On the use of Comparable Corpora to improve SMT performance

Sadaf Abdul-Rauf and **Holger Schwenk**

LIUM, University of Le Mans, FRANCE
Sadaf.Abdul-Rauf@lium.univ-lemans.fr

Abstract

We present a simple and effective method for extracting parallel sentences from comparable corpora. We employ a statistical machine translation (SMT) system built from small amounts of parallel texts to translate the source side of the non-parallel corpus. The target side texts are used, along with other corpora, in the language model of this SMT system. We then use information retrieval techniques and simple filters to create French/English parallel data from a comparable news corpora. We evaluate the quality of the extracted data by showing that it significantly improves the performance of an SMT systems.

1 Introduction

Parallel corpora have proved be an indispensable resource in Statistical Machine Translation (SMT). A parallel corpus, also called bitext, consists in bilingual texts aligned at the sentence level. They have also proved to be useful in a range of natural language processing applications like automatic lexical acquisition, cross language information retrieval and annotation projection.

Unfortunately, parallel corpora are a limited resource, with insufficient coverage of many language pairs and application domains of interest. The performance of an SMT system heavily depends on the parallel corpus used for training. Generally, more bitexts lead to better performance. Current resources of parallel corpora cover few language pairs and mostly come from one domain (proceedings of the Canadian or European Parliament, or of the United Nations). This becomes specifically problematic when SMT systems trained on such corpora are used for general translations, as the language jargon heavily used in these corpora is not appropriate for everyday life translations or translations in some other domain.

One option to increase this scarce resource could be to produce more human translations, but this is a very expensive option, in terms of both time and money. In recent work less expensive but very productive methods of creating such sentence aligned bilingual corpora were proposed. These are based on generating "parallel" texts from already available "almost parallel" or "not much parallel" texts. The term "comparable corpus" is often used to define such texts.

A comparable corpus is a collection of texts composed independently in the respective languages and combined on the basis of similarity of content (Yang and Li, 2003). The raw material for comparable documents is often easy to obtain but the alignment of individual documents is a challenging task (Oard, 1997). Multilingual news reporting agencies like AFP, Xinghua, Reuters, CNN, BBC etc. serve to be reliable producers of huge collections of such comparable corpora. Such texts are widely available from LDC, in particular the Gigaword corpora, or over the WEB for many languages and domains, e.g. Wikipedia. They often contain many sentences that are reasonable translations of each other, thus potential parallel sentences to be identified and extracted.

There has been considerable amount of work on bilingual comparable corpora to learn word translations as well as discovering parallel sentences. Yang and Lee (2003) use an approach based on dynamic programming to identify potential parallel sentences in title pairs. Longest common sub sequence, edit operations and match-based score functions are subsequently used to determine confidence scores. Resnik and Smith (2003) propose their STRAND web-mining based system and show that their approach is able to find large numbers of similar document pairs.

Works aimed at discovering parallel sentences

Proceedings of the 12th Conference of the European Chapter of the ACL, pages 16–23,
Athens, Greece, 30 March – 3 April 2009.

French: *Au total, 1,634 million d'électeurs doivent désigner les 90 députés de la prochaine législature parmi 1.390 candidats présentés par 17 partis, dont huit sont représentés au parlement.*
Query: *In total, 1,634 million voters will designate the 90 members of the next parliament among 1.390 candidates presented by 17 parties, eight of which are represented in parliament.*
Result: *Some 1.6 million voters were registered to elect the 90 members of the legislature from 1,390 candidates from 17 parties, eight of which are represented in parliament**, several civilian organisations and independent lists.***

French: *"Notre implication en Irak rend possible que d'autres pays membres de l'Otan, comme l'Allemagne par exemple, envoient un plus gros contingent" en Afghanistan, a estimé M.Belka au cours d'une conférence de presse.*
Query: *"Our involvement in Iraq makes it possible that other countries members of NATO, such as Germany, for example, send a larger contingent in Afghanistan, "said Mr.Belka during a press conference.*
Result: *"Our involvement in Iraq makes it possible for other NATO members, like Germany for example, to send troops, to send a bigger contingent to your country, "Belka said at a press conference**, with Afghan President Hamid Karzai.***

