Building and Using Multimodal Comparable Corpora for Machine Translation

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

(*Received 15 December 2014*)

Abstract

In recent decades, statistical approaches have significantly advanced the development of machine translation systems. 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. 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, built from *Euronews* and *TED* web sites. The audio is transcribed by an automatic speech recognition system, and translated with a baseline statistical machine translation system. We then use information retrieval in a large text corpus in the target language in order to extract parallel sentences/phrases. 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

Statistical machine translation (SMT) systems require a parallel corpus to train the translation model and monolingual data to build the target language model. A parallel corpus, also called bitext, consists in bilingual/multilingual texts aligned at the sentence level. Unfortunately, parallel texts are a sparse resource for many language pairs with exception of English, French, Spanish, Arabic, Chinese and some European languages (Hewavitharana and Vogel, 2011). Furthermore, these corpora are mainly derived from parliamentary proceedings and news wire texts which limits their linguistic domain. For the field of statistical machine translation, this can be problematic, because translation systems trained on data from a specific domain (*e.g.* news) will perform poorly when applied to other domains, *e.g.* scientific articles.

One way to overcome this lack of data is to exploit comparable corpora which are much more easily available (Munteanu and Marcu, 2005). A comparable corpus is a collection of texts composed independently in their respective languages and combined on the basis of similarity of content. These are bi- or multi-lingual documents that are comparable in content and form in various degrees and dimensions. Potential sources of comparable text corpora are multilingual news organizations such as Agence France Presse (AFP), Xinhua, Reuters, CNN, BBC, etc. These texts are widely available on the Web for many language pairs (Resnik and Smith, 2003). The ability to detect these parallel pairs of sentences enables the automatic creation of large 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 for these under-resourced domains. We propose to explore other sources in other modalities (audio in our case) to generate parallel texts for each domain. These kind of data are widely available on the Web for many languages.

In this paper, we explore a proposed methods for generating parallel sentences and phrases from multimodal comparable corpus (audio and text). We would expect a useful technique to meet three criteria:

- Feasibility: the multimodal comparable corpora is useful to extract parallel data.
- Good quality: the quality of the parallel data generated from multimodal corpora should be equivalent to the quality of bitext extracted from text comparable corpora.
- Effectiveness: one of our motivations is to adapt an SMT system to a specific domain, consequently, extracted bitext should improve SMT system performance over a baseline system.

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.

This article describes an experimental framework 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).

We start by giving a brief overview of related work in parallel data extraction and our multimodal comparable corpora developed to test our proposed methods. Section 4 details our parallel sentence extraction method and shows results of our experiments on TED data. In section 6 we present our methods in extracting parallel phrases and the experimental comparison between them using Euronews and TED data. We conclude with a discussion and perspectives of this work.

2 Related work

In the statistical machine translation community, there is a long-standing belief that "there are no better data than more data". Following this idea, a considerable amount of work have been undertaken for discovering parallel sentences in order to improve SMT systems. Thus, there is already an extensive literature related to the problem of comparable corpora, although from a different perspective than the one taken in this paper.

Typically, comparable corpora do not have any information regarding document pair similarity. Generally, there exist many documents in one language which do not have any corresponding document in the other language. Also, when the corresponding information among the documents is available, the documents in question are not literal translations of each other. Thus, extracting parallel data from such corpora requires special algorithms designed for such corpora.

An adaptive approach, proposed by (Zhao and Vogel, 2002), 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 is 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 apply longest common sub-sequence (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.

(Utiyama and Isahara, 2003) uses cross-language information retrieval techniques and dynamic programming 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. (AbdulRauf and Schwenk, 2011) bypass the need of the bilingual dictionary by using proper SMT translations. They also use simple measures like word error rate (WER) or translation edit rate (TER) in place of a maximum entropy classifier.

