International Journal of Advanced Intelligence
Volume 3, Number 2, pp.189-217, July, 2011.

AIA International Advanced Information Institute

The Contribution of Low Frequencies to Multilingual Sub-sentential Alignment: a Differential Associative Approach

Adrien Lardilleux

LIMSI-CNRS
BP 133, Orsay Cedex, France
Adrien.Lardilleux@limsi.fr

Yves Lepage

Waseda University—Graduate School of Information, Production and Systems
808-0135 Hibikino 2-7, Wakamatu-ku, Kitakyuusyuu-si, Fukuoka-ken, Japan
Yves.Lepage@aoni.waseda.jp

François Yvon

LIMSI-CNRS/University Paris-Sud
BP 133, Orsay Cedex, France
Francois.Yvon@limsi.fr

Received (3 January 2011)
Revised (11 April 2011)

The goal of this paper is to show that, contrary to preconceived ideas, one can efficiently take advantage of low frequency words in natural language processing. We put them to use in sub-sentential alignment, which constitutes the first step of most data-driven machine translation systems (statistical or example-based machine translation). We show that rare words can be used as a foundation in the design of a multilingual sub-sentential alignment method, using differential techniques similar to those found in example-based machine translation. This method is *truly multilingual*, in that it allows the simultaneous processing of any number of languages. Moreover, it is very simple, *anytime*, and scales up naturally. We compare our implementation, *Anymalign*, with two statistical tools proven in the domain. Although its current results are on average slightly behind those of state of the art methods in phrase-based statistical machine translation, we show that the intrinsic quality of our lexicons is actually superior to that of lexicons produced by state of the art methods.

Keywords: natural language processing; hapax legomenon; multilingualism; machine translation; alignment; rare events.

1. Introduction

Sub-sentential alignment of parallel texts is an important preliminary task for numerous natural language processing (NLP) applications like machine translation (MT), multilingual information retrieval, or word synonymy detection. In the case of machine translation, it constitutes a major preliminary step for most data-driven systems, be they statistical (SMT) or example-based (EBMT). In general, the goal

of sub-sentential alignment is to produce a translation table, a data representation format originally used in phrase-based SMT. Translation tables consist of a set of translation pairs along with associated translation scores. These scores may represent various features, like, typically, observed conditional probabilities, or may reflect some computed likelihood that a given source unit translates to a particular target unit. Other applications of sub-sentential alignment include, among others, dictionary induction and term extraction.

Sub-sentential alignment methods can be divided into two main categories: the estimative approach, introduced by Brown et al.,[1] and the associative approach, introduced by Gale and Church.[2] The former consists in building a statistical model of the bitext, the parameters of which are estimated through a global maximization process, i.e., over all sentence pairs of the input parallel corpus simultaneously. Practically, the goal is to determine the best set of correspondences (alignment links) between all source and target words[a] of each sentence pair. The latter approach relies on a device that produces a list of translation candidates, each of them being subject to an independence statistical test such as, for instance, Dice coefficient[3] or mutual information[4] (see also more recent work by Melamed[5] or Moore[6]). Those translation candidates with an association measure higher than expected under the independence assumption are assumed to be translation pairs. In this approach, the process is a local maximization process, i.e., each segment is processed independently.

On one hand, the estimative approach has gained much popularity within the MT research community, mainly because it is mathematically well-founded, and because some studies have shown that it generally outperforms associative models.[7] In addition, it is tightly integrated within the SMT framework since the apparition of the now ubiquitous IBM models,[8] the most recent freely available implementation of which is the tool MGIZA++.[9] On the other hand, the best estimative models are relatively complex when compared to the standard associative models: Tufis and Barbu[10] report that while the complexity of associative methods typically grows quadratically with the size of the vocabulary, the complexity of estimative models may be exponential. Another drawback of estimative models is that they result in numerous parameters requiring fine adjustment in order to produce the best possible results. The tools that implement such models typically have numerous options that reflect these parameters.[b] Most users do not take the time to tune them and eventually use such tools as a blackbox.[6,11] For such reasons, recent research in sub-sentential alignment has concentrated on improving associative approaches,[6] or combining simple estimative models, like IBM models 1 and 2, and HMM,[11] in order to produce alignments of similar quality, if not better, but at a much lesser

[a] In this paper, the term "word" refers to a surface form as identified by a tokenization program.
[b] For instance, MGIZA++ is made of about 30,000 lines of code, has 58 command-line options, and a standard run typically produces more than 20 temporary files. Aligning large corpora can take days.

computational cost.

The research reported here follows the latter trend, but *quality* of alignments is not our only concern. In this paper, we propose an associative sub-sentential alignment method that fills several important gaps that most studies on sub-sentential alignment have neglected until now. In particular, contrary to most sub-sentential alignment models, our method is *non-directional*. As an important consequence, it is not restricted to the processing of pairs of languages: any number of languages can be aligned simultaneously; the method is *multilingual* from design. In addition, the method relies on a single simple model, which allows for natural scaling to very large input parallel texts, allows for massive parallelism, and makes it easily accessible to non-specialists.

This paper is organized as follows. In Section 2, we evaluate the impact of rare words in sub-sentential alignment through a series of observations. These observations will serve as a starting point in the design of a sub-sentential alignment method in Section 3. Section 4 details the different steps that constitute the method, and an optimization is proposed in Section 5. Section 6 compares our implementation with state-of-the-art tools, and Section 7 concludes this work.

2. Useful facts about rare words

The study of low frequency words constitute the foundation of this work. We started to focus on them because of a simple observation: although it is well known that rare words are massively present in any text, they are underused in most NLP tasks because of their low statistical significance. We believe on the contrary that they can serve as a valuable resource due to their large number. The fact that the majority of the vocabulary of a text actually corresponds to low frequency words is usually illustrated by Zipf's law.[12,13,14] It expresses a relation between the rank of words of a text ordered in decreasing order of frequency and this frequency: the product *rank* $\times$ *frequency* is more or less constant. In other words, most words have a very low frequency (content words), while very few have a high frequency (function words). For instance, in a sample (about 350,000 sentences) of the English part of the Europarl parallel corpus,[15] words that occur at most 10 times represent 74% of the vocabulary. While some work rely on specific techniques to process rare words efficiently,[16] we will show how the study of their distribution can be used as a cornerstone of an alignment method.

2.1. *Hapaxes in corpora*

Hapaxes[c] are words that occur only once in a text. Moore[17] reports the following widespread belief:

[c]From the Greek *hápax legómenon* 'which has been uttered once'. In this paper, we use the plural *hapaxes* for convenience.

> Consider the case of two words that each occur only once in a corpus, but happen to co-occur. Conventional wisdom strongly advises suspicion of any event that occurs only once [...].

By definition, hapaxes are discarded in approaches that filter out low frequency words. This is often the case in associative alignment methods since they usually rely on a statistical significance test. For example, Cromières[18] defines a lower bound on word frequencies before considering a word for alignment; Giguet and Luquet[19] define a threshold proportional to the inverse term length.

In addition to their infrequency, another prejudice against hapaxes is that they often correspond to neologisms or misspellings.[20] Neologisms should be considered words on their own right. As for misspelled words, their quantity depends on the quality of the corpus used. According to Nishimito,[21] who interprets the results of Evert and Lüdeling,[22] each error in a corpus occurs only once in average. Misspelled words are thus typically hapaxes, but their proportion within the totality of hapaxes remains very low. In any case, they are not problematic in sub-sentential alignment from parallel corpora: a hapax, if misspelled, is still a hapax, the only consequence being that the resulting alignment will contain a misspelled word, with no impact on the alignment process itself. In the case where a frequent word gets misspelled, then we can assume that the alignments in which it intervenes will obtain very low scores since the error presumably occurs only once, and the resulting erroneous alignment will naturally be disregarded.

Even though they are generally discarded, hapaxes are very common. For example, in the above-mentioned sample of the Europarl corpus, hapaxes cover 39% of the total vocabulary. This figure is similar to those generally found in the literature.[23] It reflects two main axes. The first one is the richness of the vocabulary, i.e. the quantity of different words (word types) used in a text. Counts on Shakespeare's most read plays reveal that they contain 58% hapaxes in average.[d] The second axis reflects the degree of synthesis of the language: isolating, synthetic, or polysynthetic. The more synthetic a language, the more inflected words, hence the more word types. The proportion of hapaxes increases accordingly. As a noticeable example, Langlais et al.[24] report more than 80% hapaxes on a corpus of Inuktitut, a highly synthetic language of Canada. In such a case, rejecting hapaxes would be tantamount to consider only 20% of vocabulary, which would seriously hinder the quality of any subsequent task.

