Exploiting Comparable Corpora with TER and TERp

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

Abstract

In this paper we present an extension of a successful simple and effective method for extracting parallel sentences from comparable corpora and we apply it to an Arabic/English NIST system. We experiment with a new TERp filter, along with WER and TER filters. We also report a comparison of our approach with that of (Munteanu and Marcu, 2005) using exactly the same corpora and show performance gain by using much lesser data. Our approach employs an 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 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, a requisite resource for Statistical Machine Translation (SMT) as well as many other natural language processing applications, remain a sparse resource due to the huge expense (human as well as monetary) required for their creation. A parallel corpus, also called bitext, consists in bilingual texts aligned at the sentence level. SMT systems use parallel texts as training material and monolingual corpora for target language modeling. Though enough monolingual data is available for most language pairs, it is the parallel corpus that is a sparse resource.

The performance of an SMT system heavily depends on the parallel corpus used for training. Generally, more bitexts lead to better performance. The existing resources of parallel corpora cover a few language pairs and mostly come from one domain (proceedings of the Canadian or European Parliament, or of the United Nations). The language jargon used in such corpora is not very well suited for everyday life translations or translations of some other domain, thus a dire need arises for more parallel corpora well suited for everyday life and domain adapted translations.

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. Crowd sourcing could be another option, but this has its own costs and thus is not very practical for all cases. The world wide web can also be crawled for potential ”parallel sentences”, but most of the found bilingual texts are not direct translations of each other and not very easy to align. In recent works 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). Potential sources of comparable corpora are multilingual news reporting agencies like AFP, Xinhua, Al-Jazeera, BBC etc, or multilingual encyclopedias like Wikipedia, Encarta etc. Such comparable corpora 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

Proceedings of the 2nd Workshop on Building and Using Comparable Corpora, ACL-IJCNLP 2009, pages 46–54,
Suntec, Singapore, 6 August 2009.

each other. Reliable identification of these pairs would enable the automatic creation of large and diverse parallel corpora.

The ease of availability of these comparable corpora and the potential for parallel corpus as well as dictionary creation has sparked an interest in trying to make maximum use of these comparable resources, some of these works include dictionary learning and identifying word translations (Rapp, 1995), named entity recognition (Sproat et al., 2006), word sense disambiguation (Kaji, 2003), improving SMT performance using extracted parallel sentences (Munteanu and Marcu, 2005), (Rauf and Schwenk, 2009). 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 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 edit rate (TER) to decide whether sentences are parallel or not. We also report an extension of our work (Rauf and Schwenk, 2009) by experimenting with an additional filter TERp, and building a named entity noun dictionary using the unknown words from the SMT (section 5.2). TERp has been tried encouraged by the outperformance of TER in our previous study on French-English. We have applied our technique on a different language pair Arabic-English, versus French-English that we reported the technique earlier on. Our use of full SMT sentences, gives us an added advantage of being able to detect one of the major errors of these approaches, 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 discuss this problem as detailed in section 5.1.

We apply our technique to create a parallel corpus for the Arabic/English language pair. We show that we achieve significant improvements in the BLEU score by adding our extracted corpus to the already available human-translated corpora. We also perform a comparison of the data extracted by our approach and that by (Munteanu and Marcu, 2005) and report the results in Section 5.3.

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 5 summarizes our experimental results and the paper concludes with a discussion and perspectives of this work.

2 Task Description

In this paper, we consider the translation from Arabic into English, under the same conditions as the official NIST 2008 evaluation. The used bi-

texts include various news wire translations[1] as well as some texts from the GALE project.[2] We also added the 2002 to 2005 test data to the parallel training data (using all reference translations). This corresponds to a total of about 8M Arabic words. Our baseline system is trained on these bitexts only.

We use the 2006 NIST test data as development data and the official NIST 2008 test data as internal test set. All case sensitive BLEU scores are calculated with the NIST scoring tool with respect to four reference translations. Both data sets include texts from news wires as well as newsgroups.

