Improving Bilingual Sub-sentential Alignment by Sampling-based Transpotting

Li Gong, Aurélien Max, François Yvon

LIMSI-CNRS & Univ. Paris Sud
Orsay, France
{firstname.lastename}@limsi.fr

Abstract

In this article, we present a sampling-based approach to improve bilingual sub-sentential alignment in parallel corpora. This approach can be used to align parallel sentences on an *as needed* basis, and is able to accurately align newly available sentences. We evaluate the resulting alignments on several Machine Translation tasks. Results show that for the tasks considered here, our approach performs on par with the state-of-the-art statistical alignment pipeline giza++/Moses, and obtains superior results in a number of configurations, notably when aligning additional parallel sentence pairs carefully selected to match the test input.

1. Introduction

Sub-sentential alignment consists in identifying translation units from a sentence-aligned parallel corpus, which is a crucial component of state-of-the-art Statistical Machine Translation (SMT) technology. One of the most prominent approaches nowadays is Phrase-based Statistical Machine Translation, which is built upon the word alignment output. The problem of learning sub-sentential alignment from parallel texts is well-known, and numerous proposals have been put forward to perform this task. Those methods roughly fall into two main categories, broadly described here as the *probabilistic* and the *associative* approaches.

The probabilistic approach, introduced in [1], considers the problems of identifying *links* between words or groups of words in parallel sentences. This approach consists in defining a probabilistic model (e.g. IBM models [2]) of the parallel corpus, the parameters of which are estimated by a global optimization process which simultaneously considers all possible associations in the entire corpus. Due to its tight integration within the SMT framework, this approach is by far the most widely used. However, it is characterized by a number of shortcomings, in particular:

- Its parameters have to be estimated and optimized based on the entire parallel corpus, hence all units in the parallel corpus have to be aligned simultaneously. This makes it a time-consuming process, especially when working on large parallel corpora. In addition, many aligned parallel sentence pairs are never used to translate an input text.
- New data are constantly made available. It is a waste of resource to run the alignment process repeatedly for the whole corpus when only a proportionally low number of new sentences are added.

These shortcomings are addressed notably in [3], which uses the online EM algorithm of [4] to implement online learning for the HMM alignment model.

Associative approaches were introduced in [5]. They do not rely on an alignment model, but rather on independence statistical measures such as the Dice coefficient, mutual information [5, 6], or likelihood ratio [7]. In this approach, a local maximization process is used, where each sentence is processed independently.

An associative sub-sentential alignment method, named Anymalign, was introduced in [8, 9]. This method relies on simple comparisons on (source and target) word occurrence distribution over randomly sampled sub-corpora. Words with the same occurrence distribution over a particular sub-corpus are extracted as an association. The more often two words are associated, the better the association score between them, and the more likely they are to be mutual translations. This method was shown to produce better results than state-of-the-art methods on bilingual lexicon constitution tasks, when the evaluation is performed by comparing word associations with reference dictionaries, but failed to perform on par with state-of-the-art methods for building SMT phrase tables. It was subsequently improved in [10], in which a recursive binary segmentation algorithm is used to process the output of Anymalign so as to obtain better sub-sentential alignments at the sentence level. While this improvement yields a performance that is comparable with the statistical approach, it can do so by processing large numbers of randomly sampled sub-corpora in order to obtain an accurate association measure and a good coverage for the entire corpus.

In this work, we propose a method to adapt Anymalign in order to align the parallel sentences on a per-need basis, meaning that it can also be used to accurately align new parallel sentences as they become available. The rest of this paper is organized as follows: Section 2 describes our sampling-based alignment approach in some detail, Section 3 presents an evaluation on several, complementary Machine Translation experiments, and Section 4 discusses our main results and introduces some of our future work.

2. Description of the method

We assume that, given a parallel bilingual corpus C, we wish to align several sentence pairs in a set S: S can be a part of the entire parallel bilingual corpus ($S \subseteq C$), or can correspond to newly available data ($S \nsubseteq C$).

An association table is first extracted for sentences in S by a *sampling-based transpotting* method. This table contains only the source phrases that exist in some sentence(s) of S. Using this table, a recursive binary segmentation algorithm (as in [10]) is applied to each sentence pair of S so as to generate the desired sub-sentential alignment.