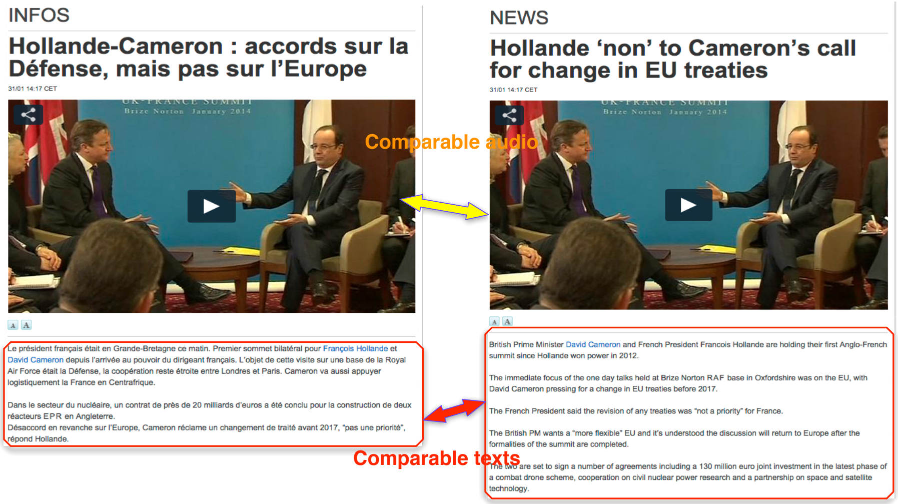
In another way, (Paulik and Waibel, 2009) demonstrated that statistical translation models can be trained in a fully automatic manner from audio recordings of human interpretation scenarios.

In this paper, we are interested in generating a parallel text from a comparable corpora composed by an audio part in one language and a text part in another language. To the best of our knowledge, no systematic empirical research exists addressing the use of comparable audio corpora to extract bitexts.

3 Building Multimodal comparable corpora

In this work, data is extracted from the available news (video and text modalities) on the *Euronews* web site.We also use TED-LIUM (Rousseau et al., 2012) corpus to build our TED multimodal comparable corpus and test our extraction methods.

3.1 Euronews



INFOS

Hollande-Cameron : accords sur la Défense, mais pas sur l'Europe

31/01 14:17 CET

Le président français était en Grande-Bretagne ce matin. Premier sommet bilatéral pour François Hollande et David Cameron depuis l'arrivée au pouvoir du dirigeant français. L'objet de cette visite sur une base de la Royal Air Force était la Défense, la coopération reste étroite entre Londres et Paris. Cameron va aussi appuyer logistiquement la France en Centrafrique.

Dans le secteur du nucléaire, un contrat de près de 20 milliards d'euros a été conclu pour la construction de deux réacteurs EPR en Angleterre.
Désaccord en revanche sur l'Europe, Cameron réclame un changement de traité avant 2017, "pas une priorité", répond Hollande.

NEWS

Hollande 'non' to Cameron's call for change in EU treaties

31/01 14:17 CET

British Prime Minister David Cameron and French President Francois Hollande are holding their first Anglo-French summit since Hollande won power in 2012.

The immediate focus of the one day talks held at Brize Norton RAF base in Oxfordshire was on the EU, with David Cameron pressing for a change in EU treaties before 2017.

The French President said the revision of any treaties was "not a priority" for France.

The British PM wants a "more flexible" EU and it's understood the discussion will return to Europe after the formalities of the summit are completed.

The two are set to sign a number of agreements including a 130 million euro joint investment in the latest phase of a combat drone scheme, cooperation on civil nuclear power research and a partnership on space and satellite technology.

Fig. 1. Example of multimodal comparable data from the *Euronews* web site.

Figure 1 shows an example of multimodal comparable data coming from the *Euronews* web site. 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.
Euronews web site 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 3.1 shows the statistics of our English/French *Euronews-LIUM* corpus created from French [1] and English news data from 2010-2012 period.

This corpus is composed of a comparable corpus, made of transcriptions (performed with the ASR system described in Section 6.1) and article content (text found on the webpage). The extracted data obtained with the extraction system described in Section 6 are also provided.