LDC provides large collections of monolingual data, namely the LDC Arabic and English Gigaword corpora. There are two text sources that do exist in Arabic and English: the AFP and XIN collection. It is likely that each corpora contains sentences which are translations of the other. We aim to extract those. We have used the XIN corpus for all of our reported results and the collection of the AFP and XIN for comparison with ISI. Table 1 summarizes the characteristics of the corpora used. Note that the English part is much larger than the Arabic one (we found the same to be the case for French-English AFP comparable corpora that we used in our previous study). The number of words are given after tokenization.

Source	Arabic	English
AFP	138M	527M
XIN	51M	140M

Table 1: Characteristics of the available comparable Gigaword corpora for the Arabic-English task (number of words).

3 Baseline SMT system

The goal of statistical machine translation (SMT) is to produce a target sentence $\mathbf{e}$ from a source sentence $\mathbf{f}$. It is today common practice to use phrases as translation units (Koehn et al., 2003; Och and Ney, 2003) and a log linear framework in order to introduce several models explaining the translation process:

$$\mathbf{e}^* = \arg\max p(\mathbf{e}|\mathbf{f}) = \arg\max_e \{exp(\sum_i \lambda_i h_i(\mathbf{e}, \mathbf{f}))\} \quad (1)$$

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 (LM).

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 target 4-gram back-off language model is trained on the English part of all bitexts as well as the whole English Gigaword corpus.

4 System Architecture

The general architecture of our parallel sentence extraction system is shown in figure 1. Starting from comparable corpora for the two languages, Arabic and English, we first translate Arabic to English using an SMT system as described in the above sections. These translated texts are then used to perform information retrieval from the English corpus, followed by simple metrics like WER, TER or TERp 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.

4.1 System for Extracting Parallel Sentences from Comparable Corpora

We start by translating the Arabic XIN and AFP texts to English using the SMT systems discussed in section 2. In our experiments we considered only the most recent texts (2001-2006, 1.7M sentences; about 65.M Arabic words for XIN). For our experiments on effect on SMT quality we use only the XIN corpus. We use the combination of AFP and XIN for comparison of sentences extracted by our approach with that of (Munteanu and Marcu, 2005). 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

[1] LDC2003T07, 2004E72, T17, T18, 2005E46 and 2006E25.

[2] LDC2005E83, 2006E24, E34, E85 and E92.

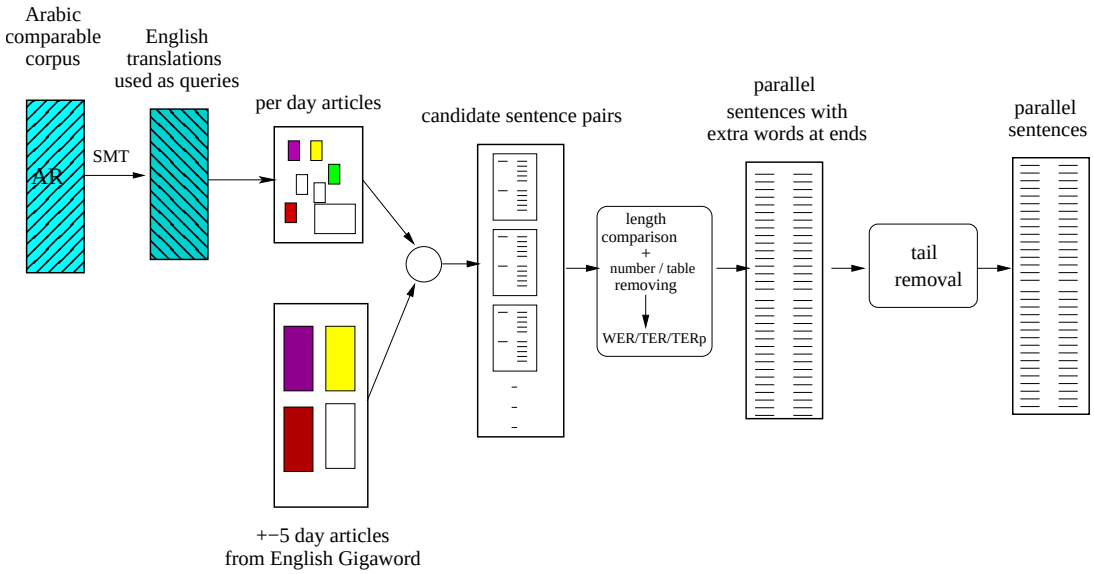

Figure 1: Architecture of the parallel sentence extraction system.

can safely say that a news item reported on day X in the Arabic 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 is to be the most accurate in terms of time and the quality of the retrieved sentences. (Munteanu and Marcu, 2005) have also worked with a $\pm$5 day window.

