Multimodal Comparable Corpora as Resources for Extracting Parallel Data: Parallel Phrases Extraction

Haithem Afli, Loïc Barrault and Holger Schwenk
Université du Maine,
Avenue Olivier Messiaen F-72085 - LE MANS, France
FirstName.LastName@lium.univ-lemans.fr

Abstract

Discovering parallel data in comparable corpora is a promising approach for overcoming the lack of parallel texts in statistical machine translation and other NLP applications. In this paper we propose an alternative to comparable corpora of texts as resources for extracting parallel data: a multimodal comparable corpus of audio and texts. We present a novel method to detect parallel phrases from such corpora based on splitting comparable sentences into fragments, called phrases. The audio is transcribed by an automatic speech recognition system, split into fragments and translated with a baseline statistical machine translation system. We then use information retrieval in a large text corpus in the target language, split also into fragments, and extract parallel phrases. We compared our method with parallel sentences extraction techniques. We evaluate the quality of the extracted data on an English to French translation task and show significant improvements over a state-of-the-art baseline.

1 Introduction

The development of a statistical machine translation (SMT) system requires one or more parallel corpora called bitexts for training the translation model and monolingual data to build the target language model. Unfortunately, parallel texts are a limited resource and they are often not available for some specific domains and language pairs. That is why, recently, there has been a huge interest in the automatic creation of parallel data. Since comparable corpora exist in large quantities and are much more easily available (Munteanu and Marcu, 2005), the ability to exploit them is highly beneficial in order to overcome the lack of parallel data. The ability to detect these parallel data enables the automatic creation of large parallel corpora.

Most of existing studies dealing with comparable corpora look for parallel data at the sentence level (Zhao and Vogel, 2002; Utiyama and Isahara, 2003; Munteanu and Marcu, 2005; Abdul-Rauf and Schwenk, 2011). However, the degree of parallelism can vary considerably, from noisy parallel texts, to quasi parallel texts (Fung and Cheung, 2004). Corpora from the last category contain none or few good parallel sentence pairs. However, there could have parallel phrases in comparable sentences that can prove to be helpful for SMT (Munteanu and Marcu, 2006). As an example, consider Figure 1, which presents two news articles with their video from the English and French editions of the Euronews website[1]. The articles report on the same event with different sentences that contain some parallel translations at the phrase level. These two documents contain in particular no exact sentence pairs, so techniques for extracting parallel sentences will not give good results. We need a method to extract parallel phrases which exist at the sub-sentential level.

For some languages, text comparable corpora may not cover all topics in some specific domains and languages. This is because potential sources of comparable corpora are mainly derived from multilingual news reporting agencies like AFP, Xinhua, Al-Jazeera, BBC etc, or multilingual encyclopedias like Wikipedia, Encarta etc. What we need is exploring other sources like audio to generate parallel data for such domains that can improve the performance of an SMT system.

In this paper, we present a method for detecting and extracting parallel data from multimodal corpora. Our method consists in extracting parallel

[1] www.euronews.com/

International Joint Conference on Natural Language Processing, pages 286–292,
Nagoya, Japan, 14-18 October 2013.

Figure 1: Example of multimodal comparable corpora from the Euronews website.

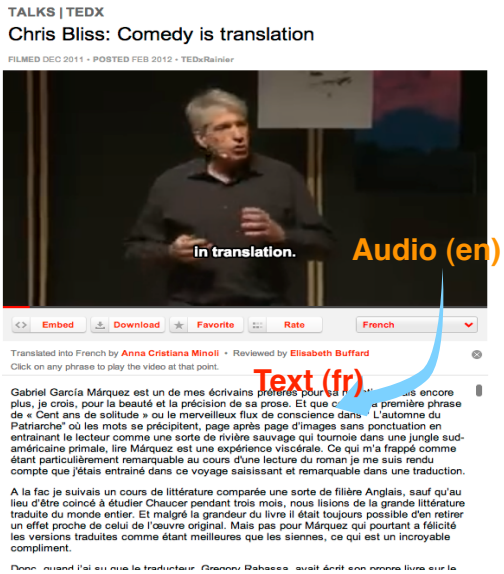

Figure 2: Example of multimodal comparable corpora from the TED website.

phrases.