2.1. Sampling-based transpotting

Our sampling-based transpotting method is inspired by `Anymalign`, which aims at extracting sub-sentential associations from multilingual, parallel corpora. `Anymalign` repeatedly draws random sub-corpora from the full parallel corpus, and extracts associations from each sub-corpora, which are used to build an association table between phrases. As each sub-corpora is independent, this process could be stopped at any time. However, large numbers of sub-corpora have to be processed in order to achieve a good coverage of the phrases in the entire corpus.

In our work, `Anymalign` is adapted in order to extract an association table for a specific list of sentence pairs S. Each sentence pair $(\mathbf{s}, \mathbf{t})$ in S is processed separately and a number N of random sub-corpora are sampled from the full parallel corpus C for each sentence. For each sub-corpora, the distribution profile is computed only for words (or phrases) occurring in s and bilingual phrases with the same profile are extracted as likely associations. The more sub-corpora are processed for each sentence pair, the more associations could be extracted, and the more accurate the association measures are. The set of all associations extracted from each sentence pair form the association table of S. In a nutshell, this procedure performs bi-sentence alignment *via* transpotting based on randomly sampled sub-corpora. The complete process is illustrated on an English-French sentence pair on Figure 1.

There are notable differences between this method and `Anymalign`:

- `Anymalign` draws random sub-corpora from the parallel corpus, and computes the occurrence distribution profile for all words of all sentence pairs in the sub-corpora, while we need to compute such profiles only for words in the sentence pair to align.[1]

- `Anymalign` is *anytime* but typically requires a large number of sub-corpora to achieve a good coverage over the entire corpus. We draw N sub-corpora for each given sentence pairs to ensure better coverage for the contents of each sentence pair to align. This allows

(1) Given a source-target sentence pair, we need to extract an association table for it:

one coke , please . ↔ *un coca, s'il vous plaît .*

⇓

(2) Draw a random sub-corpus from the parallel corpus:

	English	*French*
1	*one coffee, please .*	*un café, s'il vous plaît .*
2	*the coffee is not bad .*	*ce café est correct .*
3	*yes, one tea .*	*oui, un thé .*

⇓

(3) Compute occurrence distribution profile for words in the current sentence pair:

words with same distribution profile			*profiles*
one ,	↔	*un ,*	*[1, 0, 1]*
coke	↔	*coca*	*[0, 0, 0]*
please	↔	*s'il vous plaît*	*[1, 0, 0]*
.	↔	*.*	*[1, 1, 1]*

⇓

(4) If the source and target phrases are each contiguous, then increment the count for the corresponding phrase pair:

1. *count of (coke↔coca) plus 1*
2. *count of (please↔s'il vous plaît) plus 1*
3. *count of (. ↔ .) plus 1*

⇓

(5) Repeat steps (2) and (4) N times, so as to obtain an association table for the given sentence pair, e.g.:

source phrase		*target phrase*	*count*
one	↔	*un*	*830*
coke	↔	*coca*	*560*
one coke	↔	*un coca*	*20*
,	↔	*,*	*900*
please	↔	*s'il vous plaît*	*160*
please	↔	*s'il*	*200*
please	↔	*plaît*	*500*
.	↔	*.*	*980*

Figure 1: Illustration of the sampling-based transpotting method on an English-French sentence pair.

[1] Note that, when one's objective is in fact to align a complete parallel corpus, all counts should be kept.

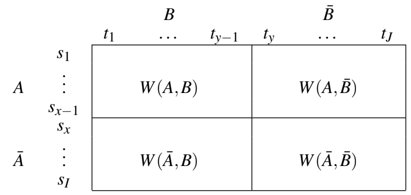


Figure 2: Schematic representation of the segmentation of a pair of sentences $S = A.\bar{A}$ and $T = B.\bar{B}$ (from [10]).

to align sentences on a per-need basis, and furthermore offers a more interpretable running time, which is now controlled by the amount of desired sampling for each sentence pair, which could e.g. depend on its length.

2.2. Sub-sentential alignment extraction

Once the association table for some sentence pairs is obtained, a recursive binary segmentation algorithm, described in [10], and inspired by the work of [11, 12], is used to generate a sub-sentential alignment for each sentence pair. Its purpose is to recursively segment the source and target sentence simultaneously on the basis of local association scores so as to find the *links* between the source and target words. It thus requires some association score $w(s,t)$ between each source word s and target word t in a sentence pair, which can be the result of the process described in Section 2.1. Then, recursive binary segmentation is guided by the sum W of the association scores between each source and target words of a block $(X,Y) \in \{A, \bar{A}\} \times \{B, \bar{B}\}$ (as shown in Figure 2):

