Mining Parallel Fragments from Comparable Texts

Mauro Cettolo, Marcello Federico, Nicola Bertoldi

FBK - Fondazione Bruno Kessler
via Sommarive 18 - 38123 Povo, Trento, Italy
{cettolo,federico,bertoldi}@fbk.eu

Abstract

This paper proposes a novel method for exploiting comparable documents to generate parallel data for machine translation. First, each source document is paired to each sentence of the corresponding target document; second, partial phrase alignments are computed within the paired texts; finally, fragment pairs across linked phrase-pairs are extracted. The algorithm has been tested on two recent challenging news translation tasks. Results show that mining for parallel fragments is more effective than mining for parallel sentences, and that comparable in-domain texts can be more valuable than parallel out-of-domain texts.

1. Introduction

Statistical machine translation (SMT) technology enables rapid construction of systems for any language pair with sufficient translated text as training material. Unfortunately, large parallel corpora simply do not exist for many socially and economically relevant language pairs. To cope with data scarcity, machine learning methods have recently been devised to train models from alternative data sources.

Instead of using parallel corpora, one could consider using comparable corpora [1], such as newspaper articles written in different languages and describing the same content, which are not direct translations of each other. Although these documents are not parallel, it often happens that some portions of them are mutual translations to some extent.

The best way to exploit comparable corpora for improving the quality of SMT systems is still an open issue which has been receiving much attention by the research community in the recent years [2, 3, 4, 5].

In this work we propose a method for collecting parallel fragments of text from comparable documents which is novel in two main aspects: (i) fragments are mined from document-sentence pairs rather than from sentence-sentence pairs, (ii) fragments are detected through phrase- rather than word-level alignments. Our approach comprises three steps: first, the text of the source documents is paired to each sentence of the target document; then, a partial phrase-based alignment between the paired texts is computed; finally, fragment pairs are extracted after joining aligned phrases.

Translation experiments that assess the utility of the extracted fragments have been conducted on the German-to-English task defined by 2010 ACL SMT Workshop, and on the Arabic-to-English task defined by the 2009 NIST MT Evaluation. Results show that by augmenting the training data with parallel fragments mined from comparable in-domain texts, the BLEU score increases up to 5% relative, and that the same BLEU value would be obtained by employing an out-of-domain parallel corpus four times larger. Moreover, our method is also able to effectively mine parallel fragments from a comparable corpus aligned at the sentence level, and allows to reach the same BLEU score after filtering out one third of its content.

The paper is organized as follows. We start with a digression about the impact of parallel, comparable, and noisy training data on SMT performance. Then, we review previous literature on the exploitation of comparable corpora for training SMT. Next, we present our fragment detection procedure in detail. Hence, we report on our experiments comparing fragment versus sentence extraction, with respect to translation performance, as well as addressing noise robustness and vocabulary coverage of fragment-based training.

2. Training SMT with Additional Data

SMT models are trained on parallel texts from which, after a word-alignment stage, counters of word- and phrase-pairs are used to estimate translation probabilities. The resulting model embeds salient or dominant features of the training data, both from the linguistic and domain perspectives. Increasing the amount of training data affects the learned probability distributions in a way that it depends on the nature of the additional texts. If the new texts are “consistent” with (i.e. they are generated from the same source of) the training data, then the translation model is reinforced. On the contrary, in the case of source mismatch, the addition of data should be applied with care. A typical example case is the *adaptation* in SMT, which considers several ways to combine in-domain and out-of-domain parallel data (see for example [6, 7, 8, 9]). Another paradigmatic experimental condition is when supplementary in-domain data are available under form of *comparable* texts. This case poses an interesting problem, which is how to optimally exploit parallel information contained in comparable data.

As a starting point of our work, in Figure 1 we have plotted three curves reporting translation performance (BLEU

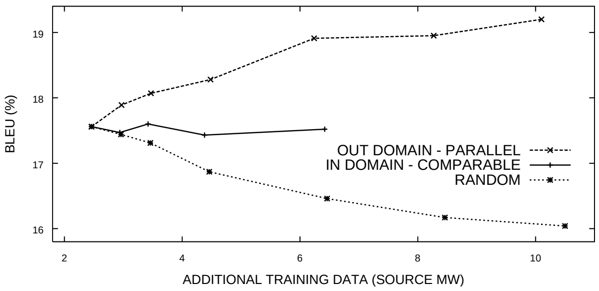


Figure 1: Effects on performance after increasing training text with different sources of bilingual data.

score) achieved by our baseline, that will be described in detail later, after augmenting the training data in three ways: (i) with out-of-domain parallel data, (ii) with in-domain (unfiltered) comparable data, (iii) with noisy bilingual data. The latter were generated by randomly aligning sentence pairs of the out-of-domain corpus and are considered just to see how quickly random noise corrupts the original translation models.

