Multimodal Comparable Corpora for Machine Translation

Haithem Afli, Loïc Barrault and Holger Schwenk
Laboratoire Informatique de l’Université du Maine (LIUM)
University of Le Mans, France
firstname.lastname@lium.univ-lemans.fr

Abstract
The construction of a statistical machine translation (SMT) system requires parallel corpus for training the translation model and monolingual data to build the target language model. A parallel corpus, also called bitext, consists in bilingual/multilingual texts. Unfortunately, parallel texts are a sparse resource for many language pairs. One way to overcome this lack of data is to exploit comparable corpora which are much more easily available. In this paper, we present the corpus developed for automatic parallel data extraction from multimodal comparable corpora, from Euronews and TED web sites. We describe the content of each corpus and how we extracted the parallel data with our new extraction system. We present the methods considered for using multimodal corpora and discuss the results on bitext extraction.

Keywords: Multimodal Comparable Corpora, Machine Translation, Parallel Data Extraction.

1. Introduction
In recent decades statistical approaches have significantly advanced the development of machine translation (MT). However, the applicability of these methods directly depends on the availability of very large quantities of parallel data. Recent works have demonstrated that a comparable corpus can compensate for the shortage of parallel corpora. However, for some languages, text comparable corpora may not cover all topics in some specific domains. One of the main challenges of our research is to build data and techniques to these under-resourced domains. What we need is to explore other sources like audio to generate parallel texts for each domain.

These kind of data are widely available on the Web for many languages.

In this paper, we present an extraction method used on multimodal comparable corpus. This corpus is then used to adapt and improve machine translation systems that suffer from resource deficiency. We, also, present Euronews-LIUM corpus which has been created within the context of our work on French DEPART project. One of its main objective is the exploitation of multimodal and multilingual data for machine translation.

The methods for improving translation quality proposed in this work rely upon multimodal comparable corpora, that is, multiple corpora in different modalities that cover the same general topics and events. We compare it with (Afli et al., 2013) method built for the same kind of data.

Our main experimental framework is designed to address two situations. The first one is when we translate data from a new domain, different from the training data. In such a condition, the translation quality is generally rather poor. The second one is when we seek to improve the quality of an SMT system already trained on the same kind of data (same domain and/or style). Data is extracted from the available news (video and text modalities) on the Euronews website2. We also used TED-LIUM (Rousseau et al., 2012) corpus to build our TED multimodal comparable corpus and test our extraction methods.

This paper is organized as follows: the first two sections present the new corpora. Section 3, contains the general extraction system architecture and some results are presented in Section 4.

2. Multimodal comparable corpora

2.1. Euronews

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

Figure 1 shows an example of multimodal comparable data coming from the Euronews website. An audio source of a political news and its text version, both in English, are available along with the equivalent news in French (audio and text modalities). The audio content in the videos are not exactly the same for each language, but are dealing with the same subject. Then, audio in one language and the text content in the other language can be considered as comparable data. This corpus can be used to extract parallel data, at the sentence and the sub-sentential level.

1http://www.projet-depart.org/
2http://www.euronews.com
Euronews website clusters news into several categories or sub-domains (e.g., Sport, Politics, etc.). These categories are preserved in the raw version of the provided corpus (but not in extracted versions). Table 1 show the statistics of our English/French Euronews-LIUM corpus created from French ³ and English news data from 2010-2012 period. This corpus⁴ is composed of a comparable corpus, made of transcriptions (performed with our ASR system, see Section 3.2.) and article content (text found on the webpage). The extracted data performed with the system described in Section 3. are also provided.

2.2. TED
TED-LIUM corpus has been created within the context of the IWSLT'11 evaluation campaign. It has been built from some video talks crawled on the TED (Technology, Entertainment, Design) website⁵. The corpus is made of 773 talks representing 118 hours of speech. We used the English audio part of this corpus and the French text part of the WIT3 parallel corpus⁶, to create the TED multimodal comparable corpus, further called TED-LIUM. Figure 2 shows an example of such multimodal comparable data.

3. Parallel data extraction
3.1. System Architecture
The basic system architecture is described in Figure 3. We begin by extracting comparable sentences with the same method of (Affi et al., 2012) called SentExtract. We can distinguish three steps in this system: automatic speech recognition (ASR), statistical machine translation (SMT) and information retrieval (IR). The ASR system accepts audio data in language L1 as input and generates an automatic transcription. This transcription is then translated by a baseline SMT system into language L2. Then, we use these translations as queries for an IR system to retrieve most similar sentences in the indexed text part of our multimodal comparable corpus. The transcribed text in language L1 and the IR result in language L2 form the comparable sentences.