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 corpus (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 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 $\pm$5 day window required 9 seconds per query vs 15 seconds per query when a $\pm$7 day window was used. We placed 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[3] was used.

The resources required by our system are minimal : translations of one side of the comparable corpus. It has already been demonstrated in (Rauf and Schwenk, 2009) that when using translations as queries, the quality of the initial SMT is not a factor for better sentence retrieval and that an SMT system trained on small amounts of human-translated data can ’retrieve’ potentially good parallel sentences.

4.2 Candidate Sentence Pair Selection

The information retrieval process gives us the potential parallel sentences per query sentence, the decision of their being parallel or not needs to be made about them. 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 initially used the same logic in our selection of the candidate sentence pairs. However our observation was that the filters that we use, WER, TER and TERp implicitly place a penalty when the length differ-

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

ence between two sentences is too large. Thus using this inherent property, we did not apply any explicit sentence length filtering.

The candidate sentences pairs are then judged based on simple filters. Our choice of filters in accordance to the task in consideration were the WER (Levenshtein distance), Translation Edit Rate (TER) and the relatively new Translation Edit Rate plus (TERp). 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. 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). TERp is an extension of Translation Edit Rate and was one of the top performing metrics at the NIST Metric MATR workshop [4]. It had the highest absolute correlation, as measured by the Pearson correlation coefficient, with human judgments in 9 of the 45 test conditions. TERp tries to address the weaknesses of TER through the use of paraphrases, morphological stemming, and synonyms, as well as edit costs that are optimized to correlate better with various types of human judgments (Snover et al., 2009). The TER filter allows shifts if the two strings (the word sequence in the translated and the IR retrieved sentence) match exactly, however TERp allows shifts if the words being shifted are exactly the same, are synonyms, stems or paraphrases of each other, or any such combination. This allows better sentence comparison by incorporation of sort of linguistic information about words.

5 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.

[4] http://www.itl.nist.gov/iad/mig/tests/metricsmatr/2008/

Bitexts	#words Arabic	BLEU Eval06	BLEU Eval08
Baseline	5.8M	42.64	39.35
+WER-10	5.8M	42.73	39.70
+WER-40	7.2M	43.34	40.59
+WER-60	14.5M	43.95	41.20
+WER-70	20.4M	43.58	41.18
+TER-30	6.5M	43.41	40.08
+TER-50	12.5M	43.90	41.45
+TER-60	17.3M	**44.30**	**41.73**
+TER-75	24.1M	43.79	41.21
+TERp-10	5.8M	42.69	39.80
+TERp-40	10.2M	43.89	41.44
+TERp-60	20.8M	43.94	41.25
+TERp-80	27.7M	43.90	41.58

Table 2: Summary of BLEU scores for the best systems selected based on various thresholds of WER, TER and TERp filters

We conducted a range of experiments by adding our extracted corpus to various combinations of already available human-translated parallel corpora. For our experiments on effect on SMT quality we use only the XIN extracted corpus. We experimented with WER, TER and TERp as filters to select the best scoring sentences. Table 2 shows some of the scores obtained based on BLEU scores on the Dev and test data as a function of the size of the added extracted corpus. The name of the bitext indicates the filter threshold used, for example, TER-50 means sentences selected based on TER filter threshold of 50. Generally, sentences selected based on TER filter showed better BLEU scores on NIST06 than their WER and TERp counter parts up to almost 21M words. Also for the same filter threshold TERp selected longer sentences, followed by TER and then WER, this fact is evident from table 2, where for the filter threshold of 60, TERp and TER select 20.8M and 17.3 words respectively, whereas WER selects 14.5M words.