French: *De son côté, Mme Nicola Duckworth, directrice d'Amnesty International pour l'Europe et l'Asie centrale, a déclaré que les ONG demanderaient à M.Poutine de mettre fin aux violations des droits de l'Homme dans le Caucase du nord.*
Query: *For its part, Mrs Nicole Duckworth, director of Amnesty International for Europe and Central Asia, said that NGOs were asking Mr Putin to put an end to human rights violations in the northern Caucasus.*
Result: *Nicola Duckworth, head of Amnesty International's Europe and Central Asia department, said the non-governmental organisations (NGOs) would call on Putin to put an end to human rights abuses in the North Caucasus**, including the war-torn province of Chechnya.***

Figure 1: Some examples of a French source sentence, the SMT translation used as query and the potential parallel sentence as determined by information retrieval. Bold parts are the extra tails at the end of the sentences which we automatically removed.

include (Utiyama and Isahara, 2003), who use cross-language information retrieval techniques and dynamic programming to extract sentences from an English-Japanese comparable corpus. They identify similar article pairs, and then, treating these pairs as parallel texts, align their sentences on a sentence pair similarity score and use DP to find the least-cost alignment over the document pair. Fung and Cheung (2004) approach the problem by using a cosine similarity measure to match foreign and English documents. They work on "very non-parallel corpora". They then generate all possible sentence pairs and select the best ones based on a threshold on cosine similarity scores. Using the extracted sentences they learn a dictionary and iterate over with more sentence pairs. Recent work by Munteanu and Marcu (2005) uses a bilingual lexicon to translate some of the words of the source sentence. These translations are then used to query the database to find matching translations using information retrieval (IR) techniques. Candidate sentences are determined based on word overlap and the decision whether a sentence pair is parallel or not is performed by a maximum entropy classifier trained on parallel sentences. Bootstrapping is used and the size of the learned bilingual dictionary is increased over iterations to get better results.

Our technique is similar to that of (Munteanu and Marcu, 2005) but we bypass the need of the bilingual dictionary by using proper SMT translations and instead of a maximum entropy classifier we use simple measures like the word error rate (WER) and the translation error rate (TER) to decide whether sentences are parallel or not. Using the full SMT sentences, we get an added advantage of being able to detect one of the major errors of this technique, also identified by (Munteanu and Marcu, 2005), i.e, the cases where the initial sentences are identical but the retrieved sentence has

a tail of extra words at sentence end. We try to counter this problem as detailed in 4.1.

We apply this technique to create a parallel corpus for the French/English language pair using the LDC Gigaword comparable corpus. We show that we achieve significant improvements in the BLEU score by adding our extracted corpus to the already available human-translated corpora.

This paper is organized as follows. In the next section we first describe the baseline SMT system trained on human-provided translations only. We then proceed by explaining our parallel sentence selection scheme and the post-processing. Section 4 summarizes our experimental results and the paper concludes with a discussion and perspectives of this work.

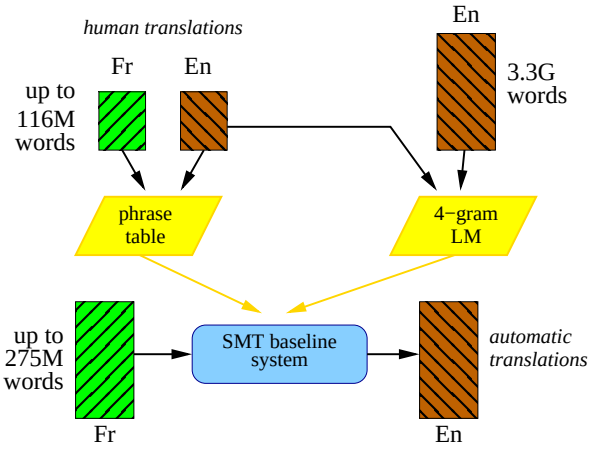

Figure 2: Using an SMT system used to translate large amounts of monolingual data.

2 Baseline SMT system

The goal of SMT is to produce a target sentence e from a source sentence f. Among all possible target language sentences the one with the highest probability is chosen:

$$e^* = \arg\max_e \Pr(e|f) \quad (1)$$

$$= \arg\max_e \Pr(f|e)\Pr(e) \quad (2)$$

where $\Pr(f|e)$ is the translation model and $\Pr(e)$ is the target language model (LM). This approach is usually referred to as the *noisy source-channel* approach in SMT (Brown et al., 1993). Bilingual corpora are needed to train the translation model and monolingual texts to train the target language model.