3.2 TED

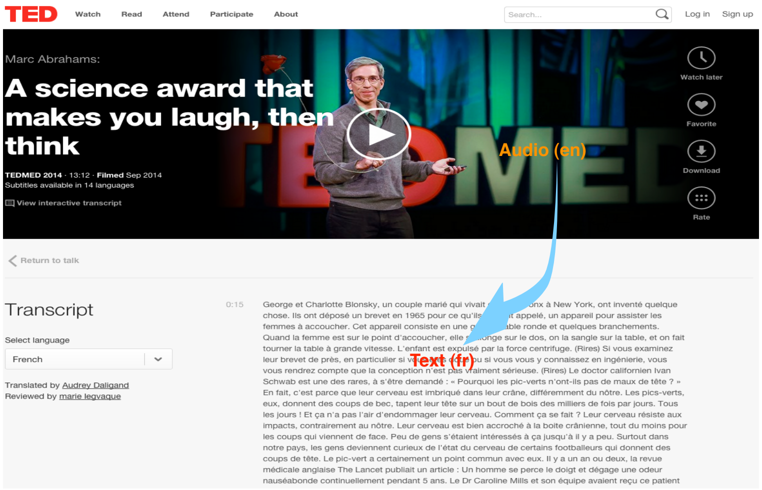
TED-LIUM corpus has been created within the context of the IWSLT'11 evaluation campaign. It has been built from some video talks crawled on the TED (Technology, Entertainment, Design) web site. The corpus is made of 773 talks representing 118 hours of speech.

[1] http://fr.euronews.com/

Sub-Domains	Audio En		Text	
	# words	# sentences	# words Fr	# words En
Business	289 k	7 k	425 K	613k
Sport	81 k	2 k	112 k	102 k
Culture	388 k	12 k	262 k	274 k
Europe	398 k	12 k	302 k	287 k
Life Style	28 k	1 k	18 k	19 k
Politics	806 k	26 k	4 M	4 M
Science	231 k	9 k	147 k	141 k
Total	2.2 M	76 K	6.2 M	6.1 M

Table 1. Size of the Euronews transcribed English audio corpus and English-French texts.

We used the English audio part of this corpus and the French text part of the $WIT3$ parallel corpus [2], to create the TED multimodal comparable corpus, further called TED-LIUM.



Translated by Audrey Daligand
Reviewed by marie legvaque

0:15 George et Charlotte Blonsky, un couple marié qui vivait [illegible] onx à New York, ont inventé quelque chose. Ils ont déposé un brevet en 1965 pour ce qu'ils [illegible] appelé, un appareil pour assister les femmes à accoucher. Cet appareil consiste en une [illegible] able ronde et quelques branchements. Quand la femme est sur le point d'accoucher, elle [illegible] onge sur le dos, on la sangle sur la table, et on fait tourner la table à grande vitesse. L'enfant est expulsé par la force centrifuge. (Rires) Si vous examinez leur brevet de près, en particulier si v[illegible] ou si vous vous y connaissez en ingénierie, vous vous rendrez compte que la conception n'est pas vraiment sérieuse. (Rires) Le doctor californien Ivan Schwab est une des rares, à s'être demandé : « Pourquoi les pic-verts n'ont-ils pas de maux de tête ? » En fait, c'est parce que leur cerveau est imbriqué dans leur crâne, différemment du nôtre. Les pics-verts, eux, donnent des coups de bec, tapent leur tête sur un bout de bois des milliers de fois par jours. Tous les jours ! Et ça n'a pas l'air d'endommager leur cerveau. Comment ça se fait ? Leur cerveau résiste aux impacts, contrairement au nôtre. Leur cerveau est bien accroché à la boite crânienne, tout du moins pour les coups qui viennent de face. Peu de gens s'étaient intéressés à ça jusqu'à il y a peu. Surtout dans notre pays, les gens deviennent curieux de l'état du cerveau de certains footballeurs qui donnent des coups de tête. Le pic-vert a certainement un point commun avec eux. Il y a un an ou deux, la revue médicale anglaise The Lancet publiait un article : Un homme se perce le doigt et dégage une odeur nauséabonde continuellement pendant 5 ans. Le Dr Caroline Mills et son équipe avaient reçu ce patient

Fig. 2. Example of multimodal comparable data from the TED web site.