In few words, the plot tells the following: by adding out-of-domain parallel data, performance improves, although the improvement rate is quite slow. No surprise that the opposite happens when training data are corrupted with random data. An intermediate behavior is observed instead when comparable in-domain text is added without any smart filtering: performance fluctuates around the baseline score, indicating that the model probably learns a mix of correct and wrong translations.

An example of comparable document used for our German-to-English translation task is shown in Figure 2. It is evident that both sides report the same news, but most of the text is not parallel. In fact, sentences conveying almost the same content in a similar manner (marked by solid links) are quite rare. Nevertheless, not only some of them can be detected automatically (dotted boxes), but also a better and more refined detection of corresponding fragments is feasible. Text in the squared brackets shows the output of our automatic technique.

The core idea of this paper is about how to extract such quasi-parallel fragments from comparable texts so that performance resulting from these data favorably compare with the curves in Figure 1.

3. Related Work

Nowadays, several international news agencies deliver content through the Web in many languages. This represents a formidable opportunity to collect comparable corpora, that is texts expressing the same content in different languages. Notice that in this work we skip the problem of aligning multilingual documents, since we assume that they already include metadata allowing this linking, as in the case considered here.

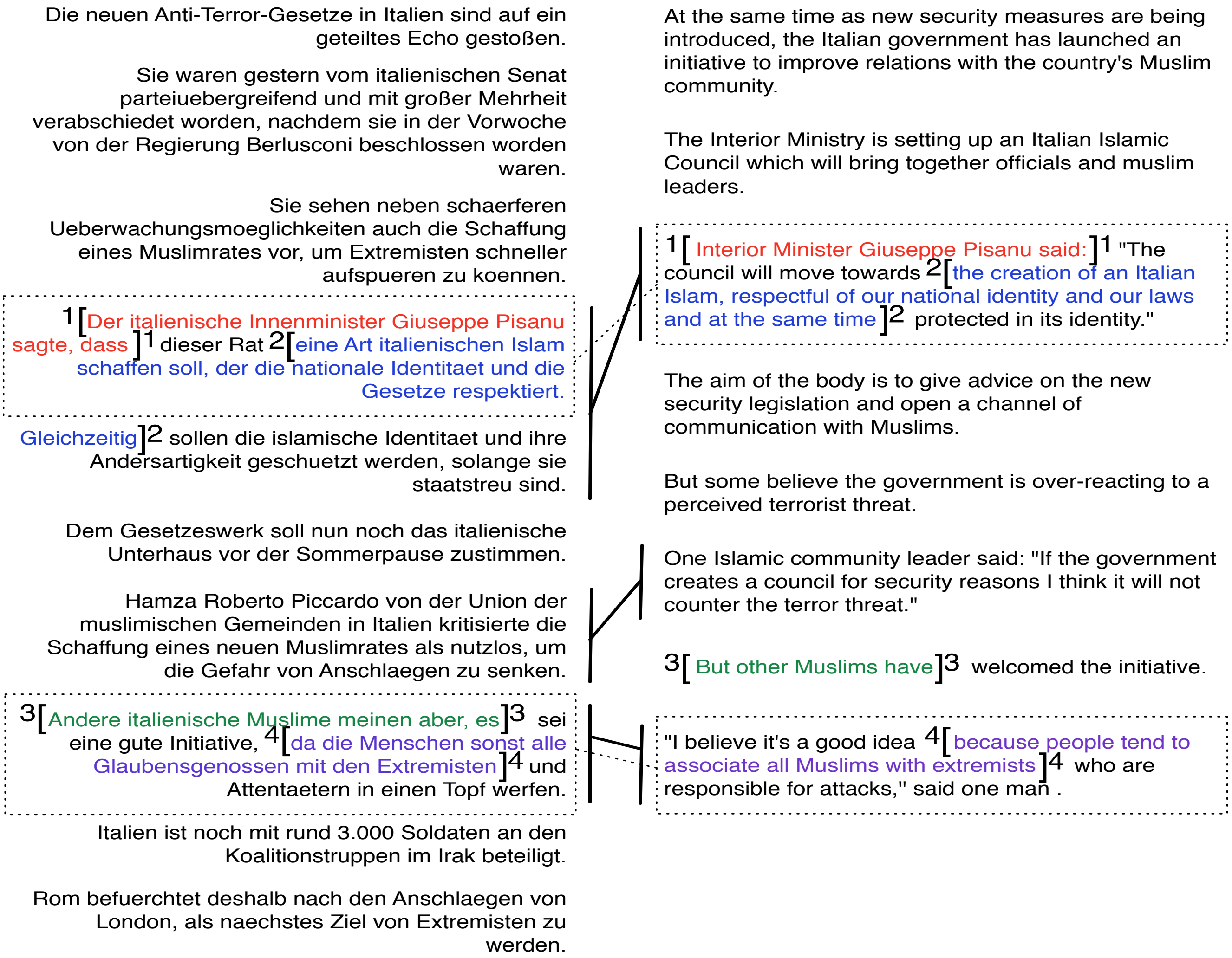
Starting from a collection of paired documents, several approaches have recently been proposed in the literature to extract parallel excerpts. Most of the techniques, if not all, share the stages of splitting documents into sentences and of pairing sentences across documents. Methods significantly differ in the successive filtering steps, that can be clustered into two main groups: procedures aiming at deciding if paired sentences are mutual translations or at extracting parallel sub-sentential fragments. In the following, we briefly describe a few works tightly related to our approach, which also report significant performance improvements.