It is today common practice to use phrases as translation units (Koehn et al., 2003; Och and Ney, 2003) instead of the original word-based approach. A phrase is defined as a group of source words $\tilde{f}$ that should be translated together into a group of target words $\tilde{e}$. The translation model in phrase-based systems includes the phrase translation probabilities in both directions, i.e. $P(\tilde{e}|\tilde{f})$ and $P(\tilde{f}|\tilde{e})$. The use of a maximum entropy approach simplifies the introduction of several additional models explaining the translation process :

$$e^* = \arg\max Pr(e|f)$$

$$= \arg\max_e \{exp(\sum_i \lambda_i h_i(e,f))\} \quad (3)$$

The feature functions h_i are the system models and the λ_i weights are typically optimized to maximize a scoring function on a development set (Och and Ney, 2002). In our system fourteen features functions were used, namely phrase and lexical translation probabilities in both directions, seven features for the lexicalized distortion model, a word and a phrase penalty, and a target language model.

The system is based on the Moses SMT toolkit (Koehn et al., 2007) and constructed as follows. First, Giza++ is used to perform word alignments in both directions. Second, phrases and lexical reorderings are extracted using the default settings of the Moses SMT toolkit. The 4-gram back-off target LM is trained on the English part of the bitexts and the Gigaword corpus of about 3.2 billion words. Therefore, it is likely that the target language model includes at least some of the translations of the French Gigaword corpus. We argue that this is a key factor to obtain good quality translations. The translation model was trained on the news-commentary corpus (1.56M words)[1] and a bilingual dictionary of about 500k entries.[2] This system uses only a limited amount of human-translated parallel texts, in comparison to the bitexts that are available in NIST evaluations. In a different versions of this system, the Europarl (40M words) and the Canadian Hansard corpus (72M words) were added.

In the framework of the EuroMatrix project, a test set of general news data was provided for the shared translation task of the third workshop on

[1] Available at `http://www.statmt.org/wmt08/shared-task.html`

[2] The different conjugations of a verb and the singular and plural form of adjectives and nouns are counted as multiple entries.

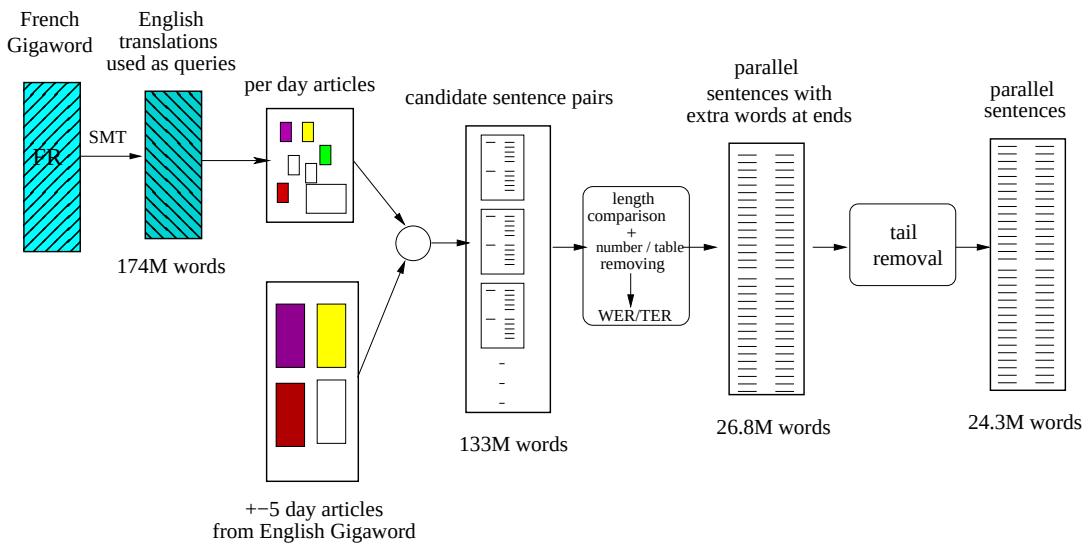

Figure 3: Architecture of the parallel sentence extraction system.