$$W(X,Y) = \sum_{s \in X, t \in Y} w(s,t) \qquad (1)$$

The best segmentation is the one which minimizes the score defined in Equation 2:

$$\text{cut}(X,Y) = W(X,\bar{Y}) + W(\bar{X},Y) \qquad (2)$$

which would indicate that the association between the words of X and $\bar{Y}$ on the one hand, and the words of $\bar{X}$ and Y on the other hand, have low association scores. Following [10], we use instead a normalized variant so as to not to encourage unbalanced segments:

$$\text{Ncut}(X,Y) = \frac{\text{cut}(X,Y)}{\text{cut}(X,Y) + 2 \times W(X,Y)} + \frac{\text{cut}(\bar{X},\bar{Y})}{\text{cut}(\bar{X},\bar{Y}) + 2 \times W(\bar{X},\bar{Y})} \qquad (3)$$

With this segmentation criterion, the binary segmentation algorithm tests every possible binary segmentation in order to find the best segmentation score, and recursively segments blocks in a greedy fashion. In our current implementation, the segmentation terminates on blocks with at least one side of length 1 token. Figure 3 shows an example of segmentation, where atomic aligned biphrases correspond to framed rectangles containing values in bold. The words in aligned biphrases are linked with each other, which forms the word-to-word alignment of the bisentence.

	un	coca	,	s'il	vous	plaît	.
one	**0.246**	ϵ	ϵ	ϵ	ϵ	ϵ	ϵ
coke	ϵ	**0.138**	ϵ	ϵ	ϵ	ϵ	ϵ
,	ϵ	ϵ	**0.624**	0.002	ϵ	ϵ	0.048
please	ϵ	ϵ	ϵ	**0.032**	**0.008**	**0.128**	ϵ
.	ϵ	ϵ	0.020	ϵ	ϵ	ϵ	**0.873**

Figure 3: Example of alignment by recursive segmentation. The number in each cell corresponds to the value of function w, with $0 < \epsilon \leq 0.001$.

2.3. Self-convergency normalization

Segmentation scores for each position of token pairs are initialized by looking up values in the association table obtained by sampling-based transpotting (see Section 2.1). Because these association scores may sometimes be unreliable and poor indicators of a translation relationship, the best-first segmentation algorithm may produce incorrect results, especially on long sentence pairs. In addition, the bilingual sentence pairs are often in some relation to each other. So, well aligned sentences can help improve the alignment of more difficult sentences.

Therefore, we propose to use the previously produced alignment to extract the source-target phrase pairs to build an updated association table. This new table can then be used for another, better informed pass of recursive segmentation. This can be repeated until the obtained alignments are stable across iterations. This is described in Algorithm 1, where distance($A - A'$) is the percentage of different links between A and A'.

Algorithm 1 Self-convergency normalization

```
Given a parallel corpus $C$ and its alignment $A$
NumIter=0
while NumIter < MaxIter do
  Extract all aligned source-target phrases from $C$ using $A$ with the same heuristic as Moses
  The extracted phrase pairs and their counts are used to build an association table $T$ (the same kind of table as the table in step 5 in Figure 1)
  Using $T$ as the input of the binary segmentation algorithm (cf. Section 2.2), a new alignment $A'$ is computed
  if distance($A - A'$) $< \epsilon$ then
    return $A$
  end if
  NumIter+=1
end while
return $A$
```

3. Experiments

3.1. Experimental settings

In this section, we describe experiments intended to test the performance of the associative sub-sentential alignment approach described in Section 2. We will focus on measuring the impact of several alignment strategies for a phrase-based SMT system. We will use the Moses toolkit [13], which can be regarded as state-of-the-art for building SMT systems. Moses will be used in all configurations to build phrase tables and reordering tables from alignment matrices, and its decoder will be used to build candidate translations during optimization (using standard MERT [14]) and testing.

Translation performance will be measured by classical corpus-based metrics, BLEU [15] and TER [16]. All results are average scores computed on the test set for 3 independent optimization runs on the development set [17].

Experiments will be conducted on three language pairs and two main corpora, and we will make use of several reference translations when possible. We will also resort to oracle decoding using a greedy, approximate local search strategy and a number of phrase-based operators [18] to get some account of the best translation score attainable given each specific phrase table. We will furthermore consider the compactness of the produced phrase tables, as it can be regarded as a desirable quality of phrase tables licencing works on phrase table pruning (see e.g. [19]), and anormally large phrase table may in fact only artificially inflate oracle results.

