Machine Translation manuscript No.
(will be inserted by the editor)

Generalizing Sampling-Based Multilingual Alignment

Adrien Lardilleux · François Yvon · Yves Lepage

Received: date / Accepted: date

Abstract Sub-sentential alignment is the process by which multi-word translation units are extracted from sentence-aligned multilingual parallel texts. This process is required, for instance, in the course of training statistical machine translation systems. Standard approaches typically rely on the estimation of several probabilistic models of increasing complexity and on the use of various heuristics, that make it possible to align, first isolated words, then, by extension, groups of words. In this paper, we explore an alternative approach, originally proposed by Lardilleux and Lepage (Proceedings of AMTA 2008, pp 125–132), which relies on a much simpler principle: the comparison of occurrence profiles in sub-corpora obtained by sampling. After analyzing the strengths and weaknesses of this approach, we show how to improve the detection of multi-word translation units and evaluate these improvements on machine translation tasks.

Keywords association measures · sub-sentential alignment · phrase-based machine translation

1 Introduction

Sub-sentential alignment consists in extracting multi-word translation units from sentence-aligned multilingual parallel texts, i.e. texts whose sentences have been related to their translations. This task constitutes the first step in the process of training data-driven machine translation systems, be they statistical or example-based. The most prominent approach

A. Lardilleux
LIMSI-CNRS, BP 133, Orsay Cedex, France
E-mail: adrien.lardilleux@limsi.fr

F. Yvon
LIMSI-CNRS & University Paris Sud, BP 133, Orsay Cedex, France

Y. Lepage
Waseda University, Graduate School of Information, Production and Systems,
808-0135 Hibikino 2-7, Wakamatu-ku, Kitakyuusyuu-si, Fukuoka-ken, Japan

nowadays is phrase-based statistical machine translation, where the core model is a translation table derived from sub-sentential mappings. This table consists in a pre-computed list of translation pairs, where each (*source*, *target*) phrase[1]pair is associated with a number of numerical values loosely reflecting the likelihood that *source* translates to *target*.

The problem of learning sub-sentential mappings (e.g. word-to-word or phrase-to-phrase associations) from sentence-aligned text corpora is far from new, and a great number of proposals have been put forward to perform this task. From a bird's eye view, methods for learning associations fall into two main categories: on the one hand, the *alignment-based* tradition, introduced by Brown et al (1988), addresses the problem of identifying *alignments* between words or groups of words in parallel sentences. This approach consists in building a statistical model of the parallel text, the parameters of which are estimated through a global maximization process which simultaneously considers all the possible associations that could exist in the parallel corpus. In practice, the goal is to determine the best set of *correspondences* (or alignment links) between source and target words in each parallel sentence pair. The most famous representatives in this category are the IBM alignment models (Brown et al 1993b) for aligning isolated words, which have given rise to an impressive series of variants and amendments (see e.g. work by Vogel et al (1996); Wu (1997); Deng and Byrne (2005); Liang et al (2006); Fraser and Marcu (2007); Ganchev et al (2008), to cite a few). Generalizing models of word alignments to models for phrase alignments proves to be a much harder problem, and in the view of the deficiencies or inefficiencies of the proposals of Marcu and Wong (2002) and of Vogel (2005), such alignments are derived by heuristically combining asymmetric word to phrase alignments obtained separately in both directions (Koehn et al 2003). Based on these alignment links, it is finally possible to evaluate the quality of each extracted pair of phrases.

On the other hand, *associative* approaches, introduced by (Gale and Church 1991), do not require the use of an alignment model. In order to detect translational equivalence, they use standard statistical measures of independence such as, for instance, the Dice coefficient, the pointwise mutual information (Cover and Thomas 1991), the mutual information (Gale and Church 1991; Fung and Church 1994), or the likelihood ratio (Dunning 1993)—see also more recent work by Melamed (2000) and Moore (2005). As the number of potential word associations grows more than linearly with the corpus size, testing all possible word associations is computationally costly, and as far as we know, nobody ever tried to evaluate all the possible phrase to phrase associations. Instead, the list of associations to be evaluated is computed using predefined patterns and filters (e.g. restricting the computation to the most frequent word n-grams). In this approach, the process is a local maximization process, where each segment is processed independently from the others. This approach generally allows the direct extraction of scored translation pairs. Work in this family includes (Gale and Church 1991), which has been since extended to non strictly parallel corpora (Fung and Church 1994; Fung and Yee 1998), to learn multi-word and term associations (Dagan and Church 1994; Gaussier and Langé 1995; Smadja et al 1996), and to alternative association measures, such as the G^2 (Gale and Church 1991) and the ϕ^2 (Dunning 1993; Moore 2004, 2005) statistics.

The alignment-based approach is the most widely used, mainly due to its tight integration within the statistical machine translation (SMT) framework, of which it constitutes a cornerstone since the introduction of the IBM models (Brown et al 1993b). Note that both

[1] As is now standard in machine translation circles, we use here the word "phrase" to denote a contiguous sequence of words of arbitrary length; we are aware that this departs from the more linguistic acceptation of the term.

approaches show complementary strengths and weaknesses, as acknowledged e.g. by the work of Johnson et al (2007), where phrase associations extracted as a result of a word alignment are then filtered out using statistical association measures.

We have recently proposed an associative sub-sentential alignment method (Lardilleux and Lepage 2008, 2009; Lardilleux 2010), which addresses a number of issues that are often overlooked. In particular, this method allows the processing of multiple languages simultaneously, and does not make any distinction between source and target; it is also amenable to massive parallelism, scales easily, and is very simple to implement. On average, this method has been shown to produce better results than state-of-the-art tools on bilingual lexicon induction tasks in terms of recall, but seems to be less efficient on phrase-based machine translation tasks (Lardilleux et al 2009) in terms of BLEU scores (Papineni et al 2002). So far, we have only proposed *hypotheses* to explain these apparently contradictory results. In this paper, we propose a more detailed analysis of our algorithm, in order to determine the origin of these differences. Based on this analysis, we introduce a generalization, resulting in significant improvements of the performance in machine translation tasks.

This paper is organized as follows: Section 2 is devoted to an overview of the original alignment method; Section 3 describes experiments to evidence the origin of its weaknesses and its possible complementarity with IBM models; in Section 4, we propose a generalization and evaluate its performance on several tasks; Section 5 then shows to what extent the method can be used to complement other word alignment tools; Conclusions and further prospects are discussed in Section 6.

2 Anymalign: an Overview

2.1 Main Principles

Our alignment method can be viewed as an emulation of associative methods, generalized beyond the extraction of mere pairs (*source*, *target*) of isolated *words*.[2] Indeed, it considers possibly discontinuous *sequences of words* of various sizes for extraction, which makes it useful for handling tricky phenomena such as separable phrasal verbs in English (see example below, which occurs in the Europarl corpus (Koehn 2005)). The main difference with traditional alignment methods is that it focuses exclusively on words that strictly occur in the same sentences of the input parallel corpus, irrespective of the language. Let us assume for instance, that in a French–English sentence-aligned parallel corpus, each occurrence of the French word *noté* is paralleled with an occurrence of the English words *written* and *down*. Assume furthermore that no other word (or word group) has the same occurrence profile as *noté* (in French) and *written* and *down* (in English), then the method will conclude that *written … down* and *noté* are mutual translations. Of course, this simple-minded principle cannot be expected to detect many associations because the number of words that have exactly the same distribution over entire parallel corpora is very limited. Instead of considering the complete input corpus as a whole, our method processes a large number of small sub-corpora, the key idea being that the smaller the corpus, the more likely it is to contain words that share the same distribution, which will then be identified by our method.