SMT (Callison-Burch et al., 2008), called *newstest2008* in the following. The size of this corpus amounts to 2051 lines and about 44 thousand words. This data was randomly split into two parts for development and testing. Note that only one reference translation is available. We also noticed several spelling errors in the French source texts, mainly missing accents. These were mostly automatically corrected using the Linux spell checker. This increased the BLEU score by about 1 BLEU point in comparison to the results reported in the official evaluation (Callison-Burch et al., 2008). The system tuned on this development data is used translate large amounts of text of French Gigaword corpus (see Figure 2). These translations will be then used to detect potential parallel sentences in the English Gigaword corpus.

3 System Architecture

The general architecture of our parallel sentence extraction system is shown in figure 3. Starting from comparable corpora for the two languages, French and English, we propose to translate French to English using an SMT system as described above. These translated texts are then used to perform information retrieval from the English corpus, followed by simple metrics like WER and TER to filter out good sentence pairs and eventually generate a parallel corpus. We show that a parallel corpus obtained using this technique helps considerably to improve an SMT system.

We shall also be trying to answer the following question over the course of this study: do we need to use the best possible SMT systems to be able to retrieve the correct parallel sentences or any ordinary SMT system will serve the purpose ?

3.1 System for Extracting Parallel Sentences from Comparable Corpora

LDC provides large collections of texts from multilingual news reporting agencies. We identified agencies that provided news feeds for the languages of our interest and chose AFP for our study.[3]

We start by translating the French AFP texts to English using the SMT systems discussed in section 2. In our experiments we considered only the most recent texts (2002-2006, 5.5M sentences; about 217M French words). These translations are then treated as queries for the IR process. The design of our sentence extraction process is based on the heuristic that considering the corpus at hand, we can safely say that a news item reported on day X in the French corpus will be most probably found in the day X-5 and day X+5 time period. We experimented with several window sizes and found the window size of ± 5 days to be the most accurate in terms of time and the quality of the retrieved sentences.

Using the ID and date information for each sentence of both corpora, we first collect all sentences from the SMT translations corresponding to the same day (query sentences) and then the corresponding articles from the English Gigaword cor-

[3]LDC corpora LDC2007T07 (English) and LDC2006T17 (French).

pus (search space for IR). These day-specific files are then used for information retrieval using a robust information retrieval system. The Lemur IR toolkit (Ogilvie and Callan, 2001) was used for sentence extraction. The top 5 scoring sentences are returned by the IR process. We found no evidence that retrieving more than 5 top scoring sentences helped get better sentences. At the end of this step, we have for each query sentence 5 potentially matching sentences as per the IR score.

The information retrieval step is the most time consuming task in the whole system. The time taken depends upon various factors like size of the index to search in, length of the query sentence etc. To give a time estimate, using a ± 5 day window required 9 seconds per query vs 15 seconds per query when a ± 7 day window was used. The number of results retrieved per sentence also had an impact on retrieval time with 20 results taking 19 seconds per query, whereas 5 results taking 9 seconds per query. Query length also affected the speed of the sentence extraction process. But with the problem at we could differentiate among important and unimportant words as nouns, verbs and sometimes even numbers (year, date) could be the keywords. We, however did place a limit of approximately 90 words on the queries and the indexed sentences. This choice was motivated by the fact that the word alignment toolkit Giza++ does not process longer sentences.

A Krovetz stemmer was used while building the index as provided by the toolkit. English stop words, i.e. frequently used words, such as "a" or "the", are normally not indexed because they are so common that they are not useful to query on. The stop word list provided by the IR Group of University of Glasgow[4] was used.

The resources required by our system are minimal : translations of one side of the comparable corpus. We will be showing later in section 4.2 of this paper that with an SMT system trained on small amounts of human-translated data we can 'retrieve' potentially good parallel sentences.