In [2] words in the source documents are translated through a bilingual lexicon; all possible sentence pairs of the two documents are passed first through a word-overlap filter and then classified as parallel or not by a maximum entropy classifier trained on (a small amount of) parallel sentences.

In [3] a SMT system, trained on a small amount of parallel data, is used to directly translate the source side of the documents. Then, instead of the maximum entropy classifier, WER and TER scores are computed at the sentence level by comparing the translations with the target side; the amount of filtered pairs can be tuned by varying the acceptance threshold.

The approach in [4] resembles [2] up to the definition of the set of candidate sentence pairs. Instead of deciding whether the two sentences are mutual translations, now they search for parallel fragments using an approach inspired by signal processing. Using a set of parameters derived from Log-Likelihood-Ratio statistics, each word is annotated with values in $[-1, +1]$ indicating the likelihood that the word has some translations in the other sentence by performing a greedy alignment. This stream of values is then treated as a signal and passed through an averaging filter. Spans that have only positive signal values and are longer than a threshold (3 words) are considered more likely to have a translation on the other side. The same process is repeated on the other translation direction, and the resulting fragment pair is assumed to be parallel.

Quirk et al. [5] try to overcome some of the limitation of the approach described in [4] in the way parallel fragments are identified. In particular, they propose two generative models for generating noisy target sentences from the source sentences. One model is employed to align words in candidate sentence pairs; fragments are then extracted from alignments by applying simple heuristics. Another model tries to directly generate fragments; in the process, three generation options are competing: source-only fragment, target-only fragment, or joint source-target fragment. The rational behind this model is that “the probability of generating source and target fragments jointly should be more likely than generating them independently if and only if they are parallel”. The latter model is definitely more complex than the former one, although their impact on SMT performance is similar.



Die neuen Anti-Terror-Gesetze in Italien sind auf ein geteiltes Echo gestoßen.

Sie waren gestern vom italienischen Senat parteiuebergreifend und mit großer Mehrheit verabschiedet worden, nachdem sie in der Vorwoche von der Regierung Berlusconi beschlossen worden waren.

Sie sehen neben schaerferen Ueberwachungsmoeglichkeiten auch die Schaffung eines Muslimrates vor, um Extremisten schneller aufspueren zu koennen.

1[Der italienische Innenminister Giuseppe Pisanu sagte, dass]1 dieser Rat 2[eine Art italienischen Islam schaffen soll, der die nationale Identitaet und die Gesetze respektiert.

Gleichzeitig]2 sollen die islamische Identitaet und ihre Andersartigkeit geschuetzt werden, solange sie staatstreu sind.

Dem Gesetzeswerk soll nun noch das italienische Unterhaus vor der Sommerpause zustimmen.

Hamza Roberto Piccardo von der Union der muslimischen Gemeinden in Italien kritisierte die Schaffung eines neuen Muslimrates als nutzlos, um die Gefahr von Anschlaegen zu senken.

3[Andere italienische Muslime meinen aber, es]3 sei eine gute Initiative, 4[da die Menschen sonst alle Glaubensgenossen mit den Extremisten]4 und Attentaetern in einen Topf werfen.

Italien ist noch mit rund 3.000 Soldaten an den Koalitionstruppen im Irak beteiligt.

Rom befuerchtet deshalb nach den Anschlaegen von London, als naechstes Ziel von Extremisten zu werden.

At the same time as new security measures are being introduced, the Italian government has launched an initiative to improve relations with the country's Muslim community.

The Interior Ministry is setting up an Italian Islamic Council which will bring together officials and muslim leaders.

1[Interior Minister Giuseppe Pisanu said:]1 "The council will move towards 2[the creation of an Italian Islam, respectful of our national identity and our laws and at the same time]2 protected in its identity."

The aim of the body is to give advice on the new security legislation and open a channel of communication with Muslims.

But some believe the government is over-reacting to a perceived terrorist threat.

One Islamic community leader said: "If the government creates a council for security reasons I think it will not counter the terror threat."

3[But other Muslims have]3 welcomed the initiative.

"I believe it's a good idea 4[because people tend to associate all Muslims with extremists]4 who are responsible for attacks," said one man .

Figure 2: The same news from EuroNews published in German and English, with manually detected semantically equivalent sentences (solid links, 3 instances), automatically detected parallel sentences (dotted boxes, 2) and fragments (squared brackets, 4).