However, the extracted sentences are of different level of quality, and a filtering step is required in order not to degrade the performance of the baseline system. Previous work make use of different kinds of techniques to filter the extracted sentences, like TER (Snover et al., 2006) between IR query and the returned sentence (Abdul-Rauf and Schwenk, 2011) or bilingual lexicon log-likelihood ratio (Munteanu and Marcu, 2006). One of the drawbacks of filtering is that it can remove a large number of sentences, which often results in a lower impact on the baseline system. Moreover, the withdrawn sentences often contain some useful parallel fragments which is interesting to extract.

³http://fr.euronews.com/
⁴available soon on our website
⁵http://www.ted.com
⁶https://wit3.fbk.eu/
As an example, consider Figure 4, which presents two paragraphs extracted from the news articles presented in Figure 1. Although the articles report on the same event and express overlapping content, the texts cannot be considered as strictly parallel. They contain no fully parallel sentences pairs, but as shown by the boxes in the figure, some parallel phrases do exist in the sub-sentential level.

We developed a parallel phrase extraction system which operates in two steps. First, parallel phrase pair candidates are detected using the IBM1 model (Brown et al., 1993). Then the candidates are filtered with probabilistic translation lexicon (learned on the baseline SMT system training data) to produce parallel phrases using log-likelihood ratio (LLR) method (see (Munteanu and Marcu, 2006) for details). Our technique is similar to that of (Afli et al., 2013) called PhrExtract, but we bypass the need of the TER filtering by using a LLR lexicon. We call this new extended system PhrExtract_LLR.

### 3.2. Baseline systems

The ASR system used in our experiments is an in-house five-pass system based on the open-source CMU Sphinx system (version 3 and 4), similar to the LIUM’08 French ASR system described in (Deléglise et al., 2009). The acoustic models were trained in the same manner, except that we added a multi-layer perceptron (MLP) using the Bottle-Neck feature extraction as described in (Grézl and Fousek, 2008).

To train the language models (LM), we used the SRILM toolkit (Stolcke, 2002). We trained a 4-gram LM on all our monolingual corpus.

The SMT system is a phrase-based system 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 with 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). To train, optimize and test our baseline MT system, we used the data presented in Table 2.

For each comparable corpus (Euronews-LIUM and TED-LIUM, we chose the most appropriate development and test corpus. devEuronews and tstEuronews are the news corpora used in the, respectively, WMT’10 and WMT’11 evaluation campaigns. devTED and tstTED are the official dev and test corpora from the IWSLT’11 international evaluation campaign.

We use the Lemur IR toolkit (Ogilvie and Callan, 2001) for the sentence extraction procedure with default settings. We first index all the French text considering each sentence as a document. This allows to use the translated sentences as queries to the IR toolkit. The IR system make use of the bag of word representation of each sentence and returns the most similar to the query. This sentence is then paired with the English query sentence. By these means we can retrieve the best matching sentences from the French side of the comparable corpus.

### 4. Results

For the sake of comparison, we ran several experiments with two methods. The first one, is PhrExtract_LLR (presented in section 3., and the second one corresponds to the method applied by (Afli et al., 2013) (called PhrExtract as in their paper). Experiments were conducted on English to French TED and Euronews tasks.

PhrExtract uses TER for filtering the result returned by IR, keeping only the phrases which have a TER score below a certain threshold determined empirically. Thus, we filter the selected sentences in each condition with different TER thresholds ranging from 0 to 100 by steps of 10.
The various SMT systems are evaluated using the BLEU score (Papineni et al., 2002).

<table>
<thead>
<tr>
<th>Methods</th>
<th># words (en)</th>
<th># words (fr)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PhrExtract (TER 60)</td>
<td>16.61M</td>
<td>13.82M</td>
</tr>
<tr>
<td>PhrExtract_LLRLR</td>
<td>1.68M</td>
<td>2.27M</td>
</tr>
</tbody>
</table>

Table 3: Number of words and sentences extracted from TED-LIUM corpus with PhrExtract and PhrExtract_LLRLR methods.

<table>
<thead>
<tr>
<th>Methods</th>
<th># words (en)</th>
<th># words (fr)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PhrExtract (TER 50)</td>
<td>2.39M</td>
<td>1.95M</td>
</tr>
<tr>
<td>PhrExtract_LLRLR</td>
<td>636.8k</td>
<td>224.1k</td>
</tr>
</tbody>
</table>

Table 4: Number of words and sentences extracted from Euronews-LIUM corpus with PhrExtract and PhrExtract_LLRLR methods.