Figure 2 shows the trend obtained in function of the number of words added. These experiments were performed by adding our extracted sentences to only 5.8M words of human-provided translations. Our best results are obtained when 11.5M of our extracted parallel sentences based on TER filter are added to 5.8M of News wire and gale parallel corpora. We gain an improvement of 1.66 BLEU points on NIST06 and 2.38 BLEU points

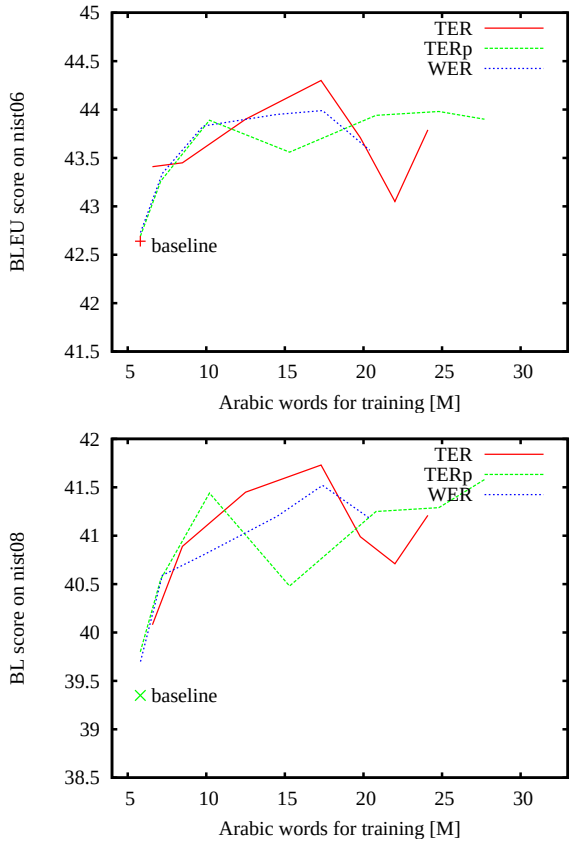

Figure 2: BLEU scores on the NIST06 (Dev, top) and NIST08 (test, bottom) data using an WER,TER or TERp filter as a function of the number of extracted Arabic words added.

on NIST08 (TER-60 in table 2).

An interesting thing to notice in figure 2 is that no filter was able to clearly outperform the others, which is contradictory to our experiments with the French-English language pair (Rauf and Schwenk, 2009), where the TER filter clearly outperformed the WER filter. WER is worse than TER but less evident here than for our previous experiments for the French-English language pair. This performance gain by using the TER filter for French-English was our main motivation for trying TERp. We expected TERp to get better results compared to WER and TER, but TER filter seems the better one among the three filters. Note that all conditions in all the experiments were identical. This gives a strong hint of language pair dependency, making the decision of suitability of a particular filter dependent on the language pair in consideration.

5.1 Sentence tail removal

Two main classes of errors are known when extracting parallel sentences from comparable corpora: 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 the advantage of having 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 3. Since this gives significant improvement in the SMT scores we used it for all our extracted sentences (Rauf and Schwenk, 2009). However, similar to our observations in the last section, the tails were much shorter as compared to our previous experiments with French-English, also most of the tails in this Arabic-English data were of type as shown in last line figure 3. This is a factor dependent on reporting agency and its scheme for reporting, i.e, whether it reports an event independently in each language or uses the translation from one language to the other .

5.2 Dictionary Creation

In our translations, we keep the unknown words as they are, i.e. in Arabic (normally a flag is used so that Moses skips them). This enables us to build a dictionary. Consider the case with translation with one unknown word in Arabic, if all the other words around align well with the English sentence that we found with IR, we could conclude the translation of the unknown Arabic word, see figure 3 line 5. We were able to make a dictionary using this scheme which was comprised mostly of proper nouns often not found in Arabic-English dictionaries. Our proper noun dictionary comprises of about 244K words, some sample words are shown in figure 4. Adding the proper nouns found by this technique to the initial SMT system should help improve translations for new sentences, as these words were before unknown to the system. However, the impact of addition of these words on translation quality is to be evaluated at the moment.