Figure 2 shows an example of such multimodal comparable data.

This corpus complete the already available TED-LIUM corpus [3], with the extracted parallel data.

[2] `https://wit3.fbk.eu/`
[3] http://www-lium.univ-lemans.fr/fr/content/corpus-ted-lium

corpus	#Talks	Speech (hours)	Gender	
			Male	Female
audio	773	118 h	82 h	36 h

Table 2. TED English audio corpus statistics.

4 Extracting parallel sentences from multimodal comparable corpora

One of our main goals is to address the situation when the data to translate is from a different domain than the training data. In such a condition, the translation quality is generally rather poor.

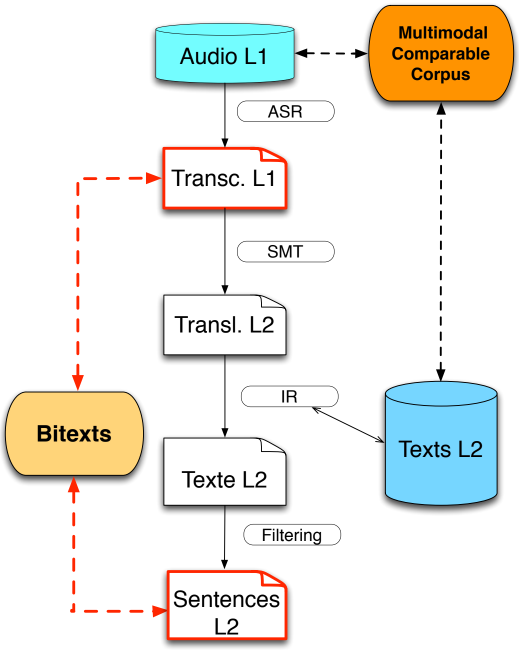


Fig. 3. Extracting parallel texts from multimodal comparable corpora

In this work we seek to improve SMT systems in domains that suffer from resource deficiency by automatically extracting bitexts from a comparable corpora which include audio and text. We propose an extension of the method described in (AbdulRauf and Schwenk, 2011). 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 language L1 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 the most similar sentences in the text part of our multimodal comparable corpus. The transcribed text in L1 and the IR result in L2 form the final bitext. We hope that the errors made by the ASR and SMT systems will not impact too severely the quality of the IR queries, and that the extracted bitext will improve the SMT system.

4.1 Task description

This framework raises several issues. Each step in the system can introduce a certain number of errors. It is important to highlight the feasibility of the approach and the impact of each module on the generated data. Thus, we conducted three different types of experiments, described in Figure 4. In the first experiment (*Exp 1*) we use the reference translations as queries for the IR system. This is the most favorable condition, it simulates the case where the ASR and the SMT systems do not commit any error. In the second experiment (*Exp 2*) we use the reference transcription as input to the SMT system. In this case, the errors only come from the SMT system since no ASR is involved. Finally, the third experiment (*Exp 3*) represents the complete proposed framework, described in section 4. It corresponds to a real scenario.