The core of the algorithm thus consists in extracting associations from multiple independent sub-corpora built by sampling. In practice, we use a distribution on the sub-corpus

[2] In this paper, the term “word” refers to a string of non-blank characters delimited by blank characters, as produced by any suitable tokenization program.

Table 1 Stages of the alignment algorithm. The numbers within brackets correspond to the numbered steps in Figure 1

```
(2) Concatenate parallel sentences from the input corpus, output one line for each input tuple of lines
Initialize an associative array of AlignmentCounts
Do
    (2) Select a sub-corpus $C$ by sampling
    (3) Index words by their vector of presence in the sentences of $C$
    (3) Words with identical distribution are clustered into a same group
    For each group of words:
        For each sentence where group appears :
            (4) Restore word order in group
            (4) AlignmentCounts[group] ++
            (4) AlignmentCounts[sentence – group] ++
Until allotted time used up or no more alignment obtained or any other criterion (or interruption by user)
(5) Compute alignment scores
```

size which concentrates most of its mass on very small sub-corpora. Indeed, small corpora are faster to process, ensure a quicker convergence toward accurate association scores, and have been shown to produce better results (Lardilleux 2010). For each sequence of words that share the same distribution in a sub-corpus, we extract two alignments: the sequence itself, and its complementary part in the sentence it appears in. The number of sub-corpora to be processed does not need to be defined in advance: the process stops according to some criteria such as the amount of elapsed time or the attained input corpus coverage; it can also be externally halted at any time by the user, and yet deliver a usable set of associations. Because the sub-corpus selection is random, the entire coverage of the initial training corpus is only guaranteed in the limit; however, the more sub-corpora are processed, the larger is the coverage of the input corpus, and the more accurate are the association measures. At the end of a run, all the alignments that have been extracted are evaluated by various numerical scores (translation probabilities and lexical weights (Koehn et al 2003)), according to the number of times they were produced (refer to Lardilleux and Lepage (2009) for details on how these computations are performed). The result is a translation table that can be directly plugged, for instance, in machine translation systems, after appropriate filtering, e.g. discarding discontinuous sequences for phrase-based systems.

2.2 Algorithmic details

Figure 1 illustrates the main stages of the method on an example trilingual text. The complete extraction algorithm is summarized in Table 1. In this description, the notation $x - y$, where x is a sentence and y a possibly discontinuous sub-part of x, denotes the complementary sub-part of y in x.

In the rest of this paper, where we primarily focus on machine translation applications, we will restrain ourselves to consider bilingual corpora, even though the ability to simultaneously extract association in more than two languages is certainly one appealing aspect of the method.

(1) Input: a multilingual parallel corpus, here Arabic–French–English.

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 .

⇓

(2) Parallel sentences are concatenated,
distinguishing words according to their original language
(here, using subscripts: 1 for Arabic, 2 for French, and 3 for English).
Selection of a random sub-corpus (here, the first three lines of the original corpus).

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$_3$.$_3$

⇓

(3) Indexation of words (compute distribution profile).
Words with identical distributions are clustered together.

	.$_1$	.$_2$	.$_3$	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$_1$	. . .
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	. . .
2	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	1	1	. . .
3	1	1	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	. . .

⇓

(4) Each group of words allows to extract two alignments
for each sentence where it appears.

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

⇓

(5) Count alignments and restore boundaries between languages.

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
		⋮			

Fig. 1 An illustration of the alignment method. In the following, we will generalize Step (3) (indexation and clustering of words)

2.3 Results

In this section, we summarize the main results and conclusions of (Lardilleux 2010). This alignment method has been evaluated on two kinds of tasks: phrase-based statistical machine translation and bilingual lexicon induction. Our implementation, *Anymalign*,[3] is compared with a phrase extraction standard pipeline using MGIZA++[4] (Gao and Vogel 2008), which is the most recent implementation of IBM models for word alignments, and tools from the Moses toolkit (Koehn et al 2007) for the extraction and scoring processes. In practice, because Anymalign may be stopped at any time, we start by executing MGIZA++ (in two directions) with its default parameters, measure its execution time, and run Anymalign for the same amount of time. The parallel corpora used in these experiments are Europarl (Koehn 2005) and excerpts of the BTEC (Takezawa et al 2002), distributed during the IWSLT machine translation evaluation campaigns (Fordyce 2007). The BTEC excerpts are made of 20,000 to 40,000 pairs of short aligned sentences (10 English words on average) and our Europarl corpus is made of 100,000 pairs of long sentences (30 English words on average).

As for the phrase-based statistical machine translation task, we compare the performance of Moses (Koehn et al 2007) with its default translation table, built from MGIZA++'s alignments, with the same system using a phrase table built with Anymalign. All phrase tables contain five common features: translation probabilities in two directions, lexical weights in two directions, and length penalty. For these experiments, we do not make use of lexical reordering and only use the default distance-based reordering model. On average, Anymalign has been found to be about two BLEU points worse than the baseline. In the best case, we observed a gain of one point in comparison with MGIZA++ (BTEC, Japanese–English); in the worst case, a loss of eight points (Europarl, Finnish–English). Overall, the differences are more pronounced with Europarl than with the BTEC.

As for the bilingual lexicon induction task, we compare the translation tables produced by the two aligners with a reference bilingual lexicon.[5] This lexicon is first filtered so as to discard pairs that do not appear in any parallel sentence, i.e. which cannot be extracted from the parallel corpus by the aligners. The vast majority of the resulting dictionary entries are simple words (considering all dictionaries, the average number of words per entry is 1.2), but a few long n-grams (up to $n = 7$) are included as well. Words in the reference can be ambiguous, i.e. a given entry may have several translations. We then define a translation table's score S, relative to this filtered reference lexicon R, as the sum of the source to target translation probabilities of those translation pairs in the translation table that are attested in the reference, divided by the number of distinct source entries in the reference:

$$S = \frac{1}{|R_S|} \sum_{(f,e)\in R} \mathrm{p}(e|f)$$

where R_S is the set of distinct source entries in the reference bilingual lexicon R and $\mathrm{p}(e|f)$ the probability that source word f translates to target word e, as reported in the translation table. If the pair (e,f) is not present in the translation table, we set $\mathrm{p}(e|f) = 0$.

Because it compares the correct outputs of the systems (more precisely, all outputs weighted by their translation probabilities) against a subset of all correct translation pairs

[3] http://www.limsi.fr/Individu/alardill/anymalign/
As of version 2.5, Anymalign implements the generalization described in Sec. 4.

[4] http://geek.kyloo.net/software/doku.php/mgiza:overview
In all subsequent experiments, we use MGIZA++ with its default parameters, which typically yield good results. We run 5 iterations of IBM model 1, 5 iterations of the HMM, and 5 iterations of models 3 and 4.

[5] Our lexicons come mainly from the XDXF website: http://xdxf.sourceforge.net

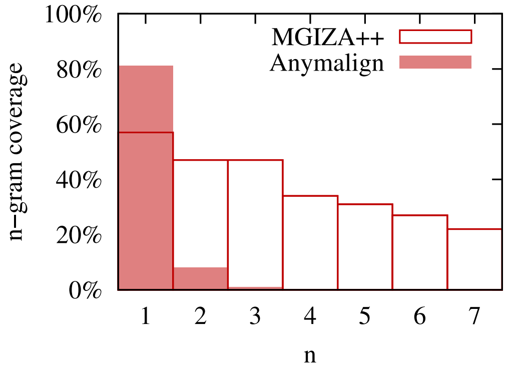

Fig. 2 Coverage of the source part of a sample of the French–English Europarl corpus by MGIZA++ and Anymalign's translation tables. Anymalign finds more associations for unigrams than the baseline, but much less for longer *n*-grams

extractable from the parallel corpus, the result is to be interpreted as a recall, in the range $[0,1]$, and is thus an indicator of lexical coverage. In our experiments, we found that, on average, Anymalign is better by 7% relatively to MGIZA++. In the best case, we obtained a relative gain of 70% (Europarl, Finnish–French); in the worst, a loss of 18% (Europarl, Swedish–Finnish). The genre of texts that constitute the corpus does not seem to have a major influence on these scores.