Two sets of experiments will be carried out in this work. The first set of experiments is designed to validate the quality of the alignment generated by our method (henceforth sba, for sampling-based alignment) on some predefined bilingual corpus against a state-of-art alignment pipeline, based on giza++ [20], using default parameters from Moses. This approach is refered to as giza++ . The second set of experiments aims to assess the ability to align new bilingual data. For this experiment, we will focus on adding sentence pairs from a very large (unaligned) bilingual corpus, chosen on the basis that they contain translations for previously out-of-vocabulary tokens. Our approach will be compared against the same alignment pipeline using the augmented parallel corpus. This strategy is however costly as it requires to retrain the complete models, so we also performed a comparison with alignments obtained using the orginal alignment models, without any retraining.

3.2. Data sets

Experiments were performed on two parallel corpora, described in Table 1: BTEC is a small English-French subpart of the Basic Travel Expression Corpus [21]; and HIT is a corpus of basic expressions built for the Beijing 2008 Olympics, used here in English, French and Chinese. We used the BTEC development set of 2003 (devel03) and BTEC test set of 2009 (test09) as our development and test set, which are described in Table 2. Note that the former has 16 reference translations available for English, and the latter has 7, allowing for a somehow more interpretable measure of performance for language pairs with English as the target language.

Corpus	# lines	#token$_{en}$	# token$_{fr}$	# token$_{zh}$
BTEC	20K	182K	207K	-
HIT	62K	600K	690K	590K
EPPS	1,982K	54,170K	59,702K	-
supp	3.3K	111K	121K	-
WMT	11,745K	317,688K	383,076K	-

Table 1: Training bitext corpora statistics

Corpus	#lines	Avg(#token$_{en}$)	#token$_{fr}$	#token$_{zh}$
devel03	506	4,098 (16 refs)	4,220	3,435
test09	469	3,928 (7 refs)	4,023	3,031

Table 2: Tuning and test sets statistics

We will describe in Section 3.4 experiments that make use of additional data extracted from the large EPPS (Europarl) English-French parallel corpus of parliamentary debates, as well as a substantially larger corpus from the translation task of the Workshop on Statistical Machine Translation (WMT)[2]: both are described in Table 1. Our development and test sets will remain the same for all experiments.

English and French texts are normalized and tokenized by our in-house tools, and Chinese texts are segmented by a CRF-based Chinese word segmenter[3].

3.3. Basic alignment task

This experiment aims to assess the quality of the sub-sentential alignment generated by our method on a full bilingual parallel corpus. We use the giza++ implementation of [22] as a competitive baseline, with default settings : 5 iterations of IBM1, HMM, IBM3, and IBM4, in both directions (source to target and target to source). As for our alignment method, its alignment quality depends on the number of sub-corpora (N) that are drawn for each sentence pair. In this work, we choose a constant value of $N = 1000$ for all sentence pairs. The self-convergency normalization process is repeated for a maximum of 10 iterations.

The results for the two alignment methods are reported in Table 3, where we compare them on 2 parallel corpus (**BTEC** and **HIT**) and their simple concatenation (**BTEC+HIT**) and 3 translation directions on the same test set.

3.3.1. In-domain evaluation

First, on the in-domain corpus, **BTEC**, we find that our approach performs better than giza++, in particular by a large margin on the single-reference English→French direction (average of +2.13 BLEU). These results are furthermore obtained using a substantially smaller phrase table (315K vs. 360K entries in the phrase tables). Oracle-BLEU also indicates a clear advantage for our approach (average

[2] http://www.statmt.org/wmt12

[3] http://nlp.stanford.edu/software/segmenter.shtml

	BTEC				**HIT**				**BTEC+HIT**			
	BLEU	oracle-BLEU	TER	# entries	BLEU	oracle-BLEU	TER	# entries	BLEU	oracle-BLEU	TER	# entries
English→French (1 reference)												
giza++	45.68	76.26	37.03	360K	39.65	68.20	44.50	1,217K	47.97	83.62	35.45	1,546K
sba	47.81	77.78	36.60	315K	39.70	68.45	43.56	921K	47.55	84.40	37.22	1,241K
French→English (7 references)												
giza++	59.50	77.23	24.59	360K	45.52	68.58	33.99	1,224K	63.69	84.00	21.95	1,551K
sba	59.92	77.50	24.22	315K	45.34	69.59	33.79	937K	64.44	83.57	22.31	1,241K
Chinese→English (7 references)												
giza++	-	-	-	-	27.88	51.69	50.76	1,139K	-	-	-	-
sba	-	-	-	-	27.85	53.05	50.93	655K	-	-	-	-

Table 3: Results of experiments where specific bilingual parallel corpora are fully aligned. Values all correspond to average scores over three decodings of the test file for 3 independent optimization runs.

of +1.52 BLEU). These last two results are possible indicators of the fact that our approach produced a better subsentential alignment of the parallel corpus: better results can be (and are) obtained although fewer phrase pairs were extracted from the corpus.