As a concluding remark, we emphasize that the proportion of hapaxes in a text is almost constant, whatever the size of this text. This is illustrated on Fig. 1. When increasing the length of a text, new occurrences of words that were previously hapaxes may be introduced, so these words are no more hapaxes; however, new hapaxes are introduced as well. The relation between hapaxes and unknown words[25,23,26]

[d] Counts available at `http://www.mta75.org/curriculum/English/Shakes/index.html` (last visited on 14/01/11).

makes them useful to estimate the behavior of MT systems on unknown words.

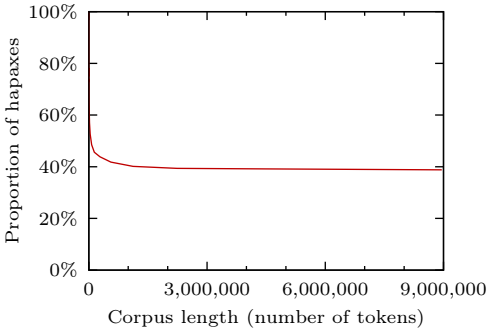

Fig. 1. Proportion of hapaxes in the English part of the Europarl corpus. The proportion is almost constant when the text has reached a certain length (here, around 1,000,000 word tokens).

2.2. *Low frequency words in sub-sentential alignment*

We now show that the majority of the best alignments obtained from parallel corpora with a standard associative alignment method mainly consists of alignment of rare words. Our goal here is not to obtain the best possible results, but to highlight some characteristics of associative methods. To this end, we use a "simple" associative method that assigns to each possible pair of words (*source*, *target*) a single score that reflects the probability that *source* and *target* are translations of each other. Amongst the many available association scores, the cosine is a classical association score used in various domains, such as named entity discovery,[27] conceptual vectors for semantic tasks,[28,29] and of course sub-sentential alignment.[19]

The cosine method consists in defining a vector space whose number of dimensions is the number of sentence pairs in the parallel input corpus. For each language, we associate each word w to a vector $\vec{w}$ whose i-th element is the number of occurrences of w in the i-th sentence. For each pair of words (w_s, w_t), we then compute the angle between their associated vectors, $\vec{w_s}$ and $\vec{w_t}$:

$$\text{angle}(\vec{w_s}, \vec{w_t}) = \text{acos}\left(\frac{\vec{w_s} \cdot \vec{w_t}}{||\vec{w_s}|| \times ||\vec{w_t}||}\right)$$

where $\vec{u} \cdot \vec{v}$ denotes the dot product of vectors $\vec{u}$ and $\vec{v}$ and $||\vec{u}||$ the norm of vector $\vec{u}$. The result is the score of the alignment (w_s, w_t), a positive real number ranging from 0 to $\pi/2$ (inclusive). A score of 0 (i.e., a cosine of 1) means that the two words are *a priori*[e] good translations because their associated vectors are collinear. A score of $\pi/2$ means they are very unlikely to be translations.

[e]For conciseness, we do not develop the fact that even alignments with angle 0 may be erroneous.

We apply the previous method on the French-Spanish parts of the above-mentioned sample of the Europarl corpus. The French vocabulary contains 73,695 word types (37% are hapaxes) and the Spanish vocabulary 93,043 (40% are hapaxes). We thus obtain $73,695 \times 93,043 \simeq 7$ billion pairs of words along with their associated scores. We then study the quality and quantity of alignments according to the frequencies of the two words they are made of.

The left graph in Fig. 2 shows the *a priori* quality of alignments according to the frequency of the source and target words. Only those alignments that consist of (rare, rare) and (frequent, frequent) pairs of words have an angle greater than $\pi/6$. All alignments for which at least one of the two words has an intermediary frequency (say, from 10 to 100,000 occurrences) have an angle greater than $\pi/3$, which corresponds to the vast white surface that occupies the majority of the figure. Note the slightly emerging diagonal between the lower left and the upper right parts of the figure, meaning that words of similar frequencies tend to align together.

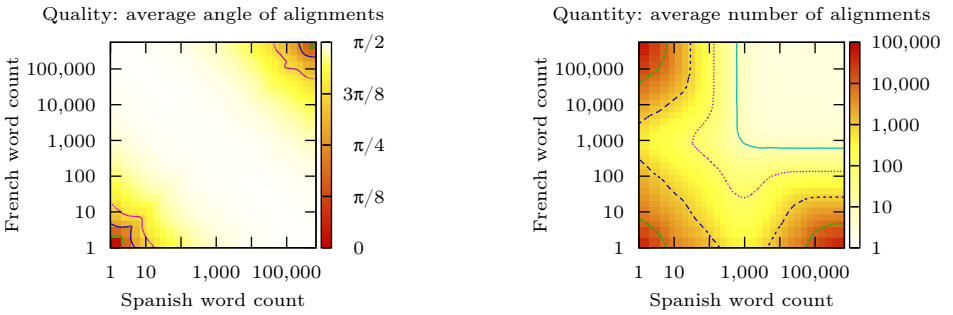

Fig. 2. Quality and quantity of word-to-word alignments obtained by the cosine method according to the count of the words they are made of. The left figure shows that the best alignments are only obtained from very rare or very frequent words (contour lines correspond to multiples of $\pi/8$). The right figure shows that the most numerous alignments involve at least one rare word (contour lines correspond to powers of 10).

The right graph in Fig. 2 shows the quantity of alignments according to the frequency of the source and target words. The most numerous alignments consist of (rare, rare), (rare, frequent), and (frequent, rare) pairs of words, while (frequent, frequent) ones are almost nonexistent. This is naturally implied by the fact that the majority of the vocabulary is made of rare words.

The conclusion of this study is that, when looking for quality *and* quantity in sub-sentential alignment, one should seriously take pairs of rare words into consideration. Neglecting them would be a mistake. See also work by Moore[17] on log-likelihood ratios and Fisher's exact Test.

2.3. *Alignment of hapaxes*

As mentioned above, even the "best" alignments (according to their angles) are not necessarily good translations. In particular, when several hapaxes occur in the same sentence, all source hapaxes get independently aligned with all target hapaxes with a null angle. On the other hand, if there is only one source hapax and one target hapax in a given pair of sentences, then the resulting alignment has much chance to be correct. In order to determine if this commonly happens, we investigate the distribution of hapaxes in our data.

The frequencies of hapaxes in our corpus are shown in Table 1. Although the number of hapaxes generally represent half of the vocabulary, they appear in a small set of sentences (9% in Spanish, 5% in French). More importantly, most sentences containing a hapax (85% in both languages) actually contain only one hapax, with an average of 1.2 hapaxes per sentence. Hence, the case where a source sentence and its translation both contain a single hapax should not be uncommon.

Table 1. Frequencies of hapaxes in our Europarl corpus: proportion of sentences that contain at least one hapax, proportion of sentences that contain exactly one hapax, amongst those that contain at least one, and average number of hapaxes per sentence.

Language	At least 1 hapax	1 hapax	Hapaxes/sentence
Spanish	9%	85%	1.2 ± 0.7
French	5%	85%	1.2 ± 0.8

Amongst those alignments with a null angle that are exclusively made of hapaxes, an important part (6,230 i.e. 21%) are alignments made of hapaxes from sentences that contain only one hapax in both languages. These alignments suffice to cover 7% of the Spanish vocabulary and 8% of the French vocabulary. Table 2 shows a sample of such alignments. These alignments are probably amongst the best: discarding them only because they were obtained from hapaxes would be a mistake. Note that Spanish and French are relatively close languages, and the results may get much worse on more distant languages. For instance, when performing the same experiment on the Finnish-English parts of our sample of Europarl, we obtained more than a half of erroneous alignments. Having said that, if one can obtain good results on close languages with such a simple method, we might expect at least as good results with a more evolved one. This is our focus in the next section.

3. Designing a sub-sentential alignment method using low frequencies

We have so far focused on 1-1 alignments in a bilingual context, in particular with hapaxes. In this section, we build on the previous results the necessary reasoning in order to design a method that is able to produce *m-n* alignments, in more than

Table 2. A sample of alignments of hapaxes from sentences that contain only one hapax. Only one alignment is erroneous, marked by a star (*fantoche* ↔ *oubliette*).

Spanish	French	Meaning
descolonizar	décolonisé	'decolonized'
predeterminarán	prédéterminerer *(misspelled word)*	'predetermine'
wallner	wallner	'(proper name)'
h-0818	h-0818	'(identifier)'
*fantoche	*oubliette	'puppet' / 'oubliette'
burns	burns	'(proper name)'
pseudojurídicos	pseudo-juridiques	'pseudo-legal'
h-0484	h-0484	'(identifier)'
antimaastrichtiana	anti-maastrichtienne	'anti-maastricht'
archiconocidos	archiconnus	'well-known'

two languages simultaneously, and for all words, whatever their frequencies.