In summary, our method is less efficient on phrase-based machine translation tasks, but produces better *word* mappings, as demonstrated by the results of a comparison with reference lexicons. As showed in previous work (Lardilleux et al 2009), this is mainly due to the limited capacity of our method to produce word *n*-gram alignments with $n \geq 2$, as illustrated in Figure 2. The goal of next section is to analyse the origin of these discrepancies.

3 A reanalysis of Anymalign

In this section, we present experiments showing that the above-mentioned, and somewhat paradoxical, results have two main causes: (1) our algorithm fails to extract groups containing both rare and frequent words, and (2) our algorithm behaves poorly on frequent words, which are the most likely to be useful in machine translation tasks. The experiments related in this section are performed on an excerpt of about 350,000 sentences from Europarl, using the Portuguese–Spanish language pair (extreme case of related languages in our experiments) and the Finnish–English language pair (extreme case of distant languages). Table 2 presents the number of words in each part of these corpora.

3.1 Rare and frequent words

As previously noted (Lardilleux et al 2009), the constraint that two words should have an identical distribution to be extracted prevents the method from extracting sequences made of words with different frequencies. For instance, a bigram composed of one frequent word and one hapax (made of, for instance, of a (frequent) determiner and a (rare) noun) can hardly

Table 2 Characteristics of corpora used for our analysis

Language	Number of words (tokens)	Vocabulary size
Portuguese	9,249,177	87,341
Spanish	9,330,199	85,366
Finnish	6,472,649	274,958
English	8,955,995	53,704

be extracted because, assuming the frequent word appears in most input corpus sentences, it would require sampling a very small sub-corpus to turn this word into a hapax, which is needed to extract both words as a sequence. However, in such a sub-corpus, most other words would also be hapaxes, and the extracted sequence would not only consist of the two words of interest, but it would also contain all the other hapaxes in the same sentence.

We further analyze this fact by studying the relationship between the frequency of a word and the size of sub-corpora from which it can be extracted. Given a source word w_s, three situations are possible:

1. in a sub-corpus that is "too small," other source words have the same distribution as w_s. Thus, w_s cannot be extracted separately.
2. in a sub-corpus whose size is "ideal," no other source word has the same distribution as w_s, and at least one target word has the same distribution. w_s can thus be extracted separately.
3. in a sub-corpus that is "too large," no other source or target word has the same distribution as w_s. In this case, w_s cannot be extracted at all.

There thus exists a range of sub-corpus sizes which enables a word to be extracted as a single unit. Of course, this range depends on the word to be extracted, and particularly on its frequency. We have empirically determined such ranges by measuring, for each source word in a parallel corpus, the average sub-corpus size needed to align it separately, as well as the average size from which it cannot be aligned at all anymore. To this end, for each word w_s we start by randomly drawing a sub-corpus consisting of one single sentence containing w_s, and test whether w_s can be aligned in it, i.e whether its frequency in the sentence differs from that of all the other words. In most cases, all source and target words will occur exactly once and thus share the same distribution, with the exception of some frequent function words that are likely to appear several times. For most words, the starting situation is thus case 1 in the enumeration above. The process is then iterated, by adding at each step one new randomly chosen sentence to the sub-corpus, and testing whether w_s can be extracted from this enlarged corpus. The process stops when no target word shares the same distribution as the source word under focus.

Each experiment produces two numbers: the minimum size from which the word can be aligned in isolation (transition from case 1 to case 2 above), and the minimum size from which the word cannot be aligned at all (transition from case 2 to case 3). This test is repeated 1,000 times for each source word, and we take the average of measures collected on all 1,000 drawings, so that each word is characterized by two average sub-corpus sizes. The results are plotted on Figure 3.

These graphs suggest two types of conclusions. First, the range of "ideal" sub-corpus sizes, i.e. the middle zone's width, greatly varies from one language pair to another. Note that the logarithmic scale makes this range seem more narrow than it actually is: the aver-

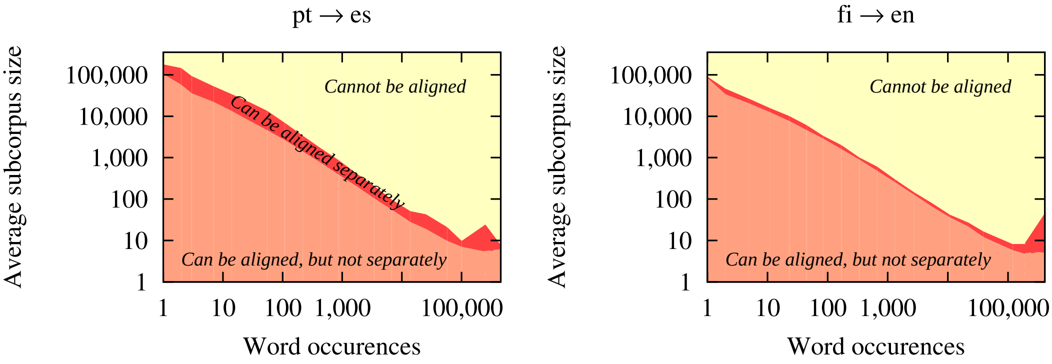

Fig. 3 Average sub-corpus sizes from which a source word can be extracted according to its frequency. In the lower zone, the word cannot be extracted separately (case 1). In the middle zone, the word can be extracted separately (case 2). In the upper zone, the word cannot be extracted at all (case 3). The slight shift of the upper limit on the right extremities of the two graphs is due to the full stop (assimilated to a word in our experiences), which is easier to extract than the other frequent words: it can be extracted separately in sub-corpora made up of about 5 to 80 sentences

age ratio between its upper and lower limits, leaving aside the logarithmic scale, is 2.2 for the Portuguese–Spanish pair, and 1.2 for the Finnish–English pair. This difference is easily explained by the differences of language morphology within each language pair. Thus, we might expect that extracting a given word by Anymalign will require the processing of more sub-corpora for the Finnish–English pair than for the Portuguese–Spanish pair, since it is less likely to randomly draw a sub-corpus having the “right size.”

The second remark is of the utmost interest within the scope of this paper: the more frequent a word, the smaller the sub-corpora from which it is extracted, and conversely. Rare words (left part of the graphs) are therefore aligned from large sub-corpora, while frequent words (right part of the graphs) are aligned from small sub-corpora. This is, for instance, the case of the the comma, which can be aligned in sub-corpora made up of 5 to 9 sentences. A consequence is that a bigram made up of a frequent word (say, a determiner) followed by a rarer word (say, a noun) can not be extracted, as it would require sub-corpora that are both small and large. This partly explains why Anymalign is unable to align long *n*-grams, as illustrated by Figure 2. We will propose an alternative in the next section.

3.2 Useful words

The second explanation for the discrepancies of Anymalign’s results observed on the two tasks discussed above can be easily explained by the fact that these tasks have a different notion of what a *useful* association actually is.