3.3.2. Multiple-reference evaluation

Looking at the opposite translation direction with 7 reference translations, French→English, we still find that our technique is superior to the baseline, although to a much more modest extent (averages of +0.42 BLEU for the one-best translation and +0.27 BLEU for the oracle). Using several reference translations can potentially help us ensure that measured improvements are more related to *actual improvements* that e.g. make translation lexically more appropriate, than to specific choices that would accidentally resemble some particular reference translation. Again, our three indicators (one-best translation, oracle translation, and phrase table size) all indicate that our approach is here superior to the baseline.

3.3.3. Out-of-domain evaluation

Moving to the slightly less in-domain **HIT** corpus (the baseline performance drops from 59.50 to 45.52 BLEU on French→English), we find that the two approaches now perform roughly in the same ballpark, with our approach still producing significantly more compact phrase tables. For the more interpretable French→English condition with 7 reference translations, we find that although BLEU cannot be used to decide between the two, the oracle value still indicates a large advantage for our sampling-based alignment (average of +1.01 BLEU). This means that it managed to extract more useful phrase pairs, but that their various scores could not be used to ensure that those would be used in the one-best hypotheses of the decoder. Given that **HIT** is of a different origin than the test corpus (**BTEC**), it is well conceivable that translation preferences or even senses can often differ, resulting in some appropriate translation hypotheses with low scores that prevent them from appearing in one-best hypotheses.

3.3.4. Larger, composite training corpus evaluation

The previous hypothesis seems to hold when considering the larger task corresponding to the concatenation of the two parallel corpora (**BTEC+HIT**), where **HIT** data outnumber **BTEC** data by more than 3:1. Results are however less clear-cut here: for instance, our approach still performs better on French→English (average of +0.75 BLEU on one-best hypotheses), but fares worse in terms of oracle performance (average of -0.43 BLEU). These results include a reflection of the fact that giza++ improves its alignment with more data, even when adding out-of-domain data [23]. At this stage of our work, we do not control which particular sentence pairs are drawn in our samples, so assessing the impact of a larger overall sentence pool cannot be done.

3.3.5. Difficult language pair evaluation

Lastly, we turn to the more difficult Chinese→English condition, which is significantly more difficult than its French→English counterpart (27.88 BLEU vs. 45.52 BLEU for the giza++ baselines). A similar pattern emerges for the two language pairs: one-best translation performance is comparable, but oracle results indicate a clear advantage for our sampling-based alignment (average of +1.36 BLEU). Furthermore, for this language pair, we find that this is obtained with significantly fewer phrase table entries (almost half as many). Chinese words may in fact be very difficult to align to English words, partly for ambiguity reasons, and many noisy translation candidates may be extracted. Additionally, many words may be left unaligned by giza++, leading to artificially large numbers of extracted phrase pairs by the default grow-diag-final-and heuristic.

3.4. Incremental alignment task

In the previous section, we have shown that our approach performs on par with the giza++ baseline on the studied configurations for full corpus alignment. We now turn to the issue of aligning new data, which in many situations could

Phrase tables				**HIT**					
main	**supplementary**								
(62K HIT)	(3.3K EPPS)	# entries	# transl.	BLEU	1g	2g	3g	4g	TER
French→English (7 references)									
`giza++`	`none`	-	-	45.52	76.5	52.2	37.8	27.1	33.99
|	`forced`	59	1,993	47.94	76.8	55.4	41.0	29.2	34.62
|	`concat`	60	1,190	48.69	78.4	56.1	41.4	29.8	33.09
|	`sba`	64	681	49.83	80.9	57.3	42.0	30.5	30.61
|	`concat++`	62	1,218	50.23	81.5	57.8	42.6	31.1	29.81
`sba`	`none`	-	-	45.34	77.0	52.1	37.4	26.9	33.79
|	`sba`	64	681	50.45	81.8	58.3	42.5	30.9	29.94

Table 4: Results of experiments where a supplementary corpus is pooled and aligned by several methods.

only be performed on demand. Indeed, considering that all input sentences in our test set could be translated independently at large intervals of time, it would certainly not be conceivable, time-wise and computation-wise, to perform a full statistical alignment of the iteratively growing bilingual corpus. We will nonetheless report evaluation results for this situation below.