Arabic: بدا الاف الموظفين فى فرز الاصوات التى تم تسجيلها فى عشرات الاف الماكينات الالك ترونية فى 855 بلدة ومدينة عبر البلاد فى الساعة الثامنة صباحا .
Query: *Thousands of officials began counting the votes registered in tens of thousands of electronic machines in 855 towns and cities across the country at 8 a.m.*
Result: *Thousands of officials began counting the votes registered in tens of thousands of electronic machines in 855 towns and cities across the country at 8 a.m.* ***thursday***.
Arabic: كان ويكرمسنغ يشير بذلك الى الجمود الحالى بين حكومته ومتمردى جبهة نمور تحرير ايلام التاميلية .
Query: ويكرمسنغ *was referring to the current stalemate between his government and the Liberation Tigers of Tamil Eelam .*
Result: *Wickremesinghe was referring to the current stalemate between his government and the Liberation Tigers of Tamil Eelam* ***(LTTE) REBELS .***
Arabic: اتخذ بونو هذا الموقف بعد ان طالب بعض المشرعين الحكومة باعادة التفكير فى التواجد العسكرى الاسبانى فى افغانستان .
Query: *Bono adopted this position after some legislators asked the government to rethink the Spanish military presence in Afghanistan .*
Result: *Bono adopted this attitude after some legislators asked the government to reconsider the Spanish military presence in Afghanistan .* ***(SPAIN-AFGHANISTAN) .***

Figure 3: Some examples of an Arabic 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.

English word from SMT	Arabic unknown word
PetroChina	بتروتشاينا
Bolotine	بولوتين
Amrozi	امروزى
Bulldozers	البولدوزورات
Schulte	شولتى
Jiuxuan	جيوشيوان
Dijmarescu	ديجماريسكو
Aliasghar Soltanieh	اليسجار سلطانيه

Figure 4: Examples of some words found by our dictionary building technique.

5.3 Comparison with previous work

LDC provides extracted parallel texts extracted with the algorithm published by (Munteanu and Marcu, 2005). This corpus contains 1.1M sentence pairs (about 35M words) which were automatically extracted and aligned from the monolingual Arabic and English Gigaword corpora, a confidence score being provided for each sentence pair. We also applied our approach on data provided by LDC, but on a different subset. Since we had used the recent data sets our corpora were till year 2006, whereas ISI's data were till year 2004. We filtered our data according to the time interval of their data (date information was provided for each sentence pair) and used them to compare the two data sets. Both AFP and XIN were used in these comparison experiments since the available ISI's data was comprised of these two collections.

To perform the comparison, we have, firstly, the ISI parallel sentences and secondly the parallel sentences extracted by using our approach using the same time frame and comparable corpora as ISI. We used our sentences as filtered by the TER filter and added them to the already available 5.8M of human-translated (as done in previous experiments). The result is shown graphically in figure 5. Adding the ISI parallel data to the 5.8M baseline parallel corpus (total 27.5M words) yielded a BLEU score of 43.59 on NIST06 Dev set and 41.84 BLEU points on NIST08 test set. Whereas we were able to achieve a BLEU score of 43.88 on NIST06 Dev and 41.35 on NIST08 test set (using a total of 16.1M words), which amounts to an increase of 0.29 BLEU points on the NIST06 Dev set. Note that this gain is achieved by using a total of only 10.3M of our extracted words as compared to 21.7M of ISI corpus to get their best result. However we were not able to improve as much on the NIST08 test corpus.