Our method and IBM models rely on opposite intuitions: the former takes advantage of word rarity in order to align them (the number of occurrences of every word is artificially and temporarily decreased by considering sub-corpora), while the latter are estimated as a result of a global optimization process, based on observations measured on the whole corpus. A consequence is that Anymalign is better at aligning words having a small number

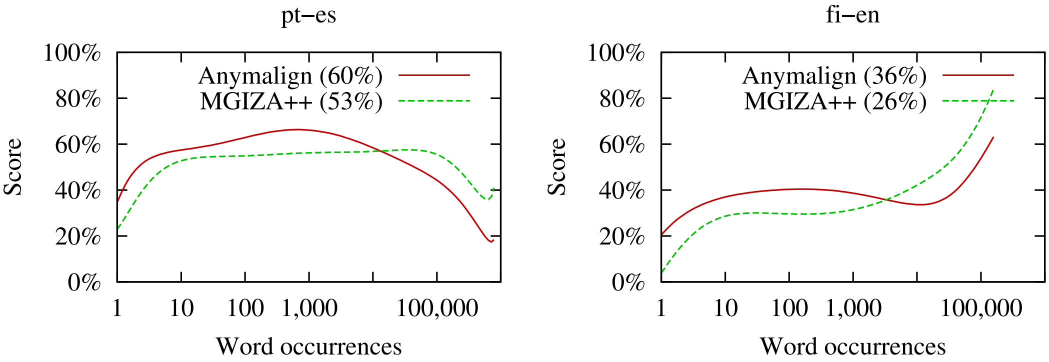

Fig. 4 Scores obtained by Anymalign and MGIZA++'s translation tables on the bilingual lexicon induction task. The scores within parentheses are global, calculated as described in the 3rd paragraph of Section 2.3. The curves present the detail of these scores, according to the number of occurrences of each alignment's source word: one score has been calculated locally for each word frequency. The curves have been smoothed for the sake of readability

of occurrences, while MGIZA++ is better at aligning more frequent ones. This is illustrated on Figure 4.

Here, we are more interested in the relative position of the curves than in their general shape: the curve for Anymalign is above that of MGIZA++ for words of about 1 to 5,000 occurrences, and below for more frequent words. This shows that Anymalign is better at extracting not only rare words, but also medium frequency words. This observation has been corroborated on other language pairs (de–en, es–en, fr–en).

However, the number of distinct rare word types is typically much higher in any text—cf. Zipf's law (Zipf 1965; Mandelbrot 1954; Montemurro 2004)—, with all the more reason in our parallel corpus as well as in the translation tables produced. And since the evaluation metrics (comparison with reference lexicons) considers word *types* rather than word *tokens*, we might expect that our method should obtain better scores on bilingual lexicon induction tasks, because the words it aligns best are overall the most numerous.

In contrast, the number of distinct frequent word types is typically much lower, but they are all the more important in machine translation because a frequent word is much more likely to reappear in the future than a rare word. This may contribute, to some extent, to explain Anymalign's low scores in machine translation. Ideally, we would like to be able to choose between alignments produced by one or the other aligner depending on the frequencies of source or target words, a goal that may be achieved, for instance, by combining the translation tables. We will further investigate this issue in Section 5, and we concentrate, for now, on aligning together words having different frequencies.

4 A generalization to arbitrary word strings

In this section, we present a generalization of our method, aiming primarily at improving its performance on phrase-based statistical machine translation tasks. This generalization retains most hypotheses of the original method: in particular, it only considers surface forms, and does not require any other resource than the input corpus. Our main objective will be to extract more word n-grams with $n \geq 2$ (see Figure 2), which means mitigating the problem of extracting together words with different frequencies (see Sec. 3.1). The following extension is already integrated in the current version of Anymalign.

4.1 Improving the indexation step

We introduce the possibility to work at a variable resolution by indexing n-grams rather than simple words. We do so without choosing a specific sentence segmentation scheme, for instance into syntactic chunks; instead we index all the (possibly) overlapping n-grams in a sentence. Consider, for instance, the following monolingual sub-corpus, made up of three sentences:

1 a b c
2 a b d e
3 a c

Here, we only display a single language for simplicity, but in the case of a multilingual corpus, we just repeat the same process for all languages in parallel (n-grams cannot span several languages). The indexation step of all n-grams in this corpus, *before* collecting the groups of same distribution that serve as a basis to extract alignments, produces the following result:

	n = 1					n = 2					n = 3			n = 4
	a	b	c	d	e	ab	ac	bc	bd	de	abc	abd	bde	abde
1	1	1	1	0	0	1	0	1	0	0	1	0	0	0
2	1	1	0	1	1	1	0	0	1	1	0	1	1	1
3	1	0	1	0	0	0	1	0	0	0	0	0	0	0

The next step of the algorithm, i.e. the identification of groups of words having the same distribution, is accordingly modified as follows. When n-grams sharing the same distribution are overlapping, then the resulting group of words is made up of the "union" of those n-grams. For instance, bigrams *bd* and *de* in the table above share the same distribution; they will be collapsed together into the larger group *bde*. In other words, groups are no longer made of words with identical distribution; instead, they correspond to maximally large n-grams with identical distribution. A given word may henceforth appear in several groups, which was not the case in the original method.

This modification raises a new problem: n-grams may block out $(n-1)$-grams, and do so recursively. In the example above, this happens for the unigram *b*, which is masked by the larger bigram *ab* which has the same distribution. As a result, *b* can no longer be extracted separately, which would be the case if we would only look at unigrams. To get the best of both worlds, it thus seems necessary to repeat the process independently for each value of n.

4.2 Three clustering strategies

In this section, we test different approaches for introducing each n-gram size during the indexation step. We compare the following three strategies:

1. process n-grams separately depending on their length (strategy *sep.* in Table 3). This means that we only extract words sequences that are made of n-grams of equal length in the source *and* the target languages. Considering the example above, we first index

only unigrams (column "$n = 1$") and use them to extract alignments as in the original method. Then, we index bigrams (column "$n = 2$") *separately* and use them to extract alignments, disregarding all indexed unigrams, and so on. Doing so is, of course, of limited efficiency for language pairs such as Finnish–English: it would be better to allow the association of a single word in the agglutinative language with several words from the isolating language.

2. progressively mixing all n-gram lengths (strategy *mix* in Table 3). As with the previous strategy, we first index only unigrams and use them to extract alignments. We then index bigrams, but we don't use them *separately* as before; instead we consider the two columns "$n = 1$" and "$n = 2$" *simultaneously*, and recreate all word groups: some are identical to groups that were extracted for $n = 1$, in which case the counts of corresponding alignments are reinforced (e.g. unigram *a*), some are new (e.g. bigram *ac*), sometimes blocking out smaller associations (e.g. bigram *ab* masks unigram *b*). This last issue does not matter anymore, though, as these will already have been extracted in the previous step. Then we add trigrams, repeat the process, and so on. Associations extracted for all values of n are finally gathered together and serve to increment the extraction statistics.
3. force the alignment of n-grams of different lengths, by iteratively processing all possible (*source*, *target*) length values (strategy *force* in Table 3). This actually enables to align numerous n-grams with different lengths in the source and the target languages. Practically, we start again by indexing source unigrams and target unigrams and use them to extract alignments. Then, we index source unigrams and target bigrams, and so on (Cartesian product of all source and target lengths). This approach is computationally more demanding than the others, as we need to repeat the indexation and extraction steps for about ${n_{\max}}^2$ times, where $n_{\max}$ is the largest n-gram considered, and is very unlikely to scale up when we consider more than two languages. It is also the more prone to learning spurious associations, as we will repeat the extraction process for a great number of times for any given sub-corpora. For instance, nothing prevents the algorithm to associate English unigrams with long Finnish n-grams, a very unlikely association given the morphology of these languages.