As we will see, our method resembles some of the features of the above mentioned works, but it differs in some aspects: we propose to look for parallel fragments in text pairs where the source side is the whole document rather than the single sentence, with the aim of improving the fragment detection recall. Moreover, the actual extraction of fragments relies on a phrase- rather than word-level alignment, exploiting in this way the job already done in building the phrase pair resource. Finally, Quirk et al. [5] have to limit the maximum fragment length to make their approach computationally tractable. Instead, we achieve the same goal without this restriction, because we exploit the bound on the phrase size which commonly exists in SMT systems.

4. Parallel Fragment Detection

In this section a novel scheme is proposed for mining parallel fragments from a bilingual document.

We assume to have access to a repository of source/target phrase pairs, where a phrase has the common meaning given in SMT, i.e. a sequence of (one or more) contiguous words. This assumption is reasonable, given that almost any kind of translation system relies on a similar bilingual resource, in the form of either a bilingual dictionary, a translation table, or translation examples. Moreover, the quite common situation we are considering is that of improving a phrase-based SMT system: so, its translation model is just what we need as repository.

The scheme consists of three modules applied in cascade: in the first, comparable bilingual texts are paired; then, the

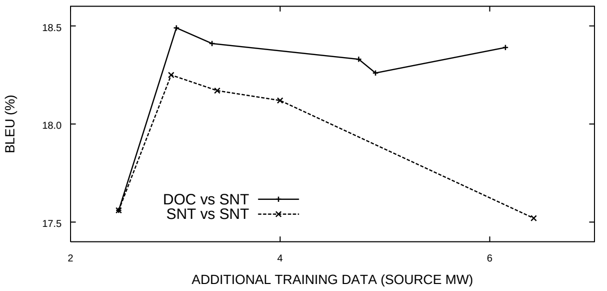


Figure 3: Effects on performance after increasing training text with fragments from different kinds of text pairing: source documents vs. target sentences and source vs. target sentences.

aligner computes the best alignment of the paired texts at phrase level given a phrase-pair repository; finally, the *extractor* exploits that alignment for detecting parallel fragments.

4.1. Text Pairing

As already stated in Section 3, we assume that documents have already been paired. At this point, documents are typically split into sentences, for example on the basis of strong punctuation, and then source and target sentences are aligned by exploiting some knowledge resource about the translation process, e.g. bilingual dictionary, alignment models, etc. Actually, we started pairing sentences relying on the IBM model 1 and obtained good results in terms of fragment extraction. However, an even better behavior was observed by means of a quite unusual pairing: the whole source document against each sentence occurring in the target document. Figure 3 highlights the gain given by the document-sentence pairing over the sentence-sentence one, supporting the adoption of that scheme in our experiments. In Section 5.4 this issue will be further commented.

4.2. Phrase-level Alignment

Given a text pair, the algorithm searches for the partial phrase-level alignment which achieves the best trade-off between the maximum coverage of both source and target texts and the minimum distortion of the source side.

More in detail, it takes a text pair $(\mathbf{f}, \mathbf{e})$ as input; in the first stage, the set of translation options $TrOpt(\mathbf{f})$ related to $\mathbf{f}$ is extracted from the repository of phrase pairs. The selected translation options are those entries of the repository whose source side matches any substring (of contiguous words) in the source text. For the sake of efficiency, a translation option also includes the actual span of the matched source phrase. Moreover, $TrOpt(\mathbf{f})$ is structured in such a way that, if queried directly using the target phrase as a key, it returns all the corresponding source spans. Note that different translation options can share the same source/target phrase pair, but refer to different source spans.

Definitions:

$\mathbf{f} = f_1^m$: source text of length $m = |\mathbf{f}|$

$\mathbf{e} = e_1^l$: target text of length $l = |\mathbf{e}|$

$[i', i]$: span of target positions

$[j', j]$: span of source positions

$\mathcal{L}(k', k)$: length of any span $[k', k]$

$\mathcal{D}(j', j)$: distortion of a source span $[j', j]$[1]

C: subsets of source positions

$\mathcal{T}(i', i)$: set of source spans $[j', j]$ such that $([j', j], e_{i'}^i) \in TrOpt(\mathbf{f})$

$Q(i, C)$: optimal score of aligning the target span $[1, i]$ to source positions C

$S(i', i, j', j)$: score of aligning a target span to a (possibly empty) source span.

DP-based Constrained Search:

$$\begin{aligned} Q(0, \emptyset) &= 0 \\ Q(i, C) &= \max_{\substack{i' \leq i \\ [j', j] \in \mathcal{T}(i', i)}} \begin{Bmatrix} Q(i', C \setminus [j', j]) + \\ +S(i', i, j', j) \end{Bmatrix} \\ Q^* &= \max_C Q(l, C) \end{aligned}$$

Model:

$$S(i', i, j', j) = \mathcal{L}(i', i) + \mathcal{L}(j', j) + \mathcal{D}(j', j)$$

Figure 4: Constrained DP-search for phrase-level alignment. For simplicity, state bookkeeping is not reported.