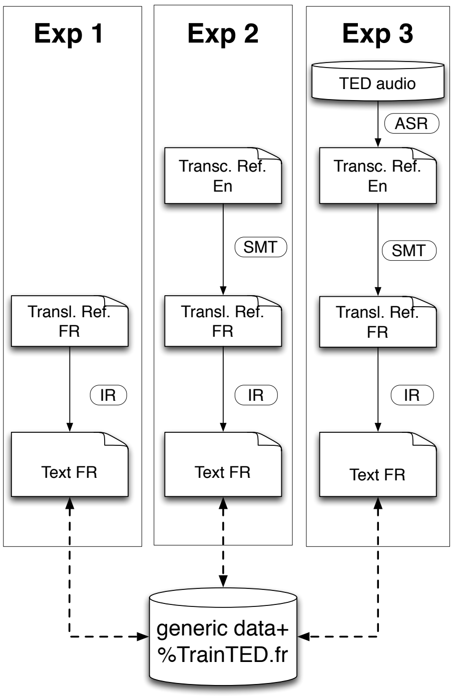


Fig. 4. Different experiments to analyze the impact of the errors of each module

Another issue is the importance of the degree of similarity between the two parts of the comparable corpora. In a real life comparable corpus, we can only expect to find matching sentences for a fraction of the sentences. Therefore, we artificially created four comparable corpora with different degrees of similarity. The source part of our comparable corpus is always the TED corpus (see next section). The target language part of the comparable corpus consists of a large generic corpus plus 25%, 50%, 75% and 100% respectively of the reference translations of the TED corpus.

For each candidate sentence pair, we need to decide whether the two sentences in the pair are mutual translations. Thus, we classify the IR result with TER (Snover et al., 2006) calculated between the query, *i.e.* the automatic translation, and the sentence selected by IR.

In all cases, an evaluation of the approach is necessary. Thus, the final parallel data extracted are re-injected into the baseline system. The various SMT systems are evaluated using the BLEU score (Papineni et al., 2002). This is the most commonly used metric in the domain of automatic machine translation, but the choice of the best metric is actually still an open research issue.

5 Experimental setup

5.1 Data description

In these experiments we used TED comparable corpus described in section 3.2. For MT training, we considered the following corpora among those available: News-Commentary (nc7) and Europarl (eparl7) corpus, the TED corpus provided by IWSLT'11 (*TEDbi*) and a subset of the French–English Gigaword corpus (ccb2). The Gigaword corpus was filtered with the same techniques described in (Rousseau et al., 2011). We transcribed all the TED audio data with the ASR system described in section 5.2 and name it *TEDasr*. Table 3 summarizes the characteristics of those different corpora. Each corpus is labeled whether it is in- or out-of domain with respect to our task.

bitexts	# words	in-domain?
nc7	3.7M	no
eparl7	56.4M	no
ccb2	1.3M	no
TEDbi	1.9M	yes
TEDasr	1.8M	yes

Table 3. MT training data.

The development corpus (dev) consists of 19 talks and represents a total of 4 hours and 13 minutes of speech. We use the same test data as provided by IWSLT'11 organizers for the speech translation task. *dev.outASR* and *test.outASR* are the automatic transcriptions of the development and test corpus respectively. The reference translations are named *dev.refSMT* and *tst.refSMT*. Table 3 summarizes the characteristics of the different corpora used in our experiments.

Name	# words En. ASR	# words Fr. SMT Reference
dev	36k	38k
test	8.7k	9.1k

Table 4. MT development and test data.

5.2 ASR system description

Our ASR system is a five-pass system based on the open-source CMU Sphinx toolkit (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 a multi-layer perceptron (MLP) is added using the Bottle-Neck feature extraction as described in (Grézl and Fousek, 2008). Table 5 shows performances of ASR system on the dev and test corpora.

Corpus	% WER
dev.outASR	19.2%
test.outASR	17.4%

Table 5. Performances of the ASR system
on dev and test data (% WER).

5.3 SMT system description

Our system is a phrase-based system (Koehn et al., 2003) which uses fourteen features functions, 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 based on the Moses SMT toolkit (Koehn et al., 2007) and 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 reordering are extracted using the default settings of the Moses toolkit. The parameters of our system were tuned on *dev.outASR*, using the MERT tool. The language model was trained with the SRI LM toolkit (Stolcke, 2002), on all the French data distributed in IWSLT 2011 evaluation campaign without the TED data. The baseline system is trained with eparl7 and nc7 bitexts.

5.4 IR system

We use the Lemur IR toolkit (Ogilvie and Callan, 2001) for the sentence extraction procedure. We first index all French text data into a database using *Indri Index*. This feature enable us to index our text documents in such a way that using the specialized *Indri*

Query Language we can use the translated sentences as queries to run TF-IDF retrieval in the database. By using these means we can retrieve the best matching sentences from the French side of the comparable corpus. The index data consist of the French part of ccb2 (described in Table 3) and different percentage of the French side of *TEDbi* as described in section 4.1.