In order to compare these three strategies, we have prepared a set of 100,000 random sub-corpora from Europarl (French–English) of various sizes, from which associations are extracted according to each strategy. This experiment was performed using a maximum n-gram length $n_{\max}$ ranging from 1 to 5 words. The resulting translation tables (5 values of $n_{\max}$ for each of the 3 strategies = 15 tables in total), obtained from this very same set of sub-corpora, are evaluated on the same tasks as previously: phrase-based statistical machine translation (evaluation criteria are BLEU and TER (Snover et al 2006)) and bilingual lexicon induction, as described in Sec. 2.3. Results are presented in Table 3. Following Clark et al (2011), because MERT is a non-deterministic optimization algorithm and results can greatly vary between runs, all BLEU and TER scores reported in subsequent machine translation evaluations are averages over 5 independent runs. Note that we did not use a sampling method to estimate the statistical significance of the observed differences, although this is standard practice. The standard deviation is always very small in our experiments, typically less than 0.1 BLEU or TER point.

As expected, the larger the maximal length of indexed n-grams, the more numerous the entries in the translation table and the longer those entries. This is because associations extracted for a given $n_{\max}$ contain those produced with a smaller $n_{\max}$. The scores in bilingual lexicon induction insignificantly improve as $n_{\max}$ increases for the two first approaches,

Table 3 Quality and characteristics of translation tables produced according to each of the three word clustering strategies, for various values of the maximal n-gram length. Lines where $n_{\max} = 1$ are identical for the three strategies and correspond to the original method. For each strategy, the best results in each column are in bold

Strat.	$n_{\max}$	Lexicon induction (%)	BLEU (%)	TER (%)	# entries	Avg. entry length
sep.	1	36.19	21.14	63.59	83,967	1.92
	2	**36.71**	22.65	61.94	277,858	2.79
	3	36.66	23.10	62.09	366,971	3.13
	4	36.60	**23.21**	**61.43**	393,453	3.24
	5	36.58	22.96	62.19	399,810	3.27
mix	1	36.19	21.14	63.59	83,967	1.92
	2	37.08	23.59	60.69	290,631	2.78
	3	37.35	**24.70**	**59.85**	398,880	3.12
	4	37.45	24.49	60.68	436,760	3.25
	5	**37.56**	24.21	59.93	448,212	3.31
force	1	**36.19**	21.14	63.59	83,967	1.92
	2	31.71	23.88	60.40	312,273	2.86
	3	30.90	**24.50**	60.67	453,429	3.24
	4	30.48	24.45	**59.97**	507,359	3.39
	5	30.25	24.27	60.03	524,091	3.45

but they significantly decrease with the third one. The gain in machine translation quality as measured by the BLEU and TER metrics, is significant with all three approaches: an increase of 2 to 3.5 BLEU points is observed compared with the original version of Anymalign ($n_{\max} = 1$), i.e. more than 10% relative. Nevertheless, the second one seems to produce marginally better results according to the three evaluation criteria. Its execution time is slightly higher than that of the first one (at worst twice as slow for 5-grams in our experiments), but it is by far faster than the third one (from the order of the hour to that of the day with 5-grams).

In the following experiments, we will adopt the second strategy (*mix*) which constitutes a good compromise between the other two. Figure 5 presents the detail of column "# entries" in Table 3 for this strategy, and is to be compared with Figure 2.

On the whole, incrementing by one the length of the indexed n-grams, i.e. moving from a particular curve to the one immediately above, increases considerably the quantity of extracted n-grams. This increase mostly concerns n-grams longer than or equal to $n_{\max}$, even though a small increase of shorter n-grams is also observed, due to the extraction of complementary sets. The most obvious case is when we move from unigrams to bigrams ($n_{\max} = 2$), which makes the quantity of bigrams output escalate, and to a lesser extent, also increases the quantities of all larger n-grams. The phenomenon also happens for larger values of $n_{\max}$ but it gets less and less significant as $n_{\max}$ increases. The graph suggests that it is useless to index n-grams made of more than 3 or 4 words, because the marginal return of those longer n-grams gets very small. This does not matter much, as the multi-word associations which are the most useful in machine translation are quite short (typically between 1 and 3 words).

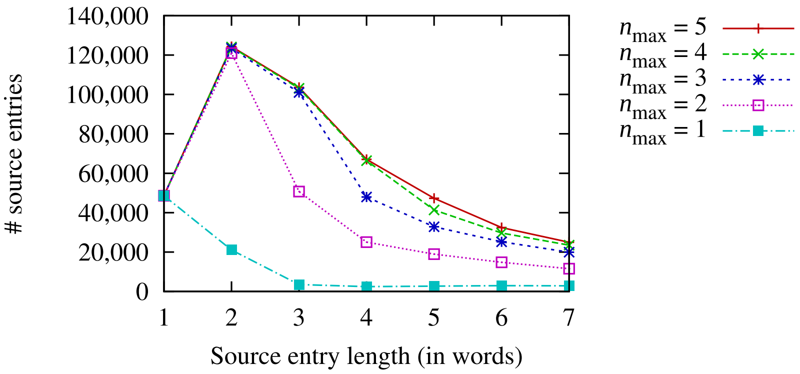

Fig. 5 Distribution of n-grams for the five translation tables obtained by the second word clustering strategies. Each curve corresponds to a line in Table 3, and the sum of values along the curve equals the number indicated in the column labeled "# entries" in that table. The lowest curve ($n_{\text{max}} = 1$) corresponds to the original method

Table 4 Corpora used for the MT evaluation

Task	Training	Development	Test	# ref. / test sentence
BTEC: ar–en	19,972	1,512	489	7
BTEC: zh–en	19,972	1,512	989	7
Europarl: fi–en, fr–en, pt–es	350,645	2,000	2,000	1

4.3 More associations yields improved translations

In this section, we systematically compare our generalized extraction algorithm with a standard MGIZA++/Moses baseline on several phrase-based statistical machine translation tasks. Table 4 presents the characteristics of the data used for each experiment: the BTEC:ar–en train-dev-test sets are made of official sets from the IWSLT 2007 evaluation campaign; for BTEC:zh–en it uses sets from the 2008 test campaign; and the Europarl dev and test data have been randomly selected from Europarl v.6 following standard evaluation practices. Results for the various conditions considered are summarized in Tables 5 and 6.

Lines with $n_{\text{max}} = 1$ correspond to the original version of Anymalign. As described in Section 2.3, since Anymalign may be stopped at any time, the stopping condition we impose depends on MGIZA++'s processing time. This time is constant whatever the value of n_{max}. Since processing time increases with this parameter, the larger this parameter, the *lower* the number of sub-corpora processed. This makes a big difference with the experiments presented in Section 4.2 where the whole set of sub-corpora to be processed was defined in advance, in which case the processing time was increasing with n_{max}. In principle, the translation tables produced for a given value of n_{max} should be larger than with any lower value, provided that the aligner runs long enough. This explains why the translation tables in Tables 5 and 6 may contain less entries than with larger values of n_{max}. In practice, these tables nevertheless contain much more long n-grams than for our original method, which yields significant score improvements, even when the translation table is smaller.