Then, the algorithm searches for the best phrase-level alignment of the pair $(\mathbf{f}, \mathbf{e})$ achievable by using $TrOpt(\mathbf{f})$. By means of the DP-based procedure sketched in Figure 4, the alignment is created incrementally by covering the target text left-to-right. At each iteration, a new target phrase $e_{i'}^i$ is aligned with any source span returned by $TrOpt(\mathbf{f})$, which does not overlap previously covered source positions. Note that the new target phrase has to be contiguous to the previous aligned ones. Finally, the optimal alignment is searched among those fully covering the target text.

The score of each single expansion takes into account the lengths of the aligned phrases and the distortion computed as the distance between the first position of the current source span and the last position of the previously aligned source span.[1]

The depicted algorithm outputs the score of the solu-

[1] To simplify the notation of the formulas, the distortion score does not include the needed dependency from the previously covered source span.

tion, but the changes to provide the actual best alignment are straightforward. Moreover, standard approximations (beam search, histogram pruning...) are omitted, but implemented.

Some aspects of the algorithm deserve to be highlighted and commented. First of all, translation probabilities are not used at all. This allows the exploitation of any phrase pair repository, even lacking of probabilities (like multiwordnets), and prevents hypotheses built on low probability phrase pairs from cutting by the beam search. In our specific case, the SMT phrase table used as repository is likely to be noisy. In order to have the repository as clean as possible, the phrase table is pruned via the algorithm described in [10].

Secondly, the use of a phrase-based translation model allows us to cover phrases instead of single words, differently from what is done in similar approaches (e.g. [5] but also [4]). This way the job done in SMT training for discovering reliable phrase pairs is exploited: if such pairs occur in the bilingual input text, they represent a solid anchor for the fragment extraction.

Finally, it is worth to noticing that the algorithm permits partial alignment: portions of either source or target texts can remain unaligned. This is achieved by (i) adding dummy translation options (i.e. a target phrase associated with empty source span and phrase) to $TrOpt(\mathbf{f})$ for each target word, and (ii) the fact that C can be a proper subset of the positions of $\mathbf{f}$. This makes our algorithm robust in the sense that it is able to handle texts not observed in training (which is the usual case).

4.3. Fragment Extraction

Given a bilingual text pair aligned at the phrase level, parallel fragments are mined by an iterative algorithm which merges pairs that are either *contiguous* (i.e. in contact) or *proximate* (i.e. close but not necessary in contact) and enough long, including the non-aligned text in between.

It starts by considering the aligned phrases as parallel blocks. At each iteration, blocks are merged if either (i) they are contiguous on one side and at most one unaligned word intervenes on the other, or (ii) they are long and proximate enough. Two thresholds limit the bounds in (ii) and, hence, control the amount of the extracted fragment pairs.

Iterations stop when no block pair can be further merged. Finally, blocks are output as parallel fragments, unless they are too short.

5. Experiments

We empirically evaluated our fragment detection method by directly measuring the impact of the extracted data on the translation quality of two SMT systems.

First, we will see that fragment mining overcomes the filtering of full sentences from noisy bilingual texts. We will also show that the extraction of parallel fragments from in-domain comparable corpora is more effective than the use of out-of-domain parallel data. Finally, we will provide evidence that our method is able to effectively mine fragments from a corpus already cleaned at the sentence level.

task	running words		dictionary size		phrase pairs
	src	tgt	src	tgt	
De-En	2.5M	2.4M	106K	54K	374K
Ar-En	6.0M	6.1M	93K	77K	1.3M

Table 1: Statistics of the De-En/Ar-En translation/reordering models: size of training texts (running words), dictionary size, and number of phrase pairs in the baselines.

5.1. Data

Experiments were conducted on two different tasks and language pairs. In the first task, German news are translated into English (De-En) according to the setup established in the Workshop on Statistical Machine Translation of the ACL 2010.[2] Parallel training data consist of a small in-domain corpus (News Commentary - NC) and a larger out-of-domain corpus (Europarl [11], version 5 - EP). News-test2008 has been used for development, while news-test2009 (TST09) and news-test2010 (TST10) for testing purposes. As comparable data, we used a set of 25,517 bilingual documents downloaded from the multilingual and pan-European television news channel EuroNews (EN),[3] for a total of 4.3 million and 4.7 million German and English words, respectively.

The second task involves the translation of news from Arabic into English (Ar-En) in the framework defined by the 2009 NIST evaluation campaign.[4] In this case, for development and testing purposes the portions containing newswires of the 2006 development set and of both 2008 and 2009 evaluation sets have been employed. We used only one of the parallel resources allowed for the constrained training condition, namely the ISI Arabic-English Automatically Extracted Parallel Text (LDC2007T08). It consists of sentence pairs extracted automatically from the Arabic and English monolingual Gigaword corpora by means of the method described in [2]. For each sentence pair, a confidence score (between 0.5 and 1.0) is provided, which is indicative of its degree of parallelism.