3.2 Candidate Sentence Pair Selection

Once we have the results from information retrieval, we proceed on to decide whether sentences are parallel or not. At this stage we choose the best scoring sentence as determined by the toolkit and pass the sentence pair through further filters. Gale and Church (1993) based their align program on the fact that longer sentences in one language tend to be translated into longer sentences in the other language, and that shorter sentences tend to be translated into shorter sentences. We also use the same logic in our initial selection of the sentence pairs. A sentence pair is selected for further processing if the length ratio is not more than 1.6. A relaxed factor of 1.6 was chosen keeping in consideration the fact that French sentences are longer than their respective English translations. Finally, we discarded all sentences that contain a large fraction of numbers. Typically, those are tables of sport results that do not carry useful information to train an SMT.

Sentences pairs conforming to the previous criteria are then judged based on WER (Levenshtein distance) and translation error rate (TER). WER measures the number of operations required to transform one sentence into the other (insertions, deletions and substitutions). A zero WER would mean the two sentences are identical, subsequently lower WER sentence pairs would be sharing most of the common words. However two correct translations may differ in the order in which the words appear, something that WER is incapable of taking into account as it works on word to word basis. This shortcoming is addressed by TER which allows block movements of words and thus takes into account the reorderings of words and phrases in translation (Snover et al., 2006). We used both WER and TER to choose the most suitable sentence pairs.

4 Experimental evaluation

Our main goal was to be able to create an additional parallel corpus to improve machine translation quality, especially for the domains where we have less or no parallel data available. In this section we report the results of adding these extracted parallel sentences to the already available human-translated parallel sentences.

We conducted a range of experiments by adding our extracted corpus to various combinations of already available human-translated parallel corpora. We experimented with WER and TER as filters to select the best scoring sentences. Generally, sentences selected based on TER filter showed better BLEU and TER scores than their WER counter parts. So we chose TER filter as standard for

[4] http://ir.dcs.gla.ac.uk/resources/linguistic_utils/stop_words

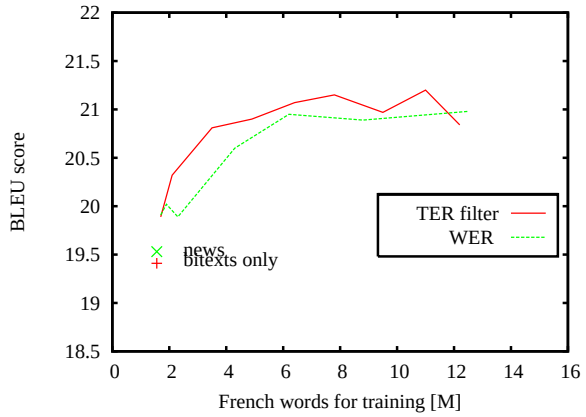

Figure 4: BLEU scores on the Test data using an WER or TER filter.

our experiments with limited amounts of human translated corpus. Figure 4 shows this WER vs TER comparison based on BLEU and TER scores on the test data in function of the size of training data. These experiments were performed with only 1.56M words of human-provided translations (news-commentary corpus).

4.1 Improvement by sentence tail removal

Two main classes of errors common in such tasks: firstly, cases where the two sentences share many common words but actually convey different meaning, and secondly, cases where the two sentences are (exactly) parallel except at sentence ends where one sentence has more information than the other. This second case of errors can be detected using WER as we have both the sentences in English. We detected the extra insertions at the end of the IR result sentence and removed them. Some examples of such sentences along with tails detected and removed are shown in figure 1. This resulted in an improvement in the SMT scores as shown in table 1.

This technique worked perfectly for sentences having TER greater than 30%. Evidently these are the sentences which have longer tails which result in a lower TER score and removing them improves performance significantly. Removing sentence tails evidently improved the scores especially for larger data, for example for the data size of 12.5M we see an improvement of 0.65 and 0.98 BLEU points on dev and test data respectively and 1.00 TER points on test data (last line table 1).

The best BLEU score on the development data is obtained when adding 9.4M words of automatically aligned bitexts (11M in total). This corresponds to an increase of about 2.88 points BLEU on the development set and an increase of 2.46 BLEU points on the test set (19.53 → 21.99) as shown in table 2, first two lines. The TER decreased by 3.07%.