Table 5 Translation results on the BTEC

	Aligner	$n_{\max}$	BLEU (%)	TER (%)	# entries
	MGIZA++		*33.66*	*46.13*	*217,512*
	Anymalign	1	26.27	51.18	170,521
ar–en	-	2	30.90	49.69	269,454
	-	3	31.80	51.46	273,197
	-	4	**33.76**	**48.80**	258,141
	MGIZA++		*15.48*	*70.45*	*141,773*
	Anymalign	1	14.72	**69.04**	158,904
zh–en	-	2	16.29	71.73	263,315
	-	3	16.55	70.58	250,292
	-	4	**16.84**	69.42	269,353

Table 6 Translation results on Europarl

			Same processing time as MGIZA++			*Total CPU time = 20 MGIZA++*		
	Aligner	$n_{\max}$	BLEU (%)	TER (%)	# entries	BLEU (%)	TER (%)	# entries
	MGIZA++		*23.06*	*62.08*	*22,070,309*			
	Anymalign	1	14.15	74.01	1,761,448	15.86	72.08	6,355,116
fi–en	-	2	16.79	**69.60**	1,122,999	18.30	68.23	4,923,519
	-	3	**16.83**	69.65	658,670	**18.45**	**68.13**	2,999,088
	-	4	16.22	70.10	426,251	18.18	68.81	1,928,166
	MGIZA++		*30.40*	*53.60*	*25,638,671*			
	Anymalign	1	25.18	58.39	1,921,848	25.55	58.07	6,626,525
fr–en	-	2	**26.69**	**57.33**	1,373,790	**27.63**	**56.56**	6,553,379
	-	3	26.49	57.45	795,070	27.57	56.60	3,921,459
	-	4	26.25	57.56	514,256	27.43	56.67	2,461,018
	MGIZA++		*38.75*	*47.37*	*32,000,140*			
	Anymalign	1	35.39	50.35	1,881,503	36.66	49.28	1,403,812
pt–es	-	2	36.03	49.63	987,884	36.72	49.10	7,295,581
	-	3	**36.73**	**49.24**	829,393	37.20	49.05	3,799,224
	-	4	**36.73**	49.31	552,664	**37.33**	**48.89**	2,472,281

For the BTEC tasks, lines where $n_{\max} = 1$ show that the original version of Anymalign achieves BLEU scores that are comparable to those of MGIZA++ in Chinese–English, and is much worse (by 7 BLEU points) in Arabic–English. For larger values of $n_{\max}$, our method outperforms MGIZA++/Moses by more than one BLEU point in Chinese–English, and achieves comparable performance in Arabic–English, where we obtain a clear gain of 7 BLEU points.

For the Europarl tasks, the performance of the original version of Anymalign are significantly below those of MGIZA++/Moses, which is consistent with the results obtained in some previous experiments. In fact, the difference was not as strongly marked in our former experiments: we observed a difference of 2 to 3 BLEU points on average, whereas it amounts here to 6 points. This change might be due the use of much larger corpora: 350,000 sentence pairs here versus 100,000 in the experiments of (Lardilleux 2010). Anymalign's translation tables are very small compared to those of MGIZA++/Moses, which seems to indicate that the amount of time for which we run our algorithm is too small. Table 6 contains on its right part a second series of results, corresponding to the execution of Anymalign for a total time equal to 20 times MGIZA++'s processing time. In practice, as Anymalign allows for mas-

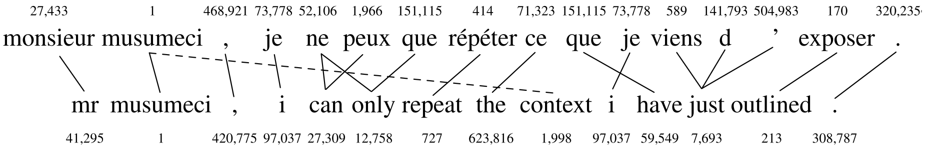

Fig. 6 An illustration of the "garbage collector effect." Numbers correspond to word counts in the training corpus. The English word *context*, for which no direct translation exists in the source sentence (its most probable translations in our corpus are *contexte* and *cadre*), was wrongly aligned with the French hapax *musumeci* (a proper name)

sive parallelism, we have distributed the processing on 140 processes and executed them on a cluster, so that the actual processing time was actually 7 times smaller than for the results on the left part of the table. The sizes of translation tables in the right part of the table are closer to those of MGIZA++, which confirms that the initial setting of the processing time was not sufficient.[6] The gain in BLEU is significant when $n_{\max}$ increases: we gain up to 3 BLEU points in Finnish–English ($n_{\max} = 3$) just by running Anymalign for a longer period of time. As can be seen in Table 6 (right part), in those conditions, the indexing of n-grams yields an increase in BLEU ranging from 1.7 point in French–English to roughly 4 points in Finnish–English. On average, the best scores of Anymalign remain 3.5 BLEU points below those of MGIZA++, closing half of the initial gap.

5 Combining translation tables

The main conclusion of the previous section is that the generalization we have introduced enables Anymalign to outperform the state-of-the-art MGIZA++ on "simple" phrase-based machine translation tasks (as exemplified by the results on BTEC, which is made of short and repetitive sentences), but it still lags behind on more difficult tasks (Europarl), in spite of significant improvements. Therefore, rather than using Anymalign on its own, we consider the benefits of using it as a complement of MGIZA++, as the two tools implement quite different approaches.

As noted in Section 3.2, Anymalign is better at aligning rare words, whereas MGIZA++ is better at aligning frequent ones. In fact, rare words are much rarer in Moses' default phrase tables obtained using MGIZA++ alignments: on average, for the 5 language pairs in our Europarl corpus (de–en, es–en, fi–en, fr–en, pt–es), only 17% of source hapaxes in the parallel corpus exactly match an entry in Moses' phrase table (29% as part of longer phrases). In comparison, this proportion rises to 61% with Anymalign (65% as part of longer phrases). The small proportion of rare words in the translation table obtained with MGIZA++'s alignments is well documented and is sometimes referred to as the "garbage collector effect" (Brown et al 1993a) (see also the recent analysis of (Graça et al 2010)): rare words in the source language tend to link with many target words for which there is no direct translation in the source language, while we would like to see such target words unaligned or aligned to other source words. Figure 6 gives an actual example on Europarl data using the French–English language pair.

[6] This raises another issue: Anymalign's stopping criterion. The current experiments suggest that our current criteria result in sub-optimal performance, even though they allow a fair comparison with other tools.

Table 7 Translation results on Europarl, using translation tables obtained by combining Moses' default phrase table with Anymalign's

	Aligner	$n_{\max}$	BLEU (%)	TER (%)	# entries
fi–en	*MGIZA++*		*23.06*	*62.08*	*22,070,309*
	Anymalign + MGIZA++	1	**23.08**	**62.15**	27,983,561
	-	2	22.53	62.59	26,408,141
	-	3	22.44	62.57	24,597,428
	-	4	22.22	65.52	27,584,655
fr–en	*MGIZA++*		*30.40*	*53.60*	*25,638,671*
	Anymalign + MGIZA++	1	30.41	53.71	31,710,782
	-	2	**30.68**	**53.62**	31,384,607
	-	3	30.10	53.89	28,887,597
	-	4	29.39	54.58	27,584,655
pt–es	*MGIZA++*		*38.75*	*47.37*	*32,000,140*
	Anymalign + MGIZA++	1	38.89	47.35	37,828,876
	-	2	38.81	**47.20**	37,409,893
	-	3	38.84	47.27	34,991,893
	-	4	**38.93**	47.27	33,825,477

In the example in Figure 6, even though the source word *musumeci* is linked here to some target words, it does not appear in the final phrase table, because of the large distance between the two target words it is linked to. In fact, a target phrase of eight words (from *musumeci* up to *context*) would be necessary for it to appear in the phrase table, but we rarely need to extract such long phrases because of their limited use.