3.1. *Less data is better data*

We previously mentioned that low frequency words were often discarded from associative alignment procedures because of their low statistical significance. Thus, only words of sufficient frequency get aligned. In order to align rare words with a method that concentrates on frequent ones, the accepted common wisdom in the field recommends to increase the amount of input data, so that the counts of rare words increase as well; once frequent, they can be aligned. However, when adding new data, new words are also added, most of them being rare words (see Fig. 1 for the case of hapaxes), which may result in an endless process. On the contrary, a method relying on the exploitation of rare words would not require to add new data. This is all the opposite, actually: *removing* data would suffice to perform alignment. Adding new data is a potentially infinite process, removing data is not.

Considering a corpus as a whole constituted of a finite number of events entails to assign a fixed probability to each of them. On the contrary, by removing data from a corpus, as we intend to do, *new corpora*, i.e., new collections of data, can be built. Many new input corpora are made available in this way. The number of subcorpora of a given input corpus of n lines is $2^n - 1$, each one having its own set of events and associated probabilities. As a result, to some extent, removing input data is tantamount to adding even more data!

3.2. *Advantages*

In addition to making low frequency words more present, extracting alignments from small subcorpora presents several advantages. We detail some of them.

3.2.1. *Modeling advantages*

Simplicity Intuitively, aligning low frequency words should be simpler than aligning high frequency ones. Deciding whether two frequent words that appear more or less in the same sentences have to be aligned is typically a decision problem. There is no such question with hapaxes, for which no approximation is required: the two words appear in the same sentence, or they do not.

Significance Rare events gain in significance, because they now happen in several different subcorpora. In other words, no event is rare anymore.

3.2.2. *Linguistic advantages*

Disambiguation A hapax can only have one meaning in the text it appears in. Consequently, when decreasing word counts by removing input data, a certain form of—temporary—disambiguation is implicitly performed, because frequent words may become hapaxes and thus unambiguous in subcorpora through this process.

Multilingualism As already mentioned, aligning several languages simultaneously will become possible. However, this is more thanks to the way we shall process rare words than because of their rarity itself: by extracting multiword units in a multilingual context assimilated to a monolingual one. This point will be further investigated in the next section.

3.2.3. *Computational advantages*

Less memory The amount of data to be processed decreases when removing some from input. Thus we can process large corpora without caring about memory resources required by the machine, and even modest computers will be able to run the program.

Massive parallelism Since the alignment relies on subcorpora processing, we can easily process different subcorpora on different processors or different machines. We just need to ensure that all processes launched are independent and that their outputs can be merged.

3.3. *Bringing together low and high frequencies*

The last issue we need to address before moving to a fully specified alignment method concerns high frequency words. One of the conclusions of Section 2 was that the best alignments are constituted of rare words in large quantities and frequent words in much smaller quantities, intermediate frequency words being unproductive of alignments (Fig. 2). We consequently investigated the possibility to produce alignments using only low frequency words, in particular hapaxes. There exists however a case where a word cannot become a hapax by removing input data: very

high frequency words remain very frequent, no matter the size of the subcorpus. Consider for instance the case of the period (assimilated to a word), that presumably appears in all sentences of a corpus. It is very difficult to align it by removing data from a corpus because the only way to make it become a hapax would be to consider a subcorpus of only one sentence. However, in such a subcorpus, almost all words also occur only once: the period is not the only hapax in both languages and therefore cannot be aligned separately. Ideally, we would like to process low and high frequency words in a uniform way.

Actually, frequent and rare words have the following in common: they align well by methods that only rely on their distribution in a corpus, as is the case with the cosine method (Section 2.2). These are words that do not translate ambiguously in the corpus used. Practically, they are words that share strictly the same distribution.

To summarize, our proposed associative method will not look for low frequencies neither for high frequencies, but for words that strictly share the same distribution, irrespective of their count. In practice, most of these words will have a very low frequency (typically hapaxes), but some others will be very high frequency words, like punctuations.

A first algorithm

We now have all we need to design a first algorithm. It consists of the three following main points:

do
(1) Select a subcorpus
(2) Extract sequences of words that share the same distribution
loop
(3) Calculate scores for alignments

These three steps will be described in detail in the two next sections. The main loop may run indefinitely. It can be interrupted by the user at any time, or when some specific criteria are met, such as elapsed time, coverage of the input corpus, number of new alignments obtained per second, etc. The number of subcorpora processed does not influence *quality*, but rather *quantity* and *significance* of the results: the longer the aligner runs, the more alignments, and the more significant their scores.

4. A detailed description of the method

4.1. *Truly multilingual processing*

In order to produce alignments in more than two languages simultaneously, we rely on a particular transformation of the input multilingual corpus. Assume we have the following Arabic-French-English toy input corpus:

1	قهوة ، من فضلك .	↔ Un café , s'il vous plaît .	↔ One coffee , please .
2	هذه قهوة ممتازة .	↔ Ce café est excellent .	↔ This coffee is excellent .
3	شاي ثقيل .	↔ Un thé fort .	↔ One strong tea .
4	قهوة ثقيلة .	↔ Un café fort .	↔ One strong coffee .

Our transformation consists in making this input corpus *monolingual*, where each sentence is the concatenation of the different translations of one sentence of the initial parallel corpus. In addition, all words are distinguished according to the language they come from, so that words with identical surface forms but from different languages are still considered different, as is the case here of the French and English periods:

1	قهوة$_1$ ،$_1$ من$_1$ فضلك$_1$.$_1$ Un$_2$ café$_2$,$_2$ s'il$_2$ vous$_2$ plaît$_2$.$_2$ One$_3$ coffee$_3$,$_3$ please$_3$.$_3$
2	هذه$_1$ قهوة$_1$ ممتازة$_1$.$_1$ Ce$_2$ café$_2$ est$_2$ excellent$_2$.$_2$ This$_3$ coffee$_3$ is$_3$ excellent$_3$.$_3$
3	شاي$_1$ ثقيل$_1$.$_1$ Un$_2$ thé$_2$ fort$_2$.$_2$ One$_3$ strong$_3$ tea$_3$.$_3$
4	قهوة$_1$ ثقيلة$_1$.$_1$ Un$_2$ café$_2$ fort$_2$.$_2$ One$_3$ strong$_3$ coffee$_3$.$_3$

For the sake of the presentation, we use subscripts to distinguish words: 1 for Arabic, 2 for French, and 3 for English. Since this corpus is an abstraction over several languages and does not imply any knowledge about these languages, we refer to it as an *alingual* corpus. This kind of corpus is the starting point for all subsequent processing.

4.2. *Alignment extraction*

The next step consists in extracting sequences of word tokens that share the same distribution in a particular subcorpus. For this, we index each word of the subcorpus according to the sentences it appears in. For simplicity, assume we start from the subcorpus made of the first three lines of the above alingual corpus. We associate to each word the vector of its occurrences, as was needed for the cosine method in Section 2.2:

	.$_1$	،$_1$	ثقيل$_1$	شاي$_1$	فضلك$_1$	قهوة$_1$	ممتازة$_1$	من$_1$	هذه$_1$	,$_2$	,$_3$	.$_1$	.$_2$	.$_3$	Ce$_2$	One$_3$	This$_3$	Un$_2$	café$_2$	coffee$_3$	est$_2$	excellent$_2$	excellent$_3$	fort$_2$	is$_3$	...
1	1	0	0	1	1	0	1	0	1	1	1	1	1	0	1	0	1	1	1	0	0	0	0	0	...
2	0	0	0	0	1	1	0	1	0	0	1	1	1	1	0	1	0	1	1	1	1	1	0	1	...
3	0	1	1	0	0	0	0	0	0	0	1	1	1	0	1	0	1	0	0	0	0	0	1	0	...

Sorting the columns makes identical vectors adjacent:

	.$_1$	.$_2$	.$_3$	قهوة$_1$	café$_2$	coffee$_3$	One$_3$	Un$_2$	،$_1$	فضلك$_1$	من$_1$	,$_3$	,$_2$	plaît$_2$	please$_3$	s'il$_2$	vous$_2$	هذه$_1$	ممتازة$_1$	Ce$_2$	This$_3$	est$_2$	excellent$_2$	...
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	...
2	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	...
3	1	1	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	...

The result is a list of groups of words that share the same distribution. They can be seen as our first alignments. In addition, since we assume that sequences of words that share the same distribution are likely translations, we can also assume that the remaining parts of the sentences are good translations as well. This principle has often been used in example-based machine translation: for instance, Cicekli and

Güvenir[30] assume that the similar parts of two source sentences of a parallel corpus are translations of the similar parts of the corresponding target sentences (same for the differences). The main difference with our work is that we do not process sentences by pairs, but by subcorpora, that may contain several sentences.