2 Extracting parallel data

2.1 Basic Idea

Figure 2 shows an example of multimodal comparable data coming from the TED website [2]. We have an audio source of a talk in English and its text translation in French. We think that we can extract parallel data from this corpora, at the sentence and the sub-sentential level.

In this work we seek to adapt and to improve machine translation systems that suffer from resource deficiency by automatically extracting parallel data in specific domains.

[2] http://www.ted.com/

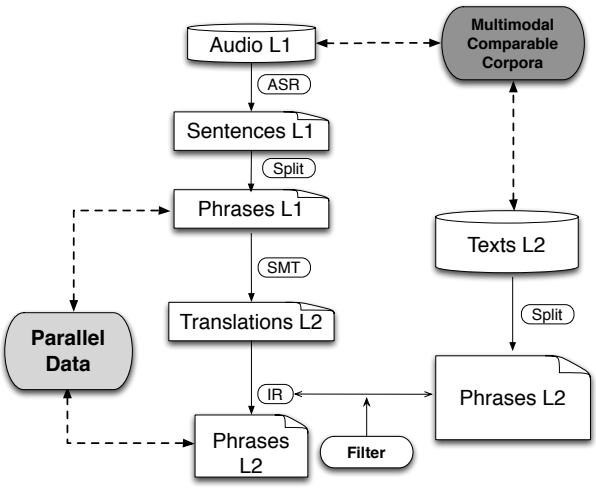

Figure 3: Principle of the parallel phrase extraction system from multimodal comparable corpora.

2.2 System Architecture

The basic system architecture is described in Figure 3. We can distinguish three steps: automatic speech recognition (ASR), statistical machine translation (SMT) and information retrieval (IR). The ASR system accepts audio data in the source language L1 and generates an automatic transcription. This transcription is then split into phrases and translated by a baseline SMT system into language L2. Then, we use these translations as queries for an IR system to retrieve most similar phrases in the texts in L2, which were previouslt split into phrases. The transcribed phrases in L1 and the IR result in L2 form the final parallel data. We hope that the errors made by the ASR and SMT systems will not impact too severely the extraction process.

Our technique is similar to that of (Munteanu and Marcu, 2006), but we bypass the need of the Log-Likelihood-Ratio lexicon by using a baseline SMT system and the TER measure (Snover et al., 2006) for filtering. We also report an extension of the work of (Afli et al., 2012) by splitting transcribed sentences and the text parts of the multimodal corpus into phrases with length between two to ten tokens. We extract from each sentence on the corpus all combinations of two to ten sequential words.

2.3 Baseline systems

Our ASR system is a five-pass system based on the open-source CMU Sphinx toolkit [3](version 3 and 4), similar to the LIUM'08 French ASR system described in (Deléglise et al., 2009). The acoustic models are trained in the same manner, except that a multi-layer perceptron (MLP) is added using the bottle-neck feature extraction as described in (Grézl and Fousek, 2008). Table 2.3 shows the performances of the ASR system on the development and test corpora.

Corpus	% WER
Development	19.2
Test	17.4

Table 1: Performance of the ASR system on development and test data.

Our SMT system is a phrase-based system (Koehn et al., 2003) based on the Moses SMT toolkit (Koehn et al., 2007). The standard fourteen feature functions are 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. It is constructed as follows. First, word alignments in both directions are calculated. We used the multi-threaded version of the GIZA++ tool (Gao and Vogel, 2008). Phrases and lexical reorderings are extracted using the default settings of the Moses toolkit. The parameters of our system were tuned on a development corpus, using the MERT tool (Och, 2003).

We use the Lemur IR toolkit (Ogilvie and Callan, 2001) for the phrases extraction procedure. We first index all the French text (after splitting it into segments) into a database using *Indri Index*. This feature enable us to index our text documents in such a way we can use the translated phrases as queries to run information retrieval in the database, with the specialized *Indri Query Language*. By these means we can retrieve the best matching phrases from the French side of the comparable corpus.

For each candidate phrases pair, we need to decide whether the two phrases are mutual translations. For this, we calculate the TER between them using the tool described in (Servan and Schwenk, 2011),[4] i.e. between automatic translation, and the phrases selected by IR.