Limit TER filter	Word tail removal	Words (M)	BLEU Dev data	BLEU Test data	TER Test data
0		1.56	19.41	19.53	63.17
10	no	1.58	19.62	19.59	63.11
	yes		19.56	19.51	63.24
20	no	1.7	19.76	19.89	62.49
	yes		**19.81**	19.75	62.80
30	no	2.1	20.29	20.32	62.16
	yes		20.16	20.22	**62.02**
40	no	3.5	20.93	20.81	61.80
	yes		**21.23**	**21.04**	**61.49**
45	no	4.9	20.98	20.90	62.18
	yes		**21.39**	**21.49**	**60.90**
50	no	6.4	21.12	21.07	61.31
	yes		**21.70**	**21.70**	**60.69**
55	no	7.8	21.30	21.15	61.23
	yes		**21.90**	**21.78**	**60.41**
60	no	9.8	21.42	20.97	61.46
	yes		**21.96**	**21.79**	**60.33**
65	no	11	21.34	21.20	61.02
	yes		**22.29**	**21.99**	**60.10**
70	no	12.2	21.21	20.84	61.24
	yes		**21.86**	**21.82**	**60.24**

Table 1: Effect on BLEU score of removing extra sentence tails from otherwise parallel sentences.

Adding the dictionary improves the baseline system (second line in Table 2), but it is not necessary any more once we have the automatically extracted data.

Having had very promising results with our previous experiments, we proceeded onto experimentation with larger human-translated data sets. We added our extracted corpus to the collection of News-commentary (1.56M) and Europarl (40.1M) bitexts. The corresponding SMT experiments yield an improvement of about 0.2 BLEU points on the Dev and Test set respectively (see table 2).

4.2 Effect of SMT quality

Our motivation for this approach was to be able to improve SMT performance by 'creating' parallel texts for domains which do not have enough or any parallel corpora. Therefore only the news-

Bitexts	total words	BLEU score Dev	BLEU score Test	TER Test
News	1.56M	19.41	19.53	63.17
News+Extracted	11M	**22.29**	**21.99**	**60.10**
News+dict	2.4M	20.44	20.18	61.16
News+dict+Extracted	13.9M	**22.40**	**21.98**	**60.11**
News+Eparl+dict	43.3M	22.27	22.35	59.81
News+Eparl+dict+Extracted	51.3M	**22.47**	**22.56**	59.83

Table 2: Summary of BLEU scores for the best systems on the Dev-data with the news-commentary corpus and the bilingual dictionary.

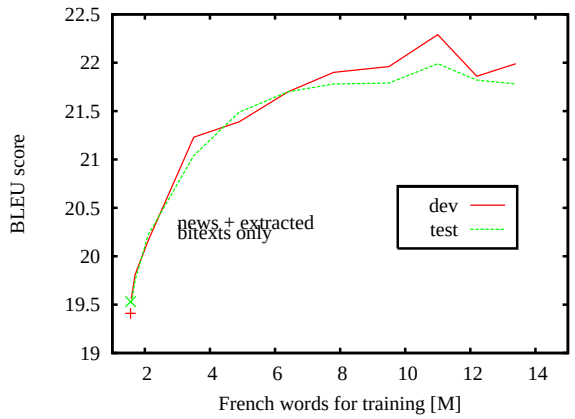

Figure 5: BLEU scores when using news-commentary bitexts and our extracted bitexts filtered using TER.

commentary bitext and the bilingual dictionary were used to train an SMT system that produced the queries for information retrieval. To investigate the impact of the SMT quality on our system, we built another SMT system trained on large amounts of human-translated corpora (116M), as detailed in section 2. Parallel sentence extraction was done using the translations performed by this big SMT system as IR queries. We found no experimental evidence that the improved automatic translations yielded better alignments of the comaprable corpus. It is however interesting to note that we achieve almost the same performance when we add 9.4M words of autoamticallly extracted sentence as with 40M of human-provided (out-of domain) translations (second versus fifth line in Table 2).

5 Conclusion and discussion

Sentence aligned parallel corpora are essential for any SMT system. The amount of in-domain parallel corpus available accounts for the quality of the translations. Not having enough or having no in-domain corpus usually results in bad translations for that domain. This need for parallel corpora, has made the researchers employ new techniques and methods in an attempt to reduce the dire need of this crucial resource of the SMT systems. Our study also contributes in this regard by employing an SMT itself and information retrieval techniques to produce additional parallel corpora from easily available comparable corpora.