By doing so, each group of words produces up to two alignments for each sentence it appears in: (1) the sequence of words made of the group itself, preserving the word order from the sentence, and (2) the complementary of this sequence in the sentence, i.e. its contexts. With our previous example,

the words:		appear in sentences:	from which we extract:
قهوة$_1$	café$_2$ coffee$_3$	1	قهوة$_1$ café$_2$ coffee$_3$ Un$_2$ _ ,$_2$ s'il$_2$ vous$_2$ plaît$_2$.$_2$ One$_3$ _ ,$_3$ please$_3$.$_3$ من$_1$ فضلك$_1$ ،$_1$.$_1$
		2	قهوة$_1$ café$_2$ coffee$_3$ Ce$_2$ _ est$_2$ excellent$_2$.$_2$ This$_3$ _ is$_3$ excellent$_3$.$_3$ هذه$_1$ _ ممتازة$_1$.$_1$
		⋮	

The same is performed for each group of words having identical distribution, and for various subcorpora. A given alignment may be obtained several times, from different subcorpora or different sentences. The global result is a list of alignments along with the number of times they were obtained. We eventually restore boundaries between languages according to word subscripts:

Arabic		French		English	Count
قهوة	↔	café	↔	coffee	2
، من فضلك .	↔	Un _ , s'il vous plaît .	↔	One _ , please .	1
هذه _ ممتازة .	↔	Ce _ est excellent .	↔	This _ is excellent .	1
		⋮			

4.3. *Scoring alignments*

The next step is to transform the list of alignments into a translation table by computing scores for each alignment. We compute two types of scores that were initially proposed by Koehn et al.[31]: translation probabilities (alignment probabilities), based on counts of alignments, and lexical weights, based on the counts of words within the alignments. We generalize these standard scores so that they can be used with multilingual alignments. In practice, given an alignment in L languages, we calculate one score per language, that reflects the likelihood that the sequence of words in this language translates simultaneously to all other sequences of the alignment (to the $L-1$ remaining languages). The same is done for both types of scores, so that $2L$ numbers are assigned to each alignment. In the case of bilingual

alignments, these numbers are the analogous of the traditional source-to-target and target-to-source probabilities.

4.3.1. *Translation probabilities*

Translation probabilities are calculated from the number of times each alignment was obtained. They reflect the probability that a given sequence of words s_i (in language i: $1 \leq i \leq L$) translates to the rest of the alignment:

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

with $\mathrm{C}(s_1, \ldots, s_L)$ the count of the alignment (rightmost column of last table of Section 4.2) and $\mathrm{C}(s_i)$ the sum of the counts of all alignments where s_i appears. See Table 3 for examples.

Table 3. Examples of multilingual alignments along with their associated translation probabilities and lexical weights. In this example, we used a subset of the English-French-German Europarl corpus. The alignments displayed are those for the English sequence *loud applause*, obtained by running our system for five minutes on 20,000 Europarl sentences. The three translation probabilities are $\mathrm{P}(e|f,g)$, $\mathrm{P}(f|e,g)$, $\mathrm{P}(g|e,f)$. On the first line, the first translation probability is $\mathrm{P}(\textit{vifs applaudissements}, \textit{lebhafter beifall} | \textit{loud applause}) = 122/(122+24+12+8+1) = 0.73$.

English (e)		French (f)		German (g)	Count	Trans. prob.
loud applause	↔	vifs applaudissements	↔	lebhafter beifall	122	0.73 0.76 0.83
loud applause	↔	vifs applaudissements	↔	starker beifall	24	0.14 0.14 0.82
loud applause	↔	vifs applaudissements	↔	(lebhafter beifall)	12	0.07 0.09 0.67
loud applause	↔	applaudissements prolongés	↔	lebhafter beifall	8	0.05 0.17 0.05
loud applause	↔		↔	beifall	1	0.01 0.00 0.01

4.3.2. *Lexical weights*

Lexical weights have been proposed by Koehn et al.[31] to validate the quality of alignments. They are known to slightly improve the quality of translations obtained by statistical machine translation systems. Given a bilingual word alignment, the goal is to check which target words each source word translates to, and to retain their translation probabilities. When a source word translates to several target words, the average of the alignment probabilities is used. The source-to-target lexical weight is then the product of these scores. The same is performed from target to source, and the result is a pair of lexical weights between 0 and 1.

We adapt this scoring method by introducing a major change. Our method does not create *links* as estimative sub-sentential alignment methods do, so we do not know which words correspond to which words within an alignment. Instead, our method directly extracts multi-words translations pairs, in which no finer-grained information is contained. Therefore, where Koehn et al.[31] compute the *average* of

the probabilities of linked words, we retain the *maximum* of the probabilities of all possible links, i.e., from a source word to *all* target words. In our multilingual context, this maximum is searched amongst all remaining languages. Lexical weights are thus calculated as follows:

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

with $\mathrm{C}(w_i, w_j)$ the co-occurrence count of words w_i and w_j on the list of alignments and $\mathrm{C}(w_i)$ the count of word w_i. All word counts are weighted by the count of the alignment in which the words occur.

5. Subcorpora selection strategy

In this section, we define a strategy whose goal is to optimize the efficiency of our method. The only parameter required is the size of subcorpora from which we extract alignments. Practically, the total number of subcorpora is too important to process them all, as this number grows exponentially with the size of the initial input corpus, and processing them all would be pointless. The distribution of subcorpora according to their size is well approximated by a Gaussian: with n the size of the input corpus, there exists only one subcorpus of size n, one of size 0, and a maximum number of subcorpora of size $n/2$. However, we will not particularly focus on these mid-sized subcorpora, because what is important is not the number of possible subcorpora, but rather the number of correct alignments they can produce.

5.1. *Definition of a probability distribution*

The strategy we adopt consists in building subcorpora by random sampling according to a particular distribution. There are three main reasons for this choice. First, it allows for a quick discovery of alignments, contrary to an approach that would require to provide the units to be aligned in input. Second, a random sampling ensures that the "natural" distribution of words will not be altered: since a corpus is generally constituted so that it represents a sample of a language,[32] a sample of this corpus should also represent a sample of this language. Eventually, this strategy is very straightforward while already producing high quality results, although we intend to move toward more elaborate strategies in further research.

Because of the randomness of this strategy, we might expect that two identical experiments yield different results. These differences are minimal in practice. Another issue is that coverage of the input corpus cannot be guaranteed unless extracting alignments from numerous subcorpora. To deal with this issue, we define a probability distribution that tries to maximize the coverage of the vocabulary of the input corpus. Practically, this is achieved by ensuring that a maximum number of sentences from the input corpus are drawn in a maximum of different subcorpora. The distribution is used solely to draw the size of subcorpora. Once a subcorpus size

is drawn according to this distribution, the corresponding number of sentences are randomly chosen from the initial input corpus according to a uniform distribution.

Let x_k be the number of subcorpora of k sentences to process. x_k must ensure that the probability that none of the sentences of a subcorpus of size k be never chosen is below a certain threshold t that reflects the coverage of the input corpus: the closer to zero, the better the coverage. Let n be the number of sentences in the initial input corpus ($1 \leq k \leq n$):

- the probability that a given sentence is chosen in a sample of size k is k/n;
- the probability that it 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 the k sentences is ever chosen is $(1 - k/n)^{kx_k}$.

The number of subcorpora of size k to be drawn by sampling is thus constrained by $(1 - k/n)^{kx_k} \leq t$, which yields:

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

This formula means that processing at least x_k random subcorpora of size x_k guarantees the coverage of the input corpus vocabulary.

Rather than defining in advance a particular degree of coverage, which implies a fixed number of subcorpora to process, we deduce from the preceding result a probability distribution to randomly draw the size of the next subcorpus from which to extract alignments:

$$\mathrm{p}(k) \propto \frac{-1}{k \log (1 - k/n)}$$

The numerator, $\log t$, was substituted for -1 because t is a constant: $t \leq 1 \Rightarrow \log t \leq 0$. In the implementation we normalize this equality so that the $\mathrm{p}(k)$ sum up to 1:

$$\sum_{k=1}^{n} \mathrm{p}(k) = 1$$

This distribution highly favors small subcorpora.[f] The purpose of the next section is to show that this is indeed beneficial for our task.

5.2. *Impact of subcorpus sizes*

In this section, we study the impact of the size of subcorpora on the alignments extracted by our method. The following experiments have been run on the French-Spanish part of our previously used sample of Europarl, which is made up of roughly 350,000 sentences.

[f]Since small values of k come up most of the time, and having $\log(1 + x) \sim x$ when x is small, the distribution is close to $1/k^2$ most of the time.