5.2. Baselines

The baseline systems are built upon the open-source MT toolkit Moses [12].[5] The translation and the lexicalized reordering models have been trained on NC for De-En, and on the subset of ISI corpus containing sentences with confidence score larger than 0.993 (ISI-0.993), for Ar-En. Phrase tables are pruned according to [10]. In all experiments, 6-gram LMs have been employed, smoothed with the improved Kneser-Ney technique [13] and computed with the IRSTLM

[2] www.statmt.org/wmt10/

[3] www.euronews.net

[4] www.itl.nist.gov/iad/mig/tests/mt/2009/

[5] www.statmt.org/moses/

task	running words	dictionary size	6-grams
De-En	1.18G	2.0M	512M
Ar-En	147M	447K	13.7M

Table 2: Statistics of the De-En and Ar-En LMs: size of training texts (running words), dictionary size, and number of 6-grams in the baselines.

training data				TST09	TST10
baseline		additional			
running src words	type	running src words	type		
2.5M	NC	-	-	16.43	17.56
2.5M	NC	0.5M	FRG(EN)	17.09	18.49
2.5M	NC	0.5M	SNT(EP)	16.52	17.89
2.5M	NC	2.0M	SNT(EP)	17.06	18.28

Table 3: Performance by adding in-domain fragments and two different amounts of out-of-domain sentences.

toolkit [14]. The monolingual resources made available by the event organizers have been exploited: for the De-En task, the English side of NC and EP, and a large corpus of news; for the Ar-En task, the English side of the allowed parallel training data (GALE, ISI, UN and others smaller). Tables 1 and 2 provide some statistics of the baseline models.

The weights of the log-linear interpolation model have been optimized on the dev sets by means of the standard MERT procedure, not only for baselines but also for all other systems employing additional parallel training data.

The translation models of the baselines have been used as repository of phrase pairs required for mining parallel fragments with our method, in particular by the algorithm in Figure 4.

5.3. Results

De-En task Figure 5 plots the BLEU score on the two test sets when additional parallel data are extracted from EN according to three different schemes. (i) Fragments are detected by our method; the amount of additional data is varied by changing the setup (thresholds) of the fragment extraction algorithm presented in Section 4.3. (ii) First, the optimal sentence alignment for each document pair is computed given the IBM model 1 (from the baseline system), and then the paired sentences are filtered according to the scheme proposed in [3] (see Section 3) on the basis of the TER score; the threshold on the TER score allows to vary the amount of additional training data. (iii) The same filtering of (ii) is applied to each possible pair of source/target sentences; this permits the same source sentence to occur more than once in the additional corpus, linked to different target sentences. In the figure, the three approaches are named FRG, SNT, and SxS, respectively.

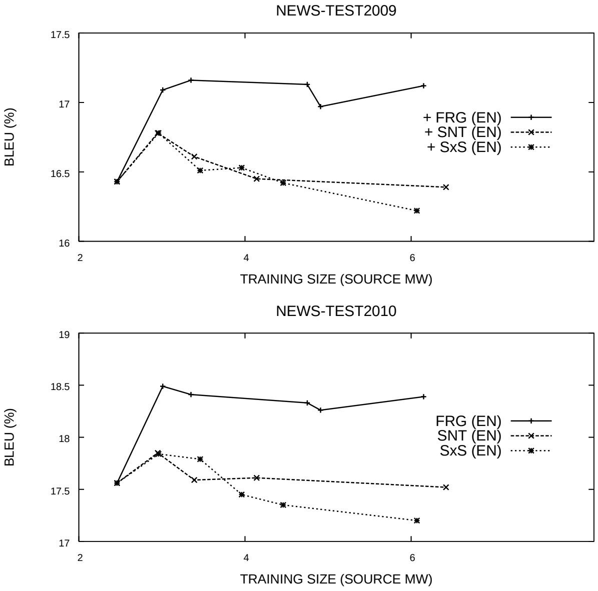


Figure 5: De-En task: BLEU score as a function of the amount of parallel training data. Additional data are automatically extracted from EN.

It is evident that our method is effective in selecting parallel data from the considered comparable corpus. The peak is reached by adding around 0.5-1 million words, that is about 10-20% of the whole EN size: hence, this could represent an estimate of the rate of quasi-parallel text inside EN.

The two attempts of filtering EN data at the sentence level via TER gave interesting but lower improvements. Note also that after the peak, performance with FRG remain more stable than with the sentence-based approaches. This suggests that fragment extraction based on the document-sentence pairing is more robust with respect to the chosen working point (setup), fact that makes its “tuning” less problematic.

Table 3 allows to compare the use of in-domain fragments selected by our method and that of out-of-domain (EP) sentences. In the first row, baseline performance are reported; the second row refers to the system trained on NC plus fragments selected by our method in its optimal setup (peaks of the FRG(EN) curves in Fig. 5). The other two rows refer to systems trained with an additional amount of sentences from EP in such a way that either the total amount of training data or the performance are the same of the system augmented with fragments. Fragments yield a 4-5% relative improvement of the baseline. Adding the same amount of data, in-domain fragments allows a higher BLEU of more than 3% relative than out-of-domain sentences; for reaching the same level of performance, more than four times of out-of-domain data is required.