3 Experiments

In our experiments, we compare our phrase extraction method (which we call *PhrExtract*) with the sentence extraction method (*SentExtract*) of (Afli et al., 2012). We use the extracted dataset by both methods as additional SMT training data, and measure the quality of the parallel data by its impact on the performance of the SMT system. Thus, the final extracated parallel data is injected into the baseline system. The various SMT systems are evaluated using the BLEU score (Papineni et al., 2002). We conducted experiments on an English to French machine translation task. All the text data is automatically split into phrases of two to ten tokens.

3.1 Data description

Our multimodal comparable corpus consists of spoken talks in English (audio) and written texts in French. The goal of the TED task is to translate public lectures from English into French. The TED corpus totals about 118 hours of speech. We call the English transcriptions of the audio part *TEDasr* witch is split into phrases (called TEDasr_split). A detailed description of the TED task can be found in (Rousseau et al., 2011).

The development corpus *DevTED* consists of 19 talks and represents a total of 4 hours and 13 minutes of speech transcribed at the sentence level. The language model is trained with the SRI LM toolkit (Stolcke, 2002), on all the available French data without the TED data. The baseline system is trained with version 7 of the News-Commentary (nc7) and Europarl (eparl7) corpus.[5] The indexed data consist of the French text part of the *TED* corpus which contains translations of the English part of the corpus. We call it *TEDbi*. It is split into phrases (called TEDbi_split). Tables 2 and 3 summarize the characteristics of the different corpora used in our experiments.

3.2 Experimental results

We first apply sentence extraction on the TED corpus with a method similar to (Afli et al., 2012). We then apply phrase extraction on the same data split

[3]Carnegie Mellon University: http://cmusphinx.sourceforge.net/

[4]http://sourceforge.net/projects/tercpp/

[5]http://www.statmt.org/europarl/

bitexts	# tokens	in-domain ?
nc7	3.7M	no
eparl7	56.4M	no
DevTED	36k	yes

Table 2: MT training and development data.

Data	# tokens	in-domain ?
TEDasr	1.8M	yes
TEDbi	1.9M	yes
TEDbi_split	80.4M	yes
TEDasr_split	82.7M	yes

Table 3: Comparable data used for the extraction experiments.

as described in 2.2. Then, both methods are compared.

As mentioned in section 2.3, the TER score is used as a metric for filtering the result of IR. We keep only the sentences or phrases which have a TER score below a certain threshold determined empirically. Thus, we filter the selected sentences or phrases in each condition with different TER thresholds ranging from 0 to 100 by steps of 10. The extracted parallel data are added to our generic training data in order to adapt the baseline system. Table 4 presents the BLEU score obtained for these different experimental conditions.

Our baseline SMT system, trained with generic bitexts achieves a BLEU score of 22.93. We can see that our new method of phrase extraction significantly improve the baseline system more than sentences extraction method until the TER threshold of 80 is reached: the BLEU score increases from 22.93 to 23.70 with the best system of our proposed method and from 22.93 to 23.40 with the best system using the classical method of sentence extraction.

The results show that the choice of the appropriate TER threshold depends on the method. We can see that for *PhrExtract* the best threshold is 60 when the best one is 80 for *SentExtract*. This last one is also an important point in the general evaluation of the two methods. In fact, we can see on Figure 4 that from this point our proposed method gives less performing results than *SentExtract* method.

This suggest to apply combination of the two methods. This corresponds to injecting the extracted phrases and sentences into the training data. The combination method is called *CombExtract*. Figure 4 presents the comparison of the different experimental conditions in term of BLEU score for each TER threshold. We can see that except for threshold 30, the curve of the combination follows in general the same trajectory of the curve of *PhrExtract*. These results show that *SentExtract* has no big impact in combination with the *PhrExtract* method and the best threshold when using *PhrExtract* is at 60.