5.2.1. *Processing time*

First, the smaller a subcorpus, the faster it is to process. Figure 3 shows that the time required to process a subcorpus is approximately linear in the number of sentences it contains. Processing 1,000 subcorpora of 100 sentences will thus typically take as much time as processing one subcorpus of 100,000 sentences. For a fair comparison between subcorpora of different sizes, the next measures will take into account the overall processing time rather than to the number of subcorpora processed.

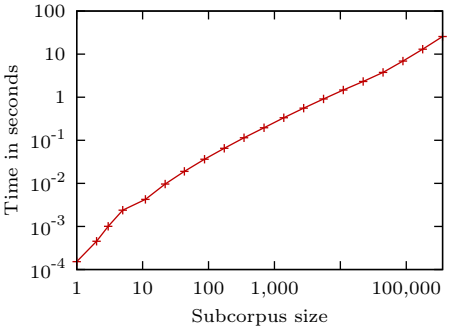

Fig. 3. Time required to process a single subcorpus according to its size (in number of sentences). Processing time increases almost linearly with the size of the subcorpus.

5.2.2. *Quantity of alignments*

Next, small subcorpora produce many more distinct alignments than larger ones, as is shown in Fig. 4. Typically, the longer the elapsed time (hence the larger the number of subcorpora processed sequentially), the more alignments obtained. As we might expect, in the lower part of the figure, the number of alignments obtained from subcorpora of only one sentence tends to be the number of sentences contained in the main corpus (roughly 350,000): these are the aligned sentences from the parallel corpus left untouched, no more no less. The fastest increase in the number of alignments is approximately obtained with subcorpora made of two to ten sentences. The number of distinct alignments then progressively decreases when the size of subcorpora increases. At the top of the figure, and despite the fact that this is hardly visible, the only subcorpus of size n produces the totality of its 45,548 possible alignments during its first run, which takes about 25 seconds.

From the implementation point of view, we do not randomly select subcorpora of extreme sizes (1 and 350,000), as it is more efficient to process them all exhaustively.

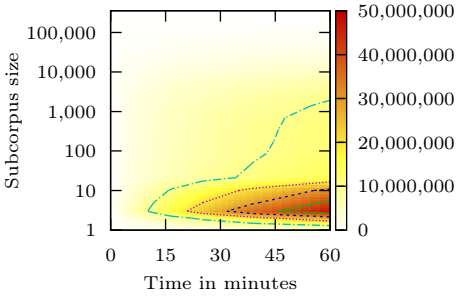

Fig. 4. Number of distinct alignments obtained by our method according to processing time and subcorpus size. Small subcorpora yield more alignments. Contour lines correspond to multiples of 10 millions.

5.2.3. *Quality of alignments*

The size of subcorpora also impacts the quality of alignments extracted, which is essential. The larger the subcorpora, the smaller the sequences of words that share the same distribution, and the larger their contexts. Fig. 5 illustrates this tendency.

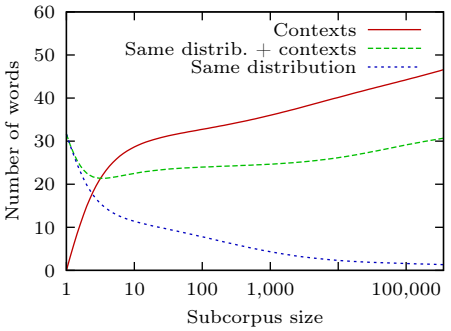

Fig. 5. Average length of alignments obtained by our method according to the size of subcorpora they have been extracted from. The smaller the subcorpus, the longer the sequences of words of equal distribution and the shorter their contexts. The curve corresponding to all alignments (same distribution + contexts) tends to follow that of the contexts because contexts are much more numerous.

In a last experiment, we count the number of alignments which occur in a French-Spanish reference lexicon according to the size of the subcorpora. This bilingual lexicon mainly consists of unigrams (source-target word pairs). The results are presented in Fig. 6. The maximum is reached with subcorpora made of roughly

1,000 sentences. This shows that relatively small subcorpora can produce small alignments that are also of good quality.

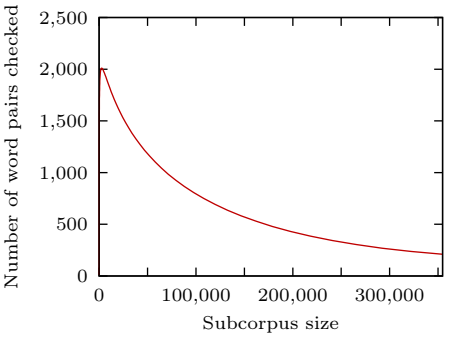

Fig. 6. Number of word alignments found in a reference bilingual lexicon according to subcorpus size. The system was run for 25 seconds in this experiment, which corresponds to the time required to process the largest subcorpus. The curve increases very quickly and reaches the maximum for subcorpora of about 1,000 sentences, then progressively decreases.

Summary: complete algorithm

Transform the multilingual parallel corpus into an alingual corpus
Initialize *AlignmentCounts*
do
  Select a subcorpus of k sentences with $\mathrm{p}(k) = \frac{-1}{k \log(1-k/n)}$
  Compute for each word its vector of presence/absence in sentences
  Sort the words according to their vectors in order to build *group*'s
  For each *group* of words:
    For each *sentence* the *group* appears in:
      Restore word order in *group*
      AlignmentCounts[group] ++
      AlignmentCounts[sentence - group] ++
until timeout **or** no new alignment is obtained **or** etc.
Compute scores for alignments

Figure 7 presents a screen shot of an alignment file produced by our tool.

6. Evaluation

All the evaluations described hereafter are *bilingual* because, to our knowledge, there exists no other multilingual aligner (with the notable exception of work by Simard[33]) or evaluation protocol. Hence, these experiments do not reflect all the

No	Freq.	Translation probabilities	Lexical weights	es	en	ar	zh	ja
1	3279	0.71 0.72 0.71 0.72 0.94	1.00 1.00 1.00 1.00 0.64	.	.	.	。	。
2	457	0.10 0.10 0.10 0.10 0.93	1.00 1.00 1.00 1.00 0.41	.	.	.	。	です 。
3	161	0.72 0.65 0.61 0.85 0.65	0.99 0.95 0.94 0.75 0.82	Dónde	Where	أين	哪	どこ
4	143	0.03 0.03 0.03 0.03 0.86	1.00 1.00 1.00 1.00 0.35	.	.	.	。	ます 。
5	125	0.93 0.94 0.92 0.93 0.91	0.99 0.95 0.99 0.90 0.88	Japón	Japan	اليابان	日本	日本
6	81	0.95 0.95 0.98 0.98 0.95	0.99 0.98 1.00 1.00 1.00	Tokio	Tokyo	طوكيو	东京	東京
7	70	0.02 0.02 0.02 0.02 0.93	0.99 1.00 0.96 0.99 0.63	.	.	.	。	です
8	57	0.01 0.01 0.01 0.01 0.55	0.99 1.00 0.96 0.99 0.71	.	.	.	。	を
9	54	0.71 0.71 1.00 0.68 0.69	1.00 1.00 1.00 0.97 1.00	pasaporte	passport	جواز	护照	パスポート
10	48	0.01 0.01 0.01 0.01 0.79	0.99 1.00 0.98 0.99 0.66	.	.	.	。	の
11	44	0.01 0.01 0.01 0.01 0.90	0.99 1.00 0.98 0.99 0.84	.	.	.	。	が
12	43	0.98 0.98 0.98 0.98 1.00	0.88 1.00 0.96 0.89 0.97	aeropuerto	airport	المطار	机场	空港
13	34	1.00 1.00 1.00 1.00 1.00	1.00 0.98 1.00 1.00 1.00	Chicago	Chicago	شيكاغو	芝加哥	シカゴ
14	33	0.75 0.97 0.59 0.59 0.60	1.00 0.84 0.99 0.94 1.00	Nueva York	New York	نيويورك	纽约	ニューヨーク
15	33	0.01 0.01 0.01 0.01 0.59	0.99 1.00 0.96 0.99 0.69	.	.	.	。	に
16	32	0.01 0.01 0.01 0.46 0.01	1.00 1.00 1.00 1.00 0.64	.	.	.	了 。	。
17	29	0.94 0.94 0.94 1.00 0.94	0.96 1.00 1.00 0.98 1.00	Boston	Boston	بوسطن	波士顿	ボストン
18	27	0.96 0.96 1.00 0.96 0.96	1.00 0.97 1.00 1.00 1.00	Londres	London	لندن	伦敦	ロンドン
19	26	0.96 0.96 0.96 0.93 0.96	1.00 1.00 1.00 1.00 0.97	Tanaka	Tanaka	تاناكا	T a n a k a	タナカ
20	22	0.10 0.09 0.08 0.54 0.09	0.99 0.95 0.94 0.26 0.82	Dónde	Where	أين	在	どこ
21	21	1.00 1.00 1.00 1.00 1.00	1.00 1.00 1.00 1.00 1.00	Yamada	Yamada	يامادا	Y a m a d a	ヤマダ
22	21	0.00 0.00 0.00 0.51 0.01	1.00 1.00 1.00 1.00 0.64	.	.	.	的 。	。
23	20	1.00 0.91 0.87 0.95 0.87	0.96 0.99 0.73 0.72 0.86	hoy	today	اليوم	今天	今日
24	19	0.90 1.00 0.86 0.86 0.66	1.00 1.00 1.00 1.00 1.00	Miami	Miami	ميامي	迈阿密	マイアミ