Ar-En task Figure 6 plots the BLEU score on the two test sets of systems built on top of the baseline by adding training data from the remaining part of the ISI corpus: (i)

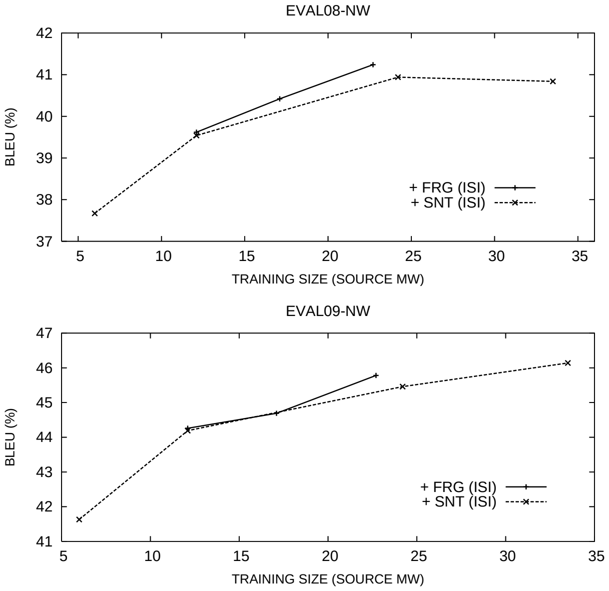


Figure 6: Ar-En task: BLEU score as a function of the amount of parallel training data. Additional parallel fragments are automatically extracted from the ISI corpus.

through our fragment selection method; (ii) simply selecting sentence pairs in the order defined by the confidence score.

Differently from the previous experiment, fragments have now been extracted from an already filtered corpus, which is considered enough clean to be included among the parallel resources of the NIST evaluation campaign. Nevertheless, our method still remains effective by permitting to reach the scores obtained with the whole ISI corpus with only two third of it (22.7M words vs. 33.5M).

5.4. Detailed Analysis

As stated in Section 4.1, our method looks for parallel fragments by pairing whole source documents with single target sentences. Differently from the commonly used sentence vs. sentence approach, it allows to extract parallel portions which are spread over more sentences. For example, in the fragment pair number 2 (Figure 2) all English words belong to the same sentence, while the German ones occur in two different sentences. This means that working at the sentence level would have prevented the identification of such a fragment pair.

Concerning the choice of working at the fragment rather than sentence level, one could argue that this prevents the proper modeling of context, as most fragments, by definition, are not full sentences. In our opinion this does not represent a real problem for the translation model. In fact, during decoding the target phrases are appended to the target string only on the basis of the source counterpart, independently from the context. Good results obtained by adding parallel fragments to training data support this belief.

Two aspects of the algorithm for fragment detection (Section 4.3) deserve to be noticed: the addition of new words to the original model and the risk of introducing noise. Concerning the first issue, the algorithm is able to generate fragments with new words thanks to the conditions defined for merging proximate blocks; notably, Figure 7 shows that adding fragments from comparable but in-domain data allows to reduce the OOV rate much faster than the parallel but out-of-domain corpus.

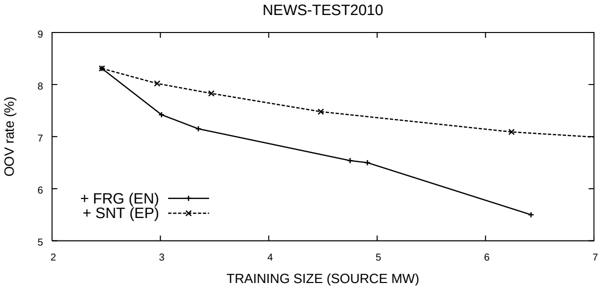


Figure 7: OOV rate of the source side of De-En newstest2010 test set adding either in-domain fragments (from EN) or out-of-domain sentences (from EP).

On the other hand, enlarging the fragments with nonaligned text does not guarantee that the words introduced in this way are really parallel. In fact, it happens that parallel fragments are noisy, as shown in Figure 2. For instance, the first fragment-pair (number 1), despite its good parallelism includes noisy tokens, like “italienische”.

From this fragment pair, the successive training steps would add new phrase pairs to the baseline translation model, such as:

Der italienische Innenminister Giuseppe Pisanu
⇒ Interior Minister Giuseppe Pisanu
Innenminister Giuseppe Pisanu ⇒ Giuseppe Pisanu
Pisanu ⇒ Pisanu

The first phrase pair is the one that does not survive the pruning step; likely, this is due to the presence of the German word “italienische” which is well known to the model but does not have a counterpart on the English side. On the contrary, the other two entries are kept in the final model. Both allow the system to know the formerly OOV name of the Italian Interior Minister. Unfortunately, the words specifying the post, which are indeed contained in the fragment, have been included into the phrase pair only on the German side, making it wrong.