Few works have previously considered the task of incremental alignment of parallel corpora [24, 25]. The focus in [25] is put on a careful selection of additional data, a reflection of the fact that not all training data can be beneficial for training and improving SMT systems [26]. For these experiments, we will concentrate on a very specific use of additional data with a conservative view[4]: sentences will be pooled from a very large, any-domain parallel corpus (`EPPS` in Table 1) on the basis that they contain at least one occurrence of a word that is out-of-vocabulary (OOV) in the baseline parallel corpus[5]. In order to study a condition where significant numbers of such OOVs exist, we used the **HIT** corpus as our main corpus, relatively to which our test set contains 79 unique OOVs (436 occurrences). Our additional training data (**EPPS**) provided matches for 65 of them. We retrieved a maximum of 100 sentences pairs for each of these 65 OOVs, which yielded an additional parallel corpus of 3,355 sentence pairs (**supp** in Table 1).

We now describe the configurations that will be compared. A main table will be used for all configurations, corresponding either to the `giza++` baseline or to our sampling-based approach. A supplementary table will be built from **supp** by various means:

- forced alignment on **supp** using the statistical models (previously) obtained on **HIT** (`forced`);
- statistical alignment on the concatenation **HIT+supp**, and extraction of the alignments on **supp** only (`concat`);
- sampling-based alignment on **supp**, sampling from the union of **HIT** and **supp** (`sba`);
- statistical alignment on the very large corpus used for experiments at WMT'12 [27], and extraction of the alignments on **supp** only (`concat++`).

As said previously, the `concat` variants cannot be considered as practical solutions for the problem at hand. Once alignments are obtained for the **supp** corpus, a separate phrase table is used by the `Moses` tools as previously, and MERT is used with the resulting two tables, where our additional table is used as backoff, for unigrams only. Therefore, our additional training data, once aligned, will only be used in practice for proposing translations for previously unknown words. Note that in this experiment we do not extract necessary information to update the lexicalized reordering models used by `Moses`.

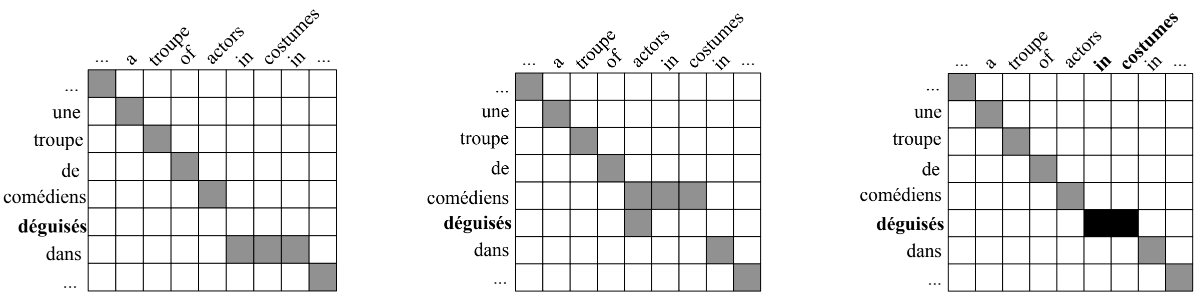
Results for this set of experiments are given in Table 4. Using `giza++` for building the main translation table, we find a very clear ranking for all the studied strategies: `concat++` > `sba` > `concat` > `forced` > `none`. The only approach that outperforms ours (average of +0.4 BLEU) is the statistical alignment technique using more than 11.7M sentence pairs[6]. `sba` outperforms `concat` (average of +1.14 BLEU) and `forced` (average of +1.89 BLEU), the latter being the most practical baseline to consider. Significant improvements can be observed on 1-gram precision, which percolate nicely to higher-order *n*-grams. We note once more that our technique produces much smaller phrase tables, and further note that the `concat` variants already significantly reduce the numerous entries produced by `forced`.

Interestingly, we manage to improve this result further by using also our sampling-based alignment technique for aligning the main parallel corpus (average of +0.62 BLEU), which furthermore happens to be even slightly superior to `concat++` (average of +0.22 BLEU, with small improvements on 1-gram and 2-gram precisions). To explain this fact, we return to our oracle results reported in Table 3 on

[4] We, however, do not have the guarantee that even if translations are correctly extracted, those will be those found in the reference translations.