5.5 Experimental Results

As mentioned in section 4, the TER score is used as a metric for filtering the result of IR. We only keep the sentences which have a TER score below a certain threshold determined empirically. Thus, in each condition, the retrieved sentences are filtered with a different TER thresholds ranging from 0 to 100. The extracted bitexts are then added to our generic training data in order to adapt the baseline system. Figures 5, 6, 7 and 8 present the BLEU score obtained for these different experimental conditions.

In *Exp2*, we use automatic translations for the IR queries. One can hope that IR itself is not too much affected by the translation errors, but this will be of course the fact for the filtering based on the TER score. (AbdulRauf and Schwenk, 2011) propose to vary the TER threshold between 0 and 100 and to keep the threshold value that maximizes the BLEU score once the corresponding extracted bitexts were injected into the generic system. We did not observe such a clear maximum in our experiments and the BLEU score increases almost continuously. Nevertheless, in order to limit the impact of noisy sentences, we decided to only keep the sentences with a TER score below the threshold of 80. One can observe that the BLEU score of the adapted system matches the one of *Exp1* in most of the cases. Therefore, we conclude that the errors induced by the SMT system have no major impact on the performance of the parallel sentence extraction algorithm. These findings are in line with those of (AbdulRauf and Schwenk, 2011).

These results show that the choice of the appropriate TER threshold depends on the type of data. Our baseline SMT system trained with generic bitext only achieves a BLEU score of 22.93. In *Exp1*, we use the reference translations as query and the IR should in theory find all the sentences in the large corpus with a TER of zero. It can happen that our generic corpus also contains some similar sentences which are “accidentally” retrieved. The four figures show that the IR does indeed work as expected: the observed improvement in the BLEU score does not depend on the TER threshold (with the exception of some noise) since all the sentences have a TER of zero. The achieved improvement depends of course on the amount of TED bitexts that are injected in our comparable corpus: the BLEU score increases from 22.93 to 24.14 when 100% is injected while we only obtain a BLEU score of 23.62 when 20% is injected. These results give us the upper bound that we could expect to get when extracting parallel sentences from this particular multimodal comparable corpus.

Finally, in *Exp3*, we use automatic speech recognition on the source side of the comparable corpus. Our ASR system has a WER of about 18%. These errors on the source side can obviously lead to wrong translations and have a negative impact on the IR process. It

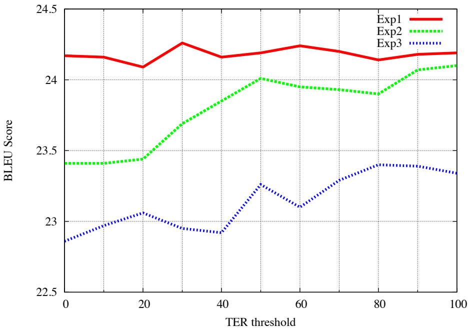


Fig. 5. BLEU score on dev using SMT systems adapted with bitexts extracted from *ccb2 + 100% TEDbi* index corpus.

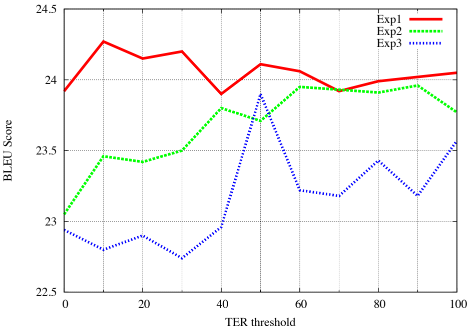


Fig. 6. BLEU score on dev using SMT systems adapted with bitexts extracted from *ccb2 + 75% TEDbi* index corpus.