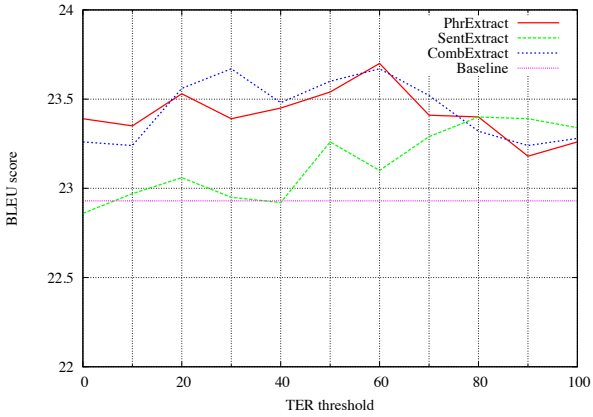

Figure 4: Performance of *PhrExtract*, *SentExtract* and their combination in term of BLEU score for each TER threshold.

This is because of the big difference on the quantity of data between the two methods as we can see in Table 4. The benefit of our method is that it can generates more quantities of parallel data than the sentence extraction method for each TER threshold, and this difference of quantities improves results of MT system until the TER threshold of 80 is reached. However, we can see in Table 4 that the quality of only 39.35k (TER 80) extracted by *SentExtract* can have exactly the same impact of 25.3M extracted by our new technique. That is why we intend to investigate in the filtering module of our system.

4 Related Work

Research on exploiting comparable corpora goes back to more than 15 years ago (Fung and Yee, 1998; Koehn and Knight, 2000; Vogel, 2003; Gaussier et al., 2004; Li and Gaussier, 2010). A lot of studies on data acquisition from comparable corpora for machine translation have been reported (Su and Babych, 2012; Hewavitharana and Vogel, 2011; Riesa and Marcu, 2012).

To the best of our knowledge (Munteanu and

TER	BLEU score SentExtract	BLEU score PhrExtract	# tokens (fr) SentExtract	# tokens (fr) PhrExtract
0	22.86	23.39	55	1.06M
10	22.97	23.35	313	1.4M
20	23.06	23.53	1.7k	2.5M
30	22.95	23.39	6.9k	4.3M
40	22.92	23.45	23.5k	7.02M
50	23.26	23.54	62.4k	11.4M
60	23.10	**23.70**	13.82k	**13.8M**
70	23.29	23.41	25.15k	18.04M
80	**23.40**	**23.40**	**39.35k**	**25.3M**
90	23.39	23.18	57.54k	35.9M
100	23.34	23.26	83.60k	45.3M
Baseline	22.93	-	60.1M	-

Table 4: Number of tokens extracted and BLEU scores on DevTED obtained with *PhrExtract* and *SentExtract* methods for each TER threshold.

Marcu, 2006) was the first attempt to extract parallel sub-sentential fragments (phrases), from comparable corpora. They used a method based on a Log-Likelihood-Ratio lexicon and a smoothing filter. They showed the effectiveness of their method to improve an SMT system from a collection of a comparable sentences. The weakness of their method is that they filter source and target fragments separately, which cannot guarantee that the extracted fragments are a good translations of each other. (Hewavitharana and Vogel, 2011) show a good result with their method based on on a pairwise correlation calculation which suppose that the source fragment has been detected.

The second type of approach in extracting parallel phrases is the alignment-based approach (Quirk et al., 2007; Riesa and Marcu, 2012). These methods are promising, but since the proposed method in (Quirk et al., 2007) do not improve significantly MT performance and model in (Riesa and Marcu, 2012) is designed for parallel data, it's hard to say that this approach is actually effective for comparable data.

This work is similar to the work by (Afli et al., 2012) where the extraction is done at the phrase level instead of the sentence level. Our methodology is the first effort aimed at detecting translated phrases on a multimodal corpora.

Since our method can extract parallel phrases from a multimodal corpus, it greatly expands the range of corpora which can be usefully exploited.

5 Conclusion

We have presented a fully automatic method for extracting parallel phrases from multimodal comparable corpora, *i.e.* the source side is available as audio stream and the target side as text. We used a framework to extract parallel data witch combine an automatic speech recognition system, a statistical machine translation system and information retrieval system. We showed by experiments conducted on English-French data, that parallel phrases extracted with this method improves significantly SMT performance. Our approach can be improved in several aspects. The automatic splitting is very simple; more advanced phrases generation might work better, and eliminate redundancy. Trying other method on filtering can also improve the precision of the method.