Fig. 7. A screen shot of a multilingual alignment file produced by our method in HTML format by running our aligner on a parallel corpus in 5 languages (only the top alignments are visible). Our implementation supports several formats: simple text, HTML, TMX, and translation table for the Moses SMT system.

potential of our method. We do not evaluate our implementation in an absolute manner but compare it with two statistical tools that are considered as state-of-the-art in the domain. They are relatively recent and are freely available software. The three tools we thus compare are the following:

Anymalign [g] This is the implementation of our method. Its main differences with the two other aligners are:

- the underlying method is close to associative alignment approaches, and the alignment extraction phase uses concepts similar to those found in some example-based machine translation systems.
- it does not create *links* between source and target words, but di-

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

rectly produces translations. Applications for which word position is important (e.g., SMT) may miss this piece of information. For other applications, this absence is likely to result in a gain in speed.
- it can be stopped at any time during its execution. Time has no influence on alignment *quality*, but on *coverage*: the longer the aligner runs, the more alignments it outputs.

Because of this last point, in practice we will first run the other aligners, measure their processing time, and run Anymalign during the same amount of time. In order to evaluate the impact of the subcorpus size selection strategy, Anymalign will be run twice: once using the distribution introduced in Section 5.1, and once with a uniform distribution (i.e. all subcorpus sizes are selected with equal probability).

MGIZA++ [h] It is the last descendant of the GIZA tradition.[34,7,9] It implements the ubiquitous IBM models,[8] and serves as a foundation of many SMT-related work. By default, it runs models IBM1, HMM, IBM3, and IBM4 for 5 iterations each. We will vary this number of iterations from 1 to 5. These models are asymmetric, so they need to be executed once in each direction (source to target and target to source) and their results must be made symmetric in order to give the best results. We perform this with the tools distributed with the Moses toolkit.[35] Note that this step is not required by Anymalign because its alignments are symmetric by design.

BerkeleyAligner [i] Introduced by Liang et al.,[11] it relies on simple models such as IBM1, IBM2, and HMM, which are trained jointly from source to target and from target to source, in order to produce better results than by training them separately as is the case with MGIZA++. The alignments it produces are thus symmetric. This tool has advanced features such as supervised word alignment, but for a fair comparison with the two other tools, we do not use these features. By default, the tool runs IBM model 1 and HMM for two iterations each. As with MGIZA++, we will vary this number from 1 to 5.

Although the three aligners are capable of parallel processing, for a fair comparison we use them on a single processor. Except for the number of iterations of the models they run, we keep their default parameters values. A summary of the programs we use in order to produce a complete translation table starting from a parallel bilingual corpus is shown in Fig. 8.

In the following, execution times will be measured by including all steps, from the input parallel corpus to the output translation table. We present detailed evaluation results using a sample of the Europarl corpus (100,000 training sentences, Spanish to French), as well as less detailed results on additional tasks.

[h] `http://geek.kyloo.net/software/doku.php/mgiza:overview`
[i] `http://nlp.cs.berkeley.edu/Main.html#wordaligner`

<table>
<tr><th>Anymalign</th><th>MGIZA++</th><th>BerkeleyAligner</th></tr>
<tr><td colspan="3">Input: a bilingual parallel corpus</td></tr>
<tr><td rowspan="4">Alignment
(<code>anymalign.py</code>)</td><td>Source-target training
(<code>mkcls</code> + <code>mgizapp</code>)</td><td rowspan="3">Joint training
(<code>berkeleyaligner.jar</code>)</td></tr>
<tr><td>Target-source training
(<code>mkcls</code> + <code>mgizapp</code>)</td></tr>
<tr><td>Symmetrization
(Moses: <code>symal</code>)</td></tr>
<tr><td colspan="2">Alignment extraction (Moses: <code>extract</code>)</td></tr>
<tr><td colspan="3">Scoring (Moses: <code>score</code>)</td></tr>
<tr><td colspan="3">Output: a translation table</td></tr>
</table>

Fig. 8. The three aligners in their respective processing chains. Anymalign does not need any pre- or post-processing tools. We use Moses with BerkeleyAligner only to extract and score alignments. MGIZA++ must be executed twice and its results made symmetric.

6.1. *A first evaluation protocol: machine translation*

Our first evaluation protocol consists in using the translation tables produced by the aligners as the main knowledge source of a phrase-based SMT system, and to evaluate the quality of the translations obtained. We use the Moses toolkit, which we already use to post-process the output of MGIZA++ and BerkeleyAligner. Since the Moses decoder relies on n-grams, we filter out discontinuous alignments produced by Anymalign. The Moses engine is used with its default parameters, and we simply replace its translation table by the one produced by each of the three aligners. We first study the behavior of the aligners according to processing time, using our Spanish-French parallel corpus. MGIZA++ and BerkeleyAligner are executed by varying their number of iterations, from 1 to 5 for each model, and their processing times are measured. Since Anymalign can be stopped at any time, we repeat the same experiment for several execution times, starting from one second. We use 500 pairs of sentences for tuning with MERT[36] and 500 for testing.

In Fig. 9, we plot the TER scores[37] obtained by Moses using the translation tables produced by the three aligners. We evaluate three versions of Anymalign, which allows us to measure the contribution of the two optimizations we proposed: lexical weights and subcorpus size selection optimization. The best results (64% TER) are obtained with MGIZA++ and BerkeleyAligner, the former being slightly faster. Their scores are relatively constant whatever the number of iterations. The scores of the "complete" version of Anymalign converge very quickly to 68% TER, which is worse than the two other aligners. Lexical weights do not improve the results much (by about 0.05% TER). They require more processing time, which explains why the curve without lexical weights is below the "complete" one for the first 30 minutes. The curve without subcorpus size selection optimization is far above all the others,

and does not converge as neatly: the optimized distribution is clearly beneficial.

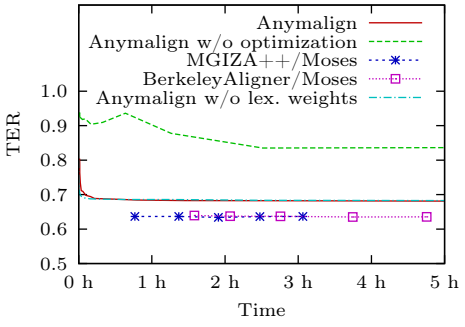

Fig. 9. Behavior of the aligners on a machine translation task. The best TER scores are the lowest. MGIZA++ and BerkeleyAligner produce comparable results. Anymalign produces slightly worse results, and lexical weights do not help much. It performs poorly without optimization of subcorpus size selection. Although this is not a critical issue for an aligner, Anymalign is by far the fastest.

Additional results are presented in Table 4. On tasks using the BTEC[38] as training corpus (short sentences: 10 English words in average), Anymalign is 1 BLEU point below the others on average. On tasks using Europarl as training corpus (longer sentences: 30 English words on average), Anymalign is 2.8 BLEU points below on average.

Table 4. BLEU scores obtained by Moses from Anymalign and MGIZA++'s phrase tables on samples of the BTEC and Europarl corpora in different languages.

Task	Anymalign	MGIZA++
IWSLT 2007 : ja → en	0.46	0.45
IWSLT 2008 : ar → en	0.37	0.41
IWSLT 2008 : zh → en	0.32	0.32
IWSLT 2008 : zh → es	0.25	0.24
Europarl: fr → en	0.25	0.29
Europarl: fr → es	0.32	0.36
Europarl: de → el	0.15	0.16
Europarl: el → de	0.14	0.16
Europarl: en → fi	0.11	0.12
Europarl: fi → en	0.16	0.21

6.2. *A second evaluation protocol: bilingual lexicon induction*

Our second protocol consists in comparing the translation tables to a bilingual reference lexicon, weighting entries by their translation probabilities. As a pre-

processing step, we filter the reference lexicon so that it only contains entries that can actually be produced by the aligners from the training corpus. Practically, an entry is kept if it is a subsequence of a pair of sentences in the training corpus. Then, we compute a score as follows: sum up all source-to-target translation probabilities for those alignments that are found in the reference lexicon, and divide by the number of distinct entries in the reference lexicon. The reason for doing this kind of evaluation is that our method dos not produce alignment links, which are necessary with standard AER evaluations.[39]

Our reference bilingual lexicons come from the XDXF website.[j] Similarly to our first protocol, we first study the behavior of our three aligners according to processing time on our Spanish-French parallel corpus. The results are presented in Fig. 10.