[5] Meaning that the word was not present in the original training data, not that no translation for it could be extracted by some technique.

[6] This alignment process took roughly 2 weeks using modern computing resources.


... a troupe of actors in costumes in ...
...
une
troupe
de
comédiens
déguisés
dans
...
... a troupe of actors in costumes in ...
...
une
troupe
de
comédiens
déguisés
dans
...
... a troupe of actors in costumes in ...
...
une
troupe
de
comédiens
déguisés
dans
...

(a) giza++ forced alignment (b) giza++ concat (c) sampling-based alignment

Figure 4: Example of matrices on French-English obtained using two `giza++` baselines and our sampling-based strategy.

HIT for French-to-English translation. We there found that one-best translation was slightly superior for the baseline (average of -0.18 BLEU), but that the oracle for our approach was superior (average of +1.01 BLEU), indicating that our approach did extract more useful phrases, but which were apparently poorly scored, possibly due to domain mismatch between training and testing. It seems that providing the decoder with translation for previously OOV words had an additional effect on the configuration where we use the phrase table obtained using our technique: such translations now seem to be selected more often, resulting e.g. in a largely improved 1-gram precision by using our additional phrase table (+4.8).

4. Discussion and future work

In this work, we have presented an extension of the work by [10] on sampling-based alignment and a number of experiments that have shown its very competitive performance. Our approach performed at worse on par with a state-of-the-art baseline implementing a probabilistic approach, and obtained superior results in a number of configurations. Its more apparent strength emerged when aligning new data containing highly useful words (words that were previously out-of-vocabulary in the available data). While it remains to be shown more formally, we hypothesize that these improvements mainly stem from the improved alignment of rare words and its cascading effects. Figure 4 illustrates a case where the rare French word *déguisés* (here: *in costumes*) was only correctly aligned by our technique, and where the negative consequences for the two `giza/moses` baselines could be important (at least, for our experiments, no translation for *déguisés* alone could be extracted from this sentence pair by `giza++` here).

The framework that we have described for targeted additional data selection from parallel corpora will be the basis for our future work. We can, by principle, work at the level of tera-scale translation [28], by accessing efficiently (using suffix arrays) large quantities of unaligned parallel corpora, and perform transpotting and phrase table construction on a per-need basis. However, considering the diversity in nature, origin and quality of all possibly additional training examples, some adaptation should be performed so as to introduce preferences for the most promising examples, and hence extracted translations. In this context, the most realistic scenario will be a follow-up to our previous work on *any-text* translation [29], where notably little or no a priori knowledge exists about (additional) training examples, and adaptation should be performed on-the-fly. Finally, it seems obvious that the search for new translations, and in particular for unknown words and phrases as well as poorly adapted phrases, should also be pursued in *less parallel* corpora (see e.g. [30]). It is then an interesting question to consider how our technique would fare and how it could be adapted to work indifferently on parallel or reasonably comparable sentence pairs.

5. Acknowledgements

This work was partially funded by the French State agency for innovation (OSEO) in the Quaero Programme.

6. 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 COLING*, 1988, pp. 71–76.

[2] P. Brown, V. Della Pietra, S. Della Pietra, and R. Mercer, “The mathematics of Statistical Machine Translation: parameter estimation,” *Comput. Linguist.*, vol. 19, no. 2, pp. 263–311, 1993.

[3] A. Levenberg, C. Callison-Burch, and M. Osborne, “Stream-based translation models for Statistical Machine Translation,” in *HLT: The 2010 Annual Conference of NAACL*, 2010, pp. 394–402.

[4] O. Cappé and E. Moulines, “On-line expectation–maximization algorithm for latent data models,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 71, no. 3, pp. 593–613, 2009.

[5] W. A. Gale and K. W. Church, "Identifying word correspondence in parallel texts," in *Proceedings of the Workshop on Speech and Natural Language*, Pacific Grove, USA, 1991, pp. 152–157.

[6] P. Fung and K. W. Church, "K-vec: A new approach for aligning parallel texts," in *Proceedings of COLING*, Kyoto, Japan, 1994, pp. 1096–1102.

[7] T. Dunning, "Accurate methods for the statistics of surprise and coincidence," *Computational linguistics*, vol. 19, no. 1, pp. 61–74, 1993.

[8] A. Lardilleux and Y. Lepage, "Sampling-based multilingual alignment," in *Proceedings of RANLP*, Borovets, Bulgaria, 2009, pp. 214–218.