6 Acknowledgments

This work has been partially funded by the French Government under the project DEPART.

References

S. Abdul-Rauf and H. Schwenk. 2011. Parallel sentence generation from comparable corpora for improved smt. *Machine Translation*.

H. Afli, L. Barrault, and H. Schwenk. 2012. Parallel texts extraction from multimodal comparable corpora. In *JapTAL*, volume 7614 of *Lecture Notes in Computer Science*, pages 40–51. Springer.

P. Deléglise, Y. Estève, S. Meignier, and T. Merlin. 2009. Improvements to the LIUM french ASR sys-

tem based on CMU Sphinx: what helps to significantly reduce the word error rate? In *Interspeech 2009*, Brighton (United Kingdom), 6-10 september.

P. Fung and P. Cheung. 2004. Multi-level bootstrapping for extracting parallel sentences from a quasi-comparable corpus. In *Proceedings of the 20th international conference on Computational Linguistics*, COLING '04.

P. Fung and L. Y. Yee. 1998. An ir approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 17th international conference on Computational linguistics - Volume 1*, COLING '98, pages 414–420.

Q. Gao and S. Vogel. 2008. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, SETQA-NLP '08, pages 49–57.

E. Gaussier, J.-M. Renders, I. Matveeva, C. Goutte, and H. Déjean. 2004. A geometric view on bilingual lexicon extraction from comparable corpora. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04.

F. Grézl and P. Fousek. 2008. Optimizing bottle-neck features for LVCSR. In *2008 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 4729–4732. IEEE Signal Processing Society.

S. Hewavitharana and S. Vogel. 2011. Extracting parallel phrases from comparable data. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, BUCC '11, pages 61–68.

P. Koehn and K. Knight. 2000. Estimating word translation probabilities from unrelated monolingual corpora using the em algorithm. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, pages 711–715. AAAI Press.

P. Koehn, Franz J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 48–54.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180.

B. Li and E. Gaussier. 2010. Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 644–652.

D. S. Munteanu and D. Marcu. 2005. Improving Machine Translation Performance by Exploiting Non-Parallel Corpora. *Computational Linguistics*, 31(4):477–504.

D. S. Munteanu and D. Marcu. 2006. Extracting parallel sub-sentential fragments from non-parallel corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, pages 81–88.

Franz J. Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 160–167, Stroudsburg, PA, USA. Association for Computational Linguistics.

P. Ogilvie and J. Callan. 2001. Experiments using the lemur toolkit. *Procedding of the Trenth Text Retrieval Conference (TREC-10)*.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318.

Q. Quirk, R. Udupa, and A. Menezes. 2007. Generative models of noisy translations with applications to parallel fragment extraction. In *In Proceedings of MT Summit XI, European Association for Machine Translation*.

J. Riesa and D. Marcu. 2012. Automatic parallel fragment extraction from noisy data. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 538–542.

A. Rousseau, F. Bougares, P. Deléglise, H. Schwenk, and Y. Estève. 2011. LIUM's systems for the IWSLT 2011 speech translation tasks. *International Workshop on Spoken Language Translation 2011*.

C. Servan and H. Schwenk. 2011. Optimising multiple metrics with mert. *The Prague Bulletin of Mathematical Linguistics (PBML)*.

S. Snover, B. Dorr, R. Schwartz, M. Micciulla, and J. Makhoul. 2006. A study of translation edit rate with targeted human annotation. *Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

A. Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *International Conference on Spoken Language Processing*, pages 257–286, November.

F. Su and B. Babych. 2012. Measuring comparability of documents in non-parallel corpora for efficient extraction of (semi-)parallel translation equivalents. In *Proceedings of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra)*, EACL 2012, pages 10–19. Association for Computational Linguistics.

M. Utiyama and H. Isahara. 2003. Reliable measures for aligning japanese-english news articles and sentences. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 72–79.

S. Vogel. 2003. Using noisy bilingual data for statistical machine translation. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics - Volume 2*, EACL '03, pages 175–178.

B. Zhao and S. Vogel. 2002. Adaptive parallel sentences mining from web bilingual news collection. In *Proceedings of the 2002 IEEE International Conference on Data Mining*, ICDM '02, Washington, DC, USA. IEEE Computer Society.