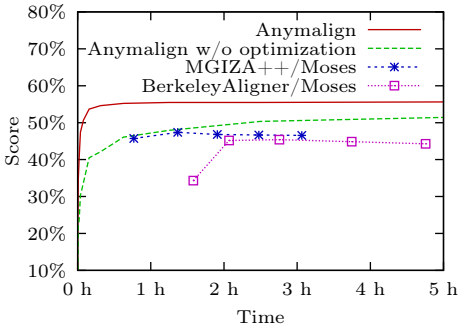

Fig. 10. Behavior of the aligners on a bilingual lexicon induction task. Anymalign produces better results than MGIZA++ and BerkeleyAligner and is much faster. Note that lexical weights are not used in this evaluation: only translation probabilities are taken into account.

Here, Anymalign's results are much better than those of the two other aligners: its score converges very quickly to 56%, while the maximum is only 47% for MGIZA++ and 45% for BerkeleyAligner. In addition, the scores obtained by Anymalign without optimized subcorpus size selection are also better; however it converges much slower. The scores of MGIZA++ and BerkeleyAligner vary very slightly, except from the first to the second iteration of BekeleyAligner.

Additional comparison between the three aligners are presented in Table 5. The Bible corpus[40] was used for this experiment with about 30,000 sentences (29 English words in average) in each of seven languages. When comparing the scores obtained by Anymalign and MGIZA++ on each of the 42 pairs of languages, in average, Anymalign's results are above by 7% relative to those of MGIZA++. On the same tasks, BerkeleyAligner's results are below by 7% relative to those of MGIZA++.

[j] `http://xdxf.sourceforge.net/`

Table 5. Anymalign's gains in score relatively to MGIZA++ on 42 bilingual lexicon induction tasks. An increase of 7% is observed in average.

	dan	eng	fin	fra	spa	swe	zho
dan		+ 3	–10	–15	+ 9	+8	– 4
eng	–15		–10	+13	+ 2	–6	– 5
fin	+ 2	+36		+70	+53	+7	+11
fra	–15	0	– 2		+ 1	–3	+ 5
spa	– 9	+15	+ 3	+13		–2	+15
swe	– 4	+ 7	–18	+19	+ 7		– 1
zho	–13	+16	0	+58	+31	+3	

In summary, Anymalign is below the two statistical tools on phrase-based SMT tasks, while it performs much better at inducing multilingual lexicons. We investigate some causes of this difference in the next section. In any case, Anymalign is much faster since it can produce almost instant results. The subcorpus size selection optimization introduced in Section 5.1 allows to greatly speed up convergence of the results.

6.3. *Discussion and future research*

The previous experiments show that Anymalign produces much better results on bilingual lexicon induction tasks than on phrase-based SMT tasks. Since the bilingual reference lexicons we use mainly consist of unigrams (1.2 words per entry in average), we naturally conclude that Anymalign produces better unigram alignments. To confirm this intuition, we repeat the SMT experiment described in Section 6.1, with the difference that we constrain the decoder to use only unigrams, i.e., the translation tables only contain word pairs. This word-based system is comparable to what was performed before the advent of phrase-based models. The new scores are 68% for Anymalign, 68% for MGIZA++, and 67% for BerkeleyAligner. As expected, the scores are worse than previously: TER increases by 4% for MGIZA++ and BerkeleyAligner. The most noticeable point is that Anymalign's score remains unchanged, and is now comparable to those of the two other aligners, while it was significantly worse in the first experiment. From this experiment, we deduce that a weakness of Anymalign concerns the production of long n-grams.

We thus investigate why our approach does not align long n-grams in sufficient quantity. To this end, we investigate the content of the phrase tables. We are particularly interested in the difference between unigrams and longer n-grams. Therefore, we simply count the number of source n-grams in the translation tables produced by Anymalign and MGIZA++, with n ranging from 1 to 7 (Moses' default maximum phrase length). The results are presented in Fig. 11.

The coverage of Anymalign's phrase table is much better on unigrams: more than 80% of the source vocabulary is covered. However, it is far behind for all remaining n-gram lengths. Experiments have shown that the larger the input corpus, the more

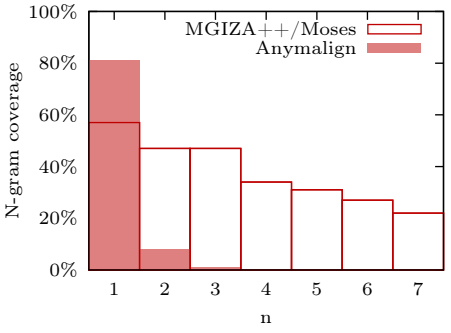

Fig. 11. Coverage of the French part of our Europarl corpus by MGIZA++ and Anymalign's phrase tables. Unigram coverage is neatly higher with Anymalign. It is however much lower on all remaining n-gram lengths, by far.

noticeable the differences. This suggests that the reason why a phrase-based MT system would be less efficient when built on top of our alignment method would not be a matter of *quality*, but rather of *quantity*: the method simply does not align n-grams with $n \geq 2$ in sufficient quantity.

In fact, the reason why long n-grams are not extracted in sufficient number by our method is related to the frequency of the words they are made of. Since the method consists in extracting sequences of words that share the same distribution, including those that come from the same language, n-grams made of words of different frequencies are difficult to extract. And since most n-grams consist of words with different frequencies (e.g., determiner + noun, pronoun + verb, etc.), most n-grams may not be extracted.

Our current research focuses on improving the results of SMT systems built on top of our alignment method. We have started to investigate two directions:

- improving the coverage of long n-grams. For this, we are generalizing our method so that the indexation phase does not rely solely on words anymore, but also on overlapping n-grams of various lengths. Our first results have shown a major increase in n-gram coverage (up to ×10 n-grams), yielding a boost of more than 4 BLEU points on some standard Europarl SMT tasks.
- combining Anymalign's phrase table with MGIZA++'s or other statistical tools. Since these aligners work in a very different way, their outputs are quite different. For instance, more than 30% of the bigrams produced by Anymalign in the previous experiment were not present in MGIZA++'s phrase table. Taking their union, and possibly giving more credit to their intersection, could result in a larger and more accurate phrase table.

7. Conclusion

In this paper, we have proposed a multilingual sub-sentential alignment method based on observations on the use of rare words in sub-sentential alignment. We have shown that the majority of the vocabulary of a text consists of rare words (typically 40% hapaxes) that appear in very few sentences (less than 10% sentences for hapaxes), and this data sparseness is precisely the reason why they are so simple to align. These observations allowed us to design a new sub-sentential alignment method. It can process any number of languages simultaneously (e.g. we could align 60 languages of KDE system messages[41]), is very simple, anytime, and allows for massive parallelism. Contrary to most preconceptions, we have shown that it is possible to safely use rare words as a basis of an NLP task, by *removing* input data. Our processes are *alingual*, which allows for multilingual, bilingual, and even monolingual processing. Our implementation, Anymalign, is free software and is available at `http://users.info.unicaen.fr/~alardill/`. It is competitive with state-of-the-art tools (in average: −2 BP in SMT tasks, but +7% in bilingual lexicon induction). Our current research focuses on improving its results in SMT, mainly by improving coverage for long n-grams.