[9] A. Lardilleux, F. Yvon, and Y. Lepage, "Generalizing sampling-based multilingual alignment," *Machine Translation*, vol. 27, no. 1, pp. 1–23, 2013.

[10] ——, "Hierarchical Sub-sentential Alignment with Anymalign," in *Proceedings of EAMT*, Trento, Italy, 2012, pp. 280–286.

[11] D. Wu, "Stochastic inversion transduction grammars and bilingual parsing of parallel corpora," *Computational linguistics*, vol. 23, no. 3, pp. 377–404, 1997.

[12] Y. Deng, S. Kumar, and W. Byrne, "Segmentation and alignment of parallel text for statistical machine translation," *Natural Language Engineering*, vol. 13, no. 03, pp. 235–260, 2006.

[13] P. Koehn, A. Birch, C. Callison-burch, M. Federico, N. Bertoldi, B. Cowan, C. Moran, C. Dyer, A. Constantin, and E. Herbst, "Moses : Open Source Toolkit for Statistical Machine Translation," in *ACL, demo session*, Prague, Czech Republic, 2007, pp. 177–180.

[14] F. J. Och, "Minimum Error Rate Training in Statistical Machine Translation," in *Proceedings of ACL*, Sapporo, Japan, 2003, pp. 160–167.

[15] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of ACL*, 2002, pp. 311–318.

[16] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A study of translation edit rate with targeted human annotation," in *Proceedings of AMTA*, Cambridge, USA, 2006, pp. 223–231.

[17] J. H. Clark, C. Dyer, A. Lavie, and N. A. Smith, "Better Hypothesis Testing for Statistical Machine Translation: Controlling for Optimizer Instability," in *Proceedings of ACL*, Portland, USA, 2011, pp. 176–181.

[18] B. Marie and A. Max, "A Study in Greedy Oracle Improvement of Translation Hypotheses," in *IWSLT, Heidelberg, Germany*, 2013.

[19] R. Zens, D. Stanton, and P. Xu, "A Systematic Comparison of Phrase Table Pruning Techniques," in *Proceedings of EMNLP*, Jeju Island, Korea, 2012, pp. 972–983.

[20] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Computational linguistics*, vol. 29, no. 1, pp. 19–51, 2003.

[21] T. Takezawa, E. Sumita, F. Sugaya, H. Yamamoto, and S. Yamamoto, "Toward a Broad-coverage Bilingual Corpus for Speech Translation of Travel Conversations in the Real World," in *Proceedings of LREC*, Las Palmas, Spain, 2002.

[22] Q. Gao and S. Vogel, "Parallel implementations of word alignment tool," in *Software Engineering, Testing, and Quality Assurance for NLP*, 2008, pp. 49–57.

[23] K. Duh, K. Sudoh, and H. Tsukada, "Analysis of translation model adaptation in Statistical Machine Translation," in *Proceedings of IWSLT*, Paris, France, 2010.

[24] Q. Gao, W. Lewis, C. Quirk, and M.-Y. Hwang, "Incremental Training and Intentional Over-fitting of Word Alignment," in *Proceedings of MT Summit*, Xiamen, China, 2011.

[25] P. Banerjee, S. K. Naskar, J. Roturier, A. Way, and J. van Genabith, "Translation Quality-Based Supplementary Data Selection by Incremental Update of Translation Models," in *Proceedings of COLING*, Mumbai, India, 2012, pp. 149–166.

[26] G. Gascó, M.-A. Rocha, G. Sanchis-Trilles, J. Andrés-Ferrer, and F. Casacuberta, "Does more data always yield better translations?" in *Proceedings of EACL*, Avignon, France, 2012, pp. 152–161.

[27] H.-S. Le, T. Lavergne, A. Allauzen, M. Apidianaki, L. Gong, A. Max, A. Sokolov, G. Wisniewski, and F. Yvon, "LIMSI @ WMT12," in *Proceedings of WMT*, Montréal, Canada, 2012, pp. 330–337.

[28] A. Lopez, "Tera-Scale Translation Models via Pattern Matching," in *Proceedings of COLING*, Manchester, UK, 2008.

[29] L. Gong, A. Max, and F. Yvon, "Towards Contextual Adaptation for Any-text Translation," in *Proceedings of IWSLT*, Hong Kong, 2012.

[30] J. Bourdaillet and P. Langlais, "Identifying Infrequent Translations by Aligning Non Parallel Sentences," in *Proceedings of AMTA*, San Diego, USA, 2012.