In Table 4 some statistics are provided on the phrase tables resulting by using either the small ISI-0.993 corpus, the whole ISI corpus (+SNT(ISI)), or ISI-0.993 plus the fragments extracted from the remaining part. The size difference of training data explains the difference of the number of entries of the unpruned phrase tables. Instead, after pruning by means of the algorithm in [10], the phrase table trained on fragments is larger than that built on the whole ISI corpus; this means that many of the new phrase pairs generated from the fragments are judged reliable by the pruning algorithm. Finally, one notation on the average length of phrases: if fragments are used, the resulting phrases are longer by 10%, which in general represents a facilitating condition for the translation process.

training	running		phrase	
data	words (M)		pairs (M)	
	source	target	unprun.	pruned
ISI-0.993	6.0	6.1	13.7	1.3
+FRG(ISI)	22.7	23.6	32.4	9.2
+SNT(ISI)	33.5	33.1	69.6	8.3

Table 4: Ar-En phrase tables statistics: amount of training running words and num. of entries before/after pruning.

6. Conclusions

In this work we have considered the problem of optimally exploiting parallel information contained in comparable data for improving SMT performance. In fact, we have shown that even if comparable data are in-domain, without the application of proper filters their inclusion in the training data does not produce positive effects.

We have then designed a novel method for mining parallel fragments from comparable documents. The experimental assessment has been done by directly measuring the impact of additional training resources on the translation/reordering models of two SMT systems. Main empirical outcomes are that: (i) the method is effective in distilling useful parallel data from comparable resources; (ii) fragments are preferable to whole sentences because they are less noisy and the alleged loss of contextual information has a little impact on the resulting models.

In the future, our priority is to apply the method to larger non parallel corpora downloaded from the Web on a daily basis. Assuming their equivalence to EuroNews, our experiments suggest that we can reasonably expect to mine about 10 to 20% of good parallel texts to be used as additional training data. Besides improving the system, this will allow us to further investigate the learning curve with larger amounts of fragments.

7. Acknowledgements

This work was supported by the EuroMatrixPlus project (IST-231720), which is funded by the EC under the 7^{th} Framework Programme for Research and Technological Development. The authors also thank Gabriele Musillo and Yashar Mehdad for their helpful comments.

8. References

[1] P. Fung and P. Cheung, “Mining very-non-parallel corpora: Parallel sentence and lexicon extraction via bootstrapping and EM,” in *Proc. of EMNLP*, Barcelona, Spain, pp. 57–63, 2004.

[2] D. S. Munteanu and D. Marcu, “Improving machine translation performance by exploiting non-parallel corpora,” *Computational Linguistics*, 31(4):477–504, ’05.

[3] S. Abdul-Rauf and H. Schwenk, “On the use of comparable corpora to improve SMT performance,” in *Proc. of EACL*, Athens, Greece, pp. 16–23, 2009.

[4] D. S. Munteanu and D. Marcu, “Extracting parallel subsentential fragments from non-parallel corpora,” *Proc. of ACL*, Sydney, Australia, pp. 81–88, 2006.

[5] C. Quirk, R. Udupa, and A. Menezes, “Generative models of noisy translations with applications to parallel fragment extraction,” in *Proc. of MT Summit*, Copenhagen, Denmark, 2007.

[6] L. Nepveu, G. Lapalme, P. Langlais, and G. Foster, “Adaptive language and translation models for interactive machine translation,” in *Proc. of EMNLP*, Barcelona, Spain, pp. 190–197, 2004.

[7] N. Ueffing, G. Haffari, and A. Sarkar, “Semi-supervised model adaptation for statistical machine translation,” *Machine Translation*, 21(2):77–94, 2007.

[8] P. Koehn and J. Schroeder, “Experiments in domain adaptation for statistical machine translation,” in *Proc. of the ACL Workshop on SMT*, Prague, Czech Republic, pp. 224–227, 2007.

[9] H. Schwenk and J. Senellart, “Translation model adaptation for an Arabic/French news translation system by lightly-supervised training,” in *Proc. of MT Summit*, Ottawa, Canada, 2009.

[10] H. Johnson, J. Martin, G. Foster, and R. Kuhn, “Improving translation quality by discarding most of the phrasetable,” in *In Proc. of EMNLP-CoNLL*, Prague, Czech Republic, pp. 967–975, 2007.

[11] P. Koehn, “Europarl: A parallel corpus for statistical machine translation,” in *Proc. of MT Summit*, Phuket, Thailand, pp. 79–86, 2005.

[12] P. Koehn et al., “Moses: Open source toolkit for statistical machine translation,” in *Proc. of ACL - Demo and Poster Sessions*, Prague, Czech Republic, pp. 177–180, 2007.

[13] S. F. Chen and J. Goodman, “An empirical study of smoothing techniques for language modeling,” *Computer Speech and Language*, 4(13):359–393, ’99.

[14] M. Federico, N. Bertoldi, and M. Cettolo, “Irstlm: an open source toolkit for handling large scale language models,” in *Proc. of Interspeech*, Melbourne, Australia, pp. 1618–1621, 2008.