On the other hand, Anymalign perfectly aligns French and English *musumeci* with translation probabilities of 1 in both directions. Therefore, in a last series of experiments, we combine Moses' default phrase table, which contains better alignments for frequent words, with Anymalign's, which contains better alignments for rare words. For this, we collect all phrase pairs from Moses' default phrase table and Anymalign's. Each phrase pair is assigned seven features: five from Moses' default phrase table (2 translation probabilities, 2 lexical weights, 1 length penalty) and two from Anymalign's (2 translation probabilities). Phrase pairs that appear in both phrase tables thus have 7 non-null feature scores. Phrase pairs that only appear in Moses' default phrase table have 5 non-null and 2 null feature scores, while phrase pairs that only appear Anymalign's phrase table have 5 null and 2 non-null feature scores. Anymalign may thus contribute in two ways: (1) by improving the coverage of the phrase table for low frequency words, and (2) by reinforcing phrase pairs extracted by the two aligners, which are more likely to be accurate as they were independently collected with the two alignment systems. The results of this experiment are reported in Table 7, and are to be compared with the right part of Table 6 (long run of Anymalign).

Combining the phrase tables yields small improvements in terms of BLEU scores (up to 1 BLEU point in French–English, $n_{\max} = 3$), which are corroborated by improvements in TER. The scores seem to fluctuate according to the value of $n_{\max}$, the best results being obtained for values of 2 or 3, and the worst with a value of 1 (often lower than Moses' phrase table used alone). This further confirms that our generalization has a positive effect on machine translation tasks.

In order to better understand those results, and try to determine the main effect of this phrase table combination strategy, we computed the proportion of source phrases used by

Table 8 Origin and frequency of phrases from the phrase table obtained by combination used by the Moses decoder. The values in this table are averages over the 12 experiments of Table 7 (the proportions are comparable whatever the language pair and the value of $n_{\max}$)

	Prop. of phrases (%)	Avg. phrase freq.	Avg. phrase length (in words)
Only Anymalign	2	3	1.09
Only Moses	16	19	2.76
Both	82	998	1.76

the decoder according to whether they occurred only in the default phrase table, only in Anymalign's, or in both. Table 8 reports these numbers along with average phrase frequencies in the source input corpus.

The decoder only uses a handful (2%) of phrases obtained only by Anymalign, and these correspond to very rare phrases, typically words occurring once in the training set and once in the test set. This corroborates the analysis in Section 3.2. Phrases found only in Moses' default phrase table are more numerous (16%), and contrary to what we might expect, they mostly correspond to medium frequency phrases (19 occurrences on average for more than 300,000 training pairs of sentences). This is because they correspond to longer phrases (2.76 words in average, while the phrases occurring only in phrase tables obtained by the generalized version of Anymalign, regardless of $n_{\max}$, are 1.09 words long on average), and are thus quite infrequent. These results suggest that most of the gains reported in Table 7 would rather result from reinforcement of phrase pairs which are extracted by the two aligners. This was confirmed in a contrastive experiment where all the entries output only by Anymalign are removed (i.e. the 2% of phrases used by the decoder as reported in Table 8). We observe a negligible loss of 0.12 BLEU point on average on the 12 experiments (4 values of $n_{\max}$ in 3 language pairs), showing that, indeed, most gains were due to the reinforcement of phrase pairs which are extracted by the two alignment systems, as described above.

6 Conclusion

This paper has presented a generalization of the sub-sentential alignment method introduced in (Lardilleux and Lepage 2008) aimed at improving its results in machine translation applications. The original method has many appealing properties: it may be stopped at any time, is easy to parallelize, and is intrinsically symmetric. It has also been shown to obtain better results than state-of-the-art approaches on bilingual lexicon induction tasks, but worse results on phrase-based statistical machine translation tasks. Based on a careful reexamination of the method, we showed that these discrepancies originate from two causes: (i) a systematic failure at extracting groups consisting of both frequent and rare words, and (ii) a tendency to extract noisy associations for frequent words, which is a primary handicapping factor for machine translation tasks. To alleviate these problems, we proposed a generalization of the indexation strategy, which amounts to considering word n-grams, rather than just single words, as the main indexation unit. Multi-word associations are now extracted through multiple n-gram indexation steps, for increasing values of n.

The result of this generalization is a clear improvement of the number of n-grams that are collected, yielding significant gains in phrase-based machine translation (up to +7 BLEU points on the Arabic–English pair). Our method is now neck and neck with state-of-the-art approaches on "simple" machine translation tasks (BTEC), and we have managed to close half the gap on more difficult tasks (Europarl). When combined with Moses' default phrase table, a modest, yet consistent improvement in BLEU is observed over several language pairs and directions. These results represent a significant achievement, given that the underlying model does not make use of any kind of alignment information.

This analysis opens several new avenues for exploration. On the one hand, we need to reconsider our sub-corpora sampling strategy in the light of the new results reported above. In particular, we need to find ways to increase the return of each sub-corpora in terms of the number of extracted alignments. This will be all the more crucial as we try to move towards *transductive translation* with "just-in-time" alignments, where associations will be extracted at test time for those words or phrases that need to be translated.

Another major source of inefficiency in our approach is the fact that the discontinuous fragments that are extracted through sampling are currently ignored and do not appear in our translation table: this means that more than half of the extracted segments are simply discarded. There are various ways to remedy this rather unsatisfactory state of affair, such as introducing alignment constraints so as to compute associations for each part of the discontinuous fragment. Another obvious idea to take advantage of these fragments is to use translation systems capable of handling these gappy units, e.g. hierarchical systems (Chiang 2007) or Dynamic Translation Memory as proposed by Biçici and Dymetman (2008).

Last, we also intend to consider generalizations of the method which can dispense with pre-tokenization of the text. As explained above, the notion of a word in our models is just a convenient way to identify the units to be indexed — it is thus tempting, albeit computationally more demanding, to index character strings, a generalization that might prove useful for some languages or language pairs, e.g. for languages which do not separate words by spaces.

Acknowledgements The work presented in this paper has been partially funded by the Cap Digital SAMAR project.

References

Biçici E, Dymetman M (2008) Dynamic translation memory: Using statistical machine translation to improve translation memory fuzzy matches. In: Proceedings of the 9th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2008), Haifa, Israel, pp 454–465, URL http://www.xrce.xerox.com/content/download/7009/52469/file/2007-046.pdf

Brown P, Cocke J, Della Pietra S, Della Pietra V, Jelinek F, Mercer R, Roossin P (1988) A statistical approach to language translation. In: Proceedings of the 12th International Conference on Computational Linguistics (Coling'88), Budapest, pp 71–76, URL http://aclweb.org/anthology-new/C/C88/C88-1016.pdf

Brown P, Della Pietra S, Della Pietra V, Goldsmith M, Hajic J, Mercer R, Mohanty S (1993a) But dictionaries are data too. In: Proceedings of the Workshop on Human Language Technologies, Plainsboro, New Jersey, USA, pp 202–205, URL http://www.aclweb.org/anthology/H93-1039

Brown P, Della Pietra S, Della Pietra V, Mercer R (1993b) The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics 19(2):263–311, URL http://aclweb.org/anthology-new/J/J93/J93-2003.pdf

Chiang D (2007) Hierarchical phrase-based translation. Computational Linguistics 33(2):201–228, URL http://aclweb.org/anthology-new/J/J07/J07-2003.pdf

Clark J, Dyer C, Lavie A, Smith N (2011) Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL 2011), Portland, Oregon, USA, pp 176–181, URL http://www.aclweb.org/anthology/P11-2031

Cover T, Thomas J (1991) Elements of Information Theory. Wiley and Sons, New York