The trend in BLEU score in figure 5 clearly

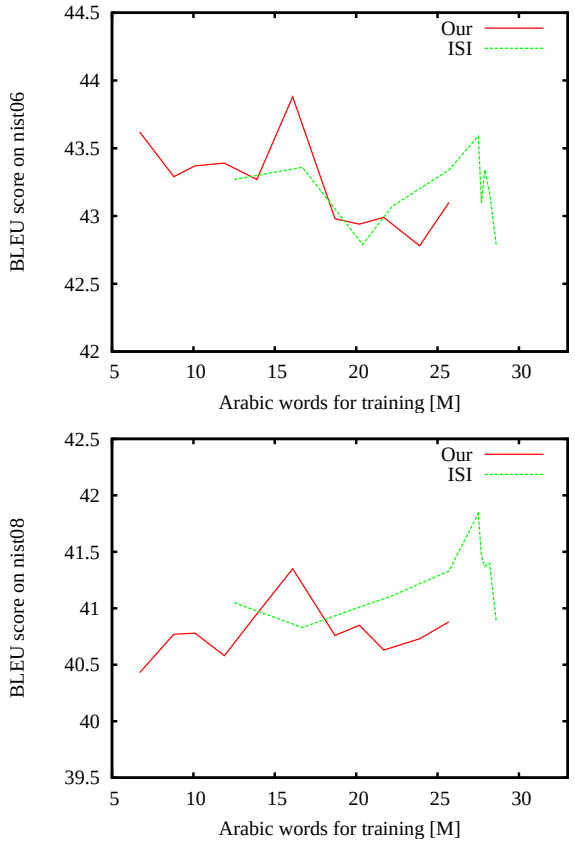

Figure 5: BLEU scores on the NIST06 and NIST08 data using the ISI parallel corpus and our comparative extracted bitexts in function of number of extracted Arabic words added.

shows that our sentence selection scheme selects good sentences, and is capable of achieving the same scores but with much less sentences. This is because in the scheme of ISI, the confidence scores provided are based on the IR and maximum entropy classifier scoring scheme, whereas our filters score the sentences based on linguistic sentence similarity, allowing us to retrieve the good sentence pairs from the bad ones. Once information retrieval is done, which is the most time consuming task in both the techniques, our approach is better able to sort out the good IR extracted sentences as is evident from the results obtained. Moreover our scheme does not require any complex operations, just simple filters which are well adapted to the problem at hand.

6 Conclusion and discussion

Sentence-aligned bilingual texts are a crucial resource to build SMT systems. For some language pairs bilingual corpora just do not exist, the existing corpora are too small to build a good SMT system or they are not of the same genre or 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 translations of the source language comparable corpus to find the corresponding parallel sentences from the target language comparable corpus. We only used a limited amount of human-provided bilingual resources. Starting with small amounts of sentence aligned bilingual data 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(NIST08) by 2.38 points for the Arabic-English language pair.

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. We have also presented a comparison with their approach and found our bitexts to achieve nice improvements using much less words. The LDC comparable corpora were used in this paper, but the same approach can be extended to extract 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).We have successfully applied our approach to French-English and Arabic-English language pairs. As this study strongly hinted towards language pair dependancy on the choice of the filter to use to select better sentences, we intend to investigate this trend in detail.

References

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.

Hiroyuki Kaji. 2003. Word sense acquisition from bilingual comparable corpora. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 32–39, Morristown, NJ, USA. Association for Computational Linguistics.

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.

Reinhard Rapp. 1995. Identifying word translations in non-parallel texts. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pages 320–322, Morristown, NJ, USA. Association for Computational Linguistics.

Sadaf Abdul Rauf and Holger Schwenk. 2009. On the use of comparable corpora to improve SMT performance. In *EACL*, pages 16–23.

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*.

Matthew Snover, Bonnie Dorr, and Richard Schwartz. 2008. Language and translation model adaptation using comparable corpora. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 857–866, Honolulu, Hawaii, October. Association for Computational Linguistics.

Matthew Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. 2009. Fluency, adequacy, or HTER? Exploring different human judgments with a tunable MT metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 259–268, Athens, Greece, March. Association for Computational Linguistics.

Richard Sproat, Tao Tao, and ChengXiang Zhai. 2006. Named entity transliteration with comparable corpora. In *ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 73–80, Morristown, NJ, USA. Association for Computational Linguistics.

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.