.31		
Our system ($t \times 2$)	0.30	2,690,782	**86%**
+ lexical weights	0.31		

Table 3: *Evaluation results on a Spanish to French machine translation task. The input corpus consists in roughly 200,000 aligned sentences (average sentence length: 31 words).*

measured. This time serves as a reference for our system, which can be stopped at any time. Three runs are performed: the first one is stopped after half of the reference time has elapsed, the second takes the same amount of time as the reference time, and the last one takes twice as long. All phrase tables have the five same features (two translation probabilities, two lexical weights, and length penalty). We systematically measure the contribution of lexical weights by removing them from the phrase tables and performing an additional run with the decoder.

Results are presented in Tables 2 and 3. We use BLEU [15] to measure translation quality. As for the Japanese to English task, the best results are obtained by running our system twice as long as the standard Giza++ training. Lexical weights give a significant performance boost. This is certainly due to the fact that our phrase tables contain noise (hence their size), that lexical weights help reduce. For example, on the third line of Table 1, the last score is very low because of the presence of brackets in the alignment.

This performance hint is not as visible on the Spanish to French task, because our phrase tables are much smaller than Moses' default. We still could come very close to Giza++'s quality.

Note that in a sample-based approach, quality is not a matter of time; coverage is: the method consists in continuously outputting "perfect alignments" and their contexts from various samples of the input corpus. The time, subcorpus, and sentence they have been extracted from do not matter: all alignments are on an equal footing from the quality point of view. The randomness of this process requires numerous subcorpora to be forged to ensure a proper coverage. However, Tables 2 and 3 show that the coverage of our phrase tables is always much higher than that obtained with Giza++, even within less time ($t/2$) or when the phrase table is smaller.

6 Conclusion

We described a complete alignment method, which allows multiple languages to be aligned simultaneously from parallel corpora. It solely relies on the use of low frequency terms. It makes it highly flexible regarding the amount of input data. The sample-based approach allows the user to interrupt the alignment process at any time and still produce high quality translation tables. Experiments show that it can match the accuracy of Giza++, while exhibiting a much higher coverage of input data, and being by far simpler.

References

[1] T. Andriamanankasina, K. Araki, and K. Tochinai. Sub-sentential alignment method by analogy. In *Proceedings of PACLIC 13*, pages 277–284, Taiwan, 1999.

[2] P. Brown, S. D. Pietra, V. D. Pietra, and R. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, 1993.

[3] I. Cicekli. Similarities and differences. In *Proceedings of SCI2000*, pages 331–337, Orlando, FL, USA, 2000.

[4] C. S. Fordyce. Overview of the IWSLT 2007 evaluation campaign. In *Proceedings of IWSLT 2007*, pages 1–12, Trento, Italy, 2007.

[5] E. Giguet and P.-S. Luquet. Multilingual lexical database generation from parallel texts in 20 european languages with endogenous resources. In *Proceedings of COLING/ACL 2006*, pages 271–278, Sydney, Australia, 2006.

[6] D. Gildea. Loosely tree-based alignment for machine translation. In *Proceedings of ACL 2003*, pages 80–87, Sapporo, Japan, 2003.

[7] P. Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit X*, Phuket, Thailand, 2005.

[8] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL 2007*, Prague, Czech Republic, 2007.

[9] P. Koehn, F. J. Och, and D. Marcu. Statistical phrase-based translation. In *Proceedings of HLT-NAACL 2003*, volume 1, pages 48–54, Edmonton, Canada, 2003.

[10] A. Lardilleux and Y. Lepage. The contribution of the notion of hapax legomena to word alignment. In *Proceedings of LTC'07*, pages 458–462, Poznań, Polland, 2007.

[11] A. Lardilleux and Y. Lepage. A truly multilingual, high coverage, accurate, yet simple, sub-sentential alignment method. In *Proceedings of AMTA 2008*, pages 125–132, Waikiki, Hawai'i, USA, 2008.

[12] I. D. Melamed. Models of translational equivalence among words. *Computational Linguistics*, 26(2):221–249, 2000.

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

[14] F. Och and H. Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29:19–51, 2003.

[15] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL 2002*, pages 311–318, Philadelphia, Pennsylvania, USA, 2002.

[16] M. Simard. Text-translation alignment: Three languages are better than two. In *Proceedings of EMNLP/VLC*, College Park, Maryland, USA, 1999.

[17] Y. Zhang and S. Vogel. An efficient phrase-to-phrase alignment model for arbitrarily long phrase and large corpora. In *Proceedings of EAMT-05*, Budapest, Hungary, 2005.