Tables 3 and 4 show the statistics of the bitexts extracted from Euronews-LIUM and TED-LIUM. One can note that the sizes of the two sides of the bilingual text extracted from Euronews-LIUM are very different (English side is almost three times larger than French size). This behaviour is not observed on the TED data, and we do not yet explain this fact which requires a more fine grain analysis of the obtained bitexts. These bitexts are injected into our generic training data in order to adapt the baseline MT system.

Tables 5 and 6 present the BLEU scores obtained with training data in order to adapt the baseline MT system. The TER threshold is set to 50 for Euronews-LIUM and 60 for TED-LIUM. This new extraction method bypass the use of the TER filtering which required many experiments in order to determine the best threshold for each task. Moreover, looking at the extracted text sizes in Tables 3 and 4, we can observe that the LLR method generates much less data while obtaining equivalent performance. This suggests that only the most relevant data is extracted by this technique.

We can see in the example in Table 7, that adding the extracted phrases can have a positive effect on translation quality.

## 5. Related Work

There has been considerable amount of work on exploiting comparable corpora, although from a different perspective than the one taken in this paper. (Zhao and Vogel, 2002) proposed an adaptive approach aims at mining parallel sentences from a bilingual comparable news collection collected from the web. A maximum likelihood criterion was used by combining sentence length models and lexicon-based models. The translation lexicon was iteratively updated using the mined parallel data to get better vocabulary coverage and translation probability estimation. In (Yang and Li, 2003), an alignment method at different levels (title, word and character) based on dynamic programming (DP) is presented. The goal is to identify the one-to-one title pairs in an English/Chinese corpus collected from the web. They applied longest common subsequence (LCS) to find the most reliable Chinese translation of an English word. (Resnik and Smith, 2003) propose a web-mining based system called STRAND and show that their approach is able to find large numbers of similar document pairs.

A cross-language information retrieval techniques is used by (Utiyama and Isahara, 2003) to extract sentences from an English/Japanese comparable corpus. They identify similar article pairs, and then, considering them as parallel texts, they align their sentences using a sentence pair similarity score and use DP to find the least-cost alignment over the document pair. (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. (Abdul-Rauf and Schwenk, 2011) bypass the need of the bilingual dictionary by using their own SMT system. They also use simple measures like word error rate (WER) or translation edit rate (TER) in place of a maximum entropy classifier.

In (Munteanu and Marcu, 2006) a first attempt to extract parallel sub-sentential fragments (phrases) from comparable corpora is presented. 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 second type of approach consist in extracting parallel phrases with an alignment-based approach (Quirk et al., 2007; Riesa and Marcu, 2012). These methods are promising, because (Cettolo et al., 2010) 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. But the proposed method in (Quirk et al., 2007) do not significantly improve MT performance and model in (Riesa and Marcu, 2012) is designed for parallel data.

So, it’s hard to say that this approach is actually effective for comparable data.

Since our method can increase the precision of the extraction method, it greatly expands the range of corpora which can be usefully exploited.

6. Conclusion

In this paper, we have presented a new multimodal corpus built to extract parallel data for SMT systems. We also presented a new system to extract parallel fragments from a multimodal comparable corpus. Experiments conducted on TED and Euronews data showed that our method significantly outperforms the existing approaches and improves MT performance both in situations of domain adaptation (TED data) and of in-domain improvement (Euronews).

This is an encouraging result which do not require any threshold empirically determined comparing to TER filtering method. Our approach can be improved in several aspects. A parallel corpus is used to generate the LLR lexicon used for filtering. An alternative method could be to construct a large bilingual dictionary from comparable corpora, and use it in the filtering module. In this case, the lexicon would benefit from containing words specific to the targeted task (in the case of adaptation). Another interesting extension is to carefully select the comparable data to be used in the extraction framework. This selection could be based on a similarity measure computed before the extraction process, and would help to improve the system performances.

7. Acknowledgements

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

8. References


Haithem Afli, Loïc Barrault, and Holger Schwenk. 2013. Multimodal comparable corpora as resources for extracting parallel data: Parallel phrases extraction. *International Joint Conference on Natural Language Processing,* October.


nual Meeting on Association for Computational Linguistics - Volume 1, ACL ’03, pages 160–167, Stroudsburg, PA, USA. Association for Computational Linguistics.