References

1. P. Brown, J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, R. Mercer, and P. Roossin. A Statistical Approach to Language Translation, in *Proceedings of the 12th International Conference on Computational Linguistics (Coling'88)*, Budapest, pp. 71–76, 1988.
2. W. Gale and K. Church. Identifying Word Correspondences in Parallel Texts, in *Proceedings of the fourth DARPA workshop on Speech and Natural Language*, Pacific Grove, pp. 152–157, Feb. 1991.
3. L. R. Dice. Measures of the Amount of Ecologic Association Between Species, *Ecology*, **26**(3), pp. 297–302, 1945.
4. P. Fung and K. Church. K-vec: A New Approach for Aligning Parallel Texts, in *Proceedings of the 15th International Conference on Computational Linguistics (Coling'94)*, **2**, Kyōto, pp. 1096–1102, Aug. 1994.
5. D. Melamed, Models of Translational Equivalence among Words, *Computational Linguistics*, **26**, pp. 221–249, June 2000.
6. R. Moore. Association-Based Bilingual Word Alignment, in *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, Ann Arbor, pp. 1–8, June 2005.
7. F. Och and H. Ney. A Systematic Comparison of Various Statistical Alignment Models, *Computational Linguistics*, **29**, pp. 19–51, Mar. 2003.
8. P. Brown, S. Della Pietra, V. Della Pietra, and R. Mercer. The Mathematics of Statistical Machine Translation: Parameter Estimation, *Computational Linguistics*, **19**(2), pp. 263–311, 1993.
9. Q. Gao and S. Vogel. Parallel Implementations of Word Alignment Tool, in *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, Columbus (Ohio, USA), pp. 49–57, June 2008.
10. D. Tufiş and A.-M. Barbu. Lexical token alignment: experiments, results and applications, in *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas de Gran Canaria, pp. 458–465, 2002.
11. P. Liang, B. Taskar, and D. Klein, Alignment by Agreement. in *Proceedings of the Human Language Technology Conference of the NAACL*, New York City, pp. 104–111, June 2006.
12. G. Zipf. *The Psycho-Biology of Language: An Introduction to Dynamic Philology.* Classic Series, Cambridge, USA: The MIT Press, 1965. Fist edition 1935.

13. B. Mandelbrot. Structure formelle des textes et communication, *Word*, **10**, pp. 1–27, 1954.
14. M. Montemurro. A generalization of the Zipf-Mandelbrot Law in Linguistics, *Nonextensive Entropy: interdisciplinary applications*, 2004. 12 pages.
15. P. Koehn., Europarl: A Parallel Corpus for Statistical Machine Translation, in *Proceedings of the tenth Machine Translation Summit (MT Summit X)*, Phuket, pp. 79–86, Sept. 2005.
16. L. Ahrenberg, M. Andersson, and M. Merkel. A Simple Hybrid Aligner for Generating Lexical Correspondences in Parallel Texts, in *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (ACL-Coling 98)*, **1**, Montreal (Quebec, Canada), pp. 29–35, Aug. 1998.
17. R. Moore. On Log-Likelihood-Ratios and the Significance of Rare Events, in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, Barcelona, pp. 333–340, 2005.
18. F. Cromières. Sub-Sentential Alignment Using Substring Co-Occurrence Counts, in *Proceedings of the COLING/ACL 2006 Student Research Workshop*, Sydney, pp. 13–18, July 2006.
19. E. Giguet and P.-S. Luquet. Multilingual Lexical Database Generation from Parallel Texts in 20 European Languages with Endogenous Resources, in *Proceedings of the Coling/ACL 2006 Main Conference Poster Sessions*, Sydney, pp. 271–278, July 2006.
20. B. Schrader. How does morphological complexity translate? A cross-linguistic case study for word alignment, in *International Conference on Linguistic Evidence 2006*, Tübingen, Feb. 2006. 3 pages.
21. E. Nishimoto. Defining New Words in Corpus Data: Productivity of English Suffixes in the British National Corpus, in *Proceedings of the 26th Annual Meeting of the Cognitive Science Society (CogSci 2004)*, Chicago, Aug. 2004. 6 pages.
22. S. Evert and A. Lüdeling. Measuring morphological productivity: Is automatic preprocessing sufficient?, in *Proceedings of the Conference on Corpus Linguistics 2001 (CL2001)*, Lancaster (UK), pp. 167–175, 2001.
23. B. Cartoni. Constance et variabilité de l'incomplétude lexicale, in *Actes de la 13e conférence sur le Traitement Automatique des Langues Naturelles (TALN/RECITAL 2006)*, Leuven, pp. 661–669, Apr. 2006.
24. P. Langlais, F. Gotti, and G. Cao. NUKTI: English-Inuktitut Word Alignment System Description, in *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, Ann Arbor, pp. 75–78, June 2005.
25. H. Baayen and R. Sproat. Estimating lexical priors for low-frequency morphologically ambiguous forms, *Computational Linguistics*, **22**, pp. 155–166, June 1996.
26. R. H. Baayen, *Word Frequency distributions*. Kluwer, 2001.
27. Y. Shinyama and S. Sekine. Named Entity Discovery Using Comparable News Articles, in *Proceedings of the 20th International Conference on Computational Linguistics (Coling'04)*, Geneva, pp. 848–853, Aug. 2004.
28. M. Lafourcade and C. Boitet. UNL lexical selection with conceptual vectors, in *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas de Gran Canaria, pp. 1958–1964, 2002.
29. P. Turney and M. Littman. Corpus-based Learning of Analogies and Semantic Relations, *Machine Learning*, **60**, pp. 251–278, Sept. 2005.
30. I. Cicekli and H. A. Güvenir. Learning Translation Templates from Bilingual Translation Examples, *Applied Intelligence*, **15**, pp. 57–76, 2001.
31. P. Koehn, F. Och, and D. Marcu. Statistical Phrase-Based Translation, in *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2003)*, Edmonton, pp. 48–54, 2003.
32. J. Sinclair and J. Ball. Preliminary Recommendations on Text Typology, tech. rep., Expert Advisory Group on Language Engineering Standards (EAGLE), June 1996. 36 pages.
33. M. Simard, Text-translation Alignment: Three Languages Are Better Than Two, in *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC)*, (College Park), pp. 2–11, 1999.
34. Y. Al-Onaizan, J. Curin, M. Jahr, K. Knight, J. Lafferty, D. Melamed, F. Och, D. Purdy, N. Smith, and D. Yarowsky. Statistical Machine Translation: Final Report, tech. rep., Johns

Hopkins University 1999 Summer Workshop (WS 99) on Language Engineering, Center for Language and Speech Processing, Baltimore, 1999. 42 pages.
35. 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 the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, Prague, pp. 177–180, June 2007.
36. F. Och, Minimum Error Rate Training in Statistical Machine Translation, in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, (Sapporo), pp. 160–167, July 2003.
37. M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. A Study of Translation Edit Rate with Targeted Human Annotation, in *Proceedings of the 7th Conference of the Association for Machine Translation of the Americas (AMTA 2006)*, Cambridge, pp. 223–231, Aug. 2006.
38. T. Takezawa, E. Sumita, F. Sugaya, H. Yamamoto, and S. 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)*, Las Palmas de Gran Canaria, pp. 147–152, 2002.
39. F. Och and H. Ney. A Comparison of Alignment Models for Statistical Machine Translation, in *Proceedings of the 18th International Conference on Computational Linguistics (Coling'00)*, Saarbrücken, pp. 1086–1090, Aug. 2000.
40. P. Resnik, M. B. Olsen, and M. Diab. The Bible as a Parallel Corpus: Annotating the "Book of 2000 Tongues", *Computers and the Humanities*, **23**(1-2), pp. 129–153, 1999.
41. J. Tiedemann. News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces, *Recent Advances in Natural Language Processing*, **5**, pp. 237–248, 2009.

Adrien Lardilleux

He received the Ph.D. degree in 2010 from the University of Caen Basse-Normandie, France. He is currently a post-doctoral researcher in the Spoken Language Processing group at LIMSI/CNRS. His research interests concern natural language processing, in particular machine translation and sub-sentential alignment.

Yves Lepage

He received his D.E.A. and Ph.D. degrees in 1989 from Grenoble university, France, in GETA under the supervision of Prof. Vauquois and Prof. Boitet. After a postdoctorate at ELSAP, university of Caen and EDF, Paris, he joined ATR labs, Japan, where he worked as an invited researcher and a senior researcher until 2006. In 2003 he got the habilitation for his habilitation thesis entitled "Of this kind of analogies that renders an account in linguistics". In October 2006, he got the qualification for full professorship from the National Board of French Universities in both linguistics and computer science and became full professor at the University of Caen Basse-Normandie in October 2006. He joined Waseda University, graduate school of Information, Production and Systems in April 2010. His research interests are in Natural Language Processing, Machine Translation, and in particular Example-Based Machine Translation. He is a member of the French and the Japanese Natural Language Processing Associations. He is a member of the board of the French Natural Language Processing Association, ATALA, and one of the four editors-in-chief of the French journal on Natural Language Processing, TAL.

François Yvon

He is Professor in Computer Science at the University of Paris-Sud 11 and member of the Spoken Language Processing group of the LIMSI/CNRS. He was previously (1996-2007) associate professor in the department of computer science at Telecom ParisTech. François Yvon holds a Ph.D in Computer Science and engineering degrees from the Telecom ParisTech and an engineering degree from the Ecole Polytechnique. His main research interests include analogy-based and statistical language learning, speech recognition and synthesis, and statistical machine translation.