Dagan I, Church K (1994) Termight: identifying and translating technical terminology. In: Proceedings of the fourth conference on Applied natural language processing, Stuttgart, pp 34–40, URL http://www.mt-archive.info/ANLP-1994-Dagan.pdf

Deng Y, Byrne W (2005) HMM word and phrase alignment for statistical machine translation. In: Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), Vancouver, British Columbia, Canada, pp 169–176, URL http://www.aclweb.org/anthology/H/H05/H05-1022

Dunning T (1993) Accurate methods for the statistics of surprise and coincidence. Computational Linguistics 19(1):61–74, URL http://portal.acm.org/citation.cfm?id=972450.972454

Fordyce CS (2007) Overview of the iwslt 2007 evaluation campaign. In: Proceedings of the 4th International Workshop on Spoken Language Translation (IWSLT 2007), Trente, pp 1–12, URL http://www.mt-archive.info/IWSLT-2007-Fordyce.pdf

Fraser A, Marcu D (2007) Getting the structure right for word alignment: LEAF. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), Prague, pp 51–60, URL http://www.aclweb.org/anthology/D/D07/D07-1006

Fung P, Church K (1994) K-vec: A new approach for aligning parallel texts. In: Proceedings of the 15th International Conference on Computational Linguistics (Coling'94), Kyōto, vol 2, pp 1096–1102, URL http://aclweb.org/anthology-new/C/C94/C94-2178.pdf

Fung P, Yee LY (1998) An IR approach for translating new words from nonparallel, comparable texts. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Montreal, vol 1, pp 414–420, URL http://www.aclweb.org/anthology/P98-1069

Gale W, Church K (1991) Identifying word correspondences in parallel texts. In: Proceedings of the fourth DARPA workshop on Speech and Natural Language, Pacific Grove, pp 152–157, URL http://www.aclweb.org/anthology/H/H91/H91-1026.pdf

Ganchev K, Graça J, Taskar B (2008) Better alignments = better translations? In: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-08: HLT), Columbus, Ohio, pp 986–993, URL http://www.aclweb.org/anthology/P/P08/P08-1112.pdf

Gao Q, Vogel S (2008) Parallel implementations of word alignment tool. In: Software Engineering, Testing, and Quality Assurance for Natural Language Processing, Columbus (Ohio, USA), pp 49–57, URL http://www.aclweb.org/anthology/W/W08/W08-0509.pdf

Gaussier E, Langé JM (1995) Modèles statistiques pour l'extraction de lexiques bilingues. Traitement Automatique des Langues 36(1-2):133–155

Graça J, Ganchev K, Taskar B (2010) Learning tractable word alignment models with complex constraints. Computational Linguistics 36(3):481–504, URL http://www.aclweb.org/anthology/J/J10/J10-3007.pdf

Johnson H, Martin J, Foster G, Kuhn R (2007) Improving translation quality by discarding most of the phrasetable. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), Prague, pp 967–975, URL http://www.aclweb.org/anthology/D/D07/D07-1103.pdf

Koehn P (2005) Europarl: A parallel corpus for statistical machine translation. In: Proceedings of the tenth Machine Translation Summit (MT Summit X), Phuket, pp 79–86, URL http://www.mt-archive.info/MTS-2005-Koehn.pdf

Koehn P, Och F, Marcu D (2003) Statistical phrase-based translation. In: Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2003), Edmonton, pp 48–54, URL http://aclweb.org/anthology-new/N/N03/N03-1017.pdf

Koehn P, Hoang H, Birch A, Callison-Burch C, Federico M, Bertoldi N, Cowan B, Shen W, Moran C, Zens R, Dyer C, Bojar O, Constantin A, Herbst E (2007) 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, URL http://aclweb.org/anthology-new/P/P07/P07-2045.pdf

Lardilleux A (2010) Contribution des basses fréquences à l'alignement sous-phrastique multilingue : une approche différentielle. PhD thesis, université de Caen Basse-Normandie, URL http://tel.archives-ouvertes.fr/tel-00107056/fr/, 204 pages

Lardilleux A, Lepage Y (2008) A truly multilingual, high coverage, accurate, yet simple, sub-sentential alignment method. In: Proceedings of the 8th Conference of the Association for Machine Translation in the Americas (AMTA 2008), Waikiki, pp 125–132, URL `http://hal.archives-ouvertes.fr/hal-00368737/fr/`

Lardilleux A, Lepage Y (2009) Sampling-based multilingual alignment. In: Proceedings of Recent Advances in Natural Language Processing (RANLP 2009), Borovets, pp 214–218, URL `http://hal.archives-ouvertes.fr/hal-00439789/fr/`

Lardilleux A, Chevelu J, Lepage Y, Putois G, Gosme J (2009) Lexicons or phrase tables? an investigation in sampling-based multilingual alignment. In: Proceedings of the 3rd Workshop on Example-Based Machine Translation (EBMT3), Dublin, pp 45–52, URL `http://hal.archives-ouvertes.fr/hal-00439806/fr/`

Liang P, Taskar B, Klein D (2006) Alignment by agreement. In: Proceedings of the Human Language Technology Conference of the NAACL, New York City, pp 104–111, URL `http://www.aclweb.org/anthology/N/N06/N06-1014`

Mandelbrot B (1954) Structure formelle des textes et communication. Word 10:1–27

Marcu D, Wong D (2002) A phrase-based, joint probability model for statistical machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2002), Philadelphie, pp 133–139, URL `http://www.aclweb.org/anthology/W02-1018`

Melamed D (2000) Models of translational equivalence among words. Computational Linguistics 26(2):221–249

Montemurro M (2004) A generalization of the zipf-mandelbrot law in linguistics. Nonextensive Entropy: interdisciplinary applications 12 pages

Moore R (2004) 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

Moore R (2005) Association-based bilingual word alignment. In: Proceedings of the ACL Workshop on Building and Using Parallel Texts, Ann Arbor, pp 1–8, URL `http://www.aclweb.org/anthology/W/W05/W05-0801.pdf`

Papineni K, Roukos S, Ward T, Zhu WJ (2002) BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002), Philadelphie, pp 311–318, URL `http://www.aclweb.org/anthology/P02-1040`

Smadja F, Hatzivassiloglou V, McKeown K (1996) Translating collocations for bilingual lexicons: A statistical approach. Computational Linguistics 22(1):1–38, URL `http://aclweb.org/anthology-new/J/J96/J96-1001.pdf`

Snover M, Dorr B, Schwartz R, Micciulla L, Makhoul J (2006) 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, URL `http://www.mt-archive.info/AMTA-2006-Snover.pdf`

Takezawa T, Sumita E, Sugaya F, Yamamoto H, Yamamoto S (2002) 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, URL `http://gandalf.aksis.uib.no/lrec2002/pdf/305.pdf`

Vogel S (2005) PESA: Phrase pair extraction as sentence splitting. In: Proceedings of the tenth Machine Translation Summit (MT Summit X), Phuket, pp 251–258, URL `http://www.mt-archive.info/MTS-2005-Vogel.pdf`

Vogel S, Ney H, Tillman C (1996) Hmm-based word alignment in statistical translation. In: Proceedings of the 16th International Conference on Computational Linguistics (Coling'96), Copenhague, pp 836–841, URL `http://aclweb.org/anthology-new/C/C96/C96-2141.pdf`

Wu D (1997) Stochastic inversion transduction grammar and bilingual parsing of parallel corpora. Computational Linguistics 23(3):377–404, URL `http://www.aclweb.org/anthology/J/J97/J97-3002.pdf`

Zipf G (1965) The Psycho-Biology of Language: An Introduction to Dynamic Philology. Classic Series, The MIT Press, Cambridge, USA, fist edition 1935