We use automatic translations of comparable corpus of one language (source) to find the corresponding parallel sentence from the comparable corpus in the other language (target). We only used a limited amount of human-provided bilingual resources. Starting with about a total 2.6M words of sentence aligned bilingual data and a bilingual dictionary, large amounts of monolingual data are translated. These translations are then employed to find the corresponding matching sentences in the target side corpus, using information retrieval methods. Simple filters are used to determine whether the retrieved sentences are parallel or not. By adding these retrieved parallel sentences to already available human translated parallel corpora we were able to improve the BLEU score on the test set by almost 2.5 points. Almost one point BLEU of this improvement was obtained by removing additional words at the end of the aligned sentences in the target language.

Contrary to the previous approaches as in (Munteanu and Marcu, 2005) which used small amounts of in-domain parallel corpus as an initial resource, our system exploits the target language side of the comparable corpus to attain the same goal, thus the comparable corpus itself helps to better extract possible parallel sentences. The Gigaword comparable corpora were used in this paper, but the same approach can be extended to ex-

tract parallel sentences from huge amounts of corpora available on the web by identifying comparable articles using techniques such as (Yang and Li, 2003) and (Resnik and Y, 2003).

This technique is particularly useful for language pairs for which very little parallel corpora exist. Other probable sources of comparable corpora to be exploited include multilingual encyclopedias like Wikipedia, encyclopedia Encarta etc. There also exist domain specific comparable corpora (which are probably potentially parallel), like the documentations that are done in the national/regional language as well as English, or the translations of many English research papers in French or some other language used for academic proposes.

We are currently working on several extensions of the procedure described in this paper. We will investigate whether the same findings hold for other tasks and language pairs, in particular translating from Arabic to English, and we will try to compare our approach with the work of Munteanu and Marcu (2005). The simple filters that we are currently using seem to be effective, but we will also test other criteria than the WER and TER. Finally, another interesting direction is to iterate the process. The extracted additional bitexts could be used to build an SMT system that is better optimized on the Gigaword corpus, to translate again all the sentence from French to English, to perform IR and the filtering and to extract new, potentially improved, parallel texts. Starting with some million words of bitexts, this process may allow to build at the end an SMT system that achieves the same performance than we obtained using about 40M words of human-translated bitexts (news-commentary + Europarl).

6 Acknowledgments

This work was partially supported by the Higher Education Commission, Pakistan through the HEC Overseas Scholarship 2005 and the French Government under the project INSTAR (ANR JCJC06 143038). Some of the baseline SMT systems used in this work were developed in a cooperation between the University of Le Mans and the company SYSTRAN.

References

P. Brown, S. Della Pietra, Vincent J. Della Pietra, and R. Mercer. 1993. The mathematics of statistical machine translation. *Computational Linguistics*, 19(2):263–311.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *Third Workshop on SMT*, pages 70–106.

Pascale Fung and Percy Cheung. 2004. Mining very-non-parallel corpora: Parallel sentence and lexicon extraction via bootstrapping and em. In Dekang Lin and Dekai Wu, editors, *EMNLP*, pages 57–63, Barcelona, Spain, July. Association for Computational Linguistics.

William A. Gale and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrased-based machine translation. In *HLT/NACL*, pages 127–133.

Philipp Koehn et al. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL, demonstration session*.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504.

Douglas W. Oard. 1997. Alternative approaches for cross-language text retrieval. In *In AAAI Symposium on Cross-Language Text and Speech Retrieval. American Association for Artificial Intelligence*.

Franz Josef Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *ACL*, pages 295–302.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignement models. *Computational Linguistics*, 29(1):19–51.

Paul Ogilvie and Jamie Callan. 2001. Experiments using the Lemur toolkit. In *In Proceedings of the Tenth Text Retrieval Conference (TREC-10)*, pages 103–108.

Philip Resnik and Noah A. Smith Y. 2003. The web as a parallel corpus. *Computational Linguistics*, 29:349–380.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *ACL*.

Masao Utiyama and Hitoshi Isahara. 2003. Reliable measures for aligning Japanese-English news articles and sentences. In Erhard Hinrichs and Dan Roth, editors, *ACL*, pages 72–79.

Christopher C. Yang and Kar Wing Li. 2003. Automatic construction of English/Chinese parallel corpora. *J. Am. Soc. Inf. Sci. Technol.*, 54(8):730–742.