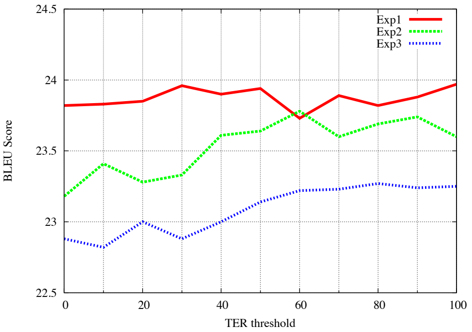


Fig. 7. BLEU score on dev using SMT systems adapted with bitexts extracted from *ccb2 + 50% TEDbi* index corpus.

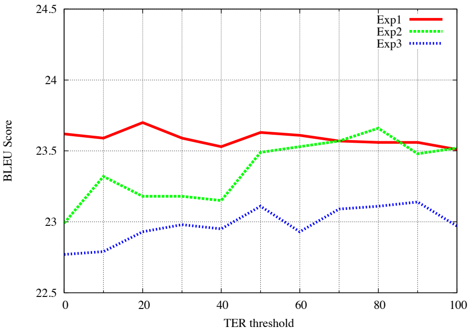


Fig. 8. BLEU score on dev using SMT systems adapted with bitexts extracted from *ccb2 + 25% TEDbi* index corpus.

is important to note that these automatic transcriptions represent the source side of our extracted parallel corpus. By these means, eventual transcription errors should less affect the translation system since it is unlikely that wrong source phrases will be used to translate other texts. We observed in our experiments that the extracted sentences do improve the SMT system. The performance in this stage is only 0.5 BLEU points below those obtained in *Exp1* or *Exp2*.

Table 6 lists the adaptation results of the baseline system in different conditions. It shows that starting with a BLEU score of 23.96% on the test set for the baseline system, adaptation with automatically extracted in-domain bitext resulted in better results in all conditions (between 1.18 in *Exp1* and 0.73 BLEU points in *Exp3*).

Table 7 provides an analysis of the performance in function of the degree of parallelism of the comparable corpus. Remember that the whole TED corpus (in text version) amounts

Experiment	Dev	Test
Baseline	22.93	23.96
Exp1	24.14	25.14
Exp2	23.90	25.15
Exp3	23.40	24.69

Table 6. BLEU scores on dev and test after adaptation of a baseline system with bitexts extracted in conditions *Exp1, Exp2 and Exp3* (100% TEDbi).

to about 1.8M words. We were able to automatically extract about 400k words of new bitexts, *i.e.* a little more than 20%. If less data is injected, the amount of extracted data decreases linearly.

Experiments	Dev	Test	# injected words
Baseline	22.93	23.96	-
25% TEDbi	23.11	24.40	∼110k
50% TEDbi	23.27	24.58	∼215k
75% TEDbi	23.43	24.42	∼293k
100% TEDbi	23.40	24.69	∼393k

Table 7. BLEU scores for different degrees of parallelism of the comparable corpus.

We argue that this is an encouraging result since we automatically aligned source audio in one language with texts in another language, without the need of human intervention to transcribe and translate the data. The TED corpus contains only 118 hours of speech. There are many domains for which much larger amounts of untranscribed audio in one language and related texts in another language are available, for instance news. However, the quantity of extracted sentences is still insufficient in comparison with the rejected data (by TER filtering). Figure 9 illustrates the selection process as a binary selection process where the non-selected sentences are simply discarded. In order to make a better use of the rejected sentences, we proposed to use them as unsupervised data. Unsupervised adaptation method ((Schwenk, 2008)) is described in Figure 10. A baseline (generic) SMT is used to translate some sentences, which are then filtered using the translation score (provided by moses decoder in our case). The data is then added to the generic SMT training data in order to train the adapted system.

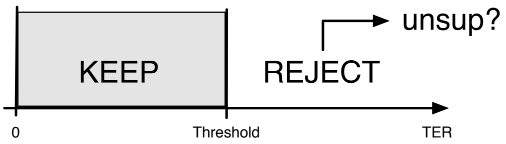


Fig. 9. Principle of parallel sentence extraction with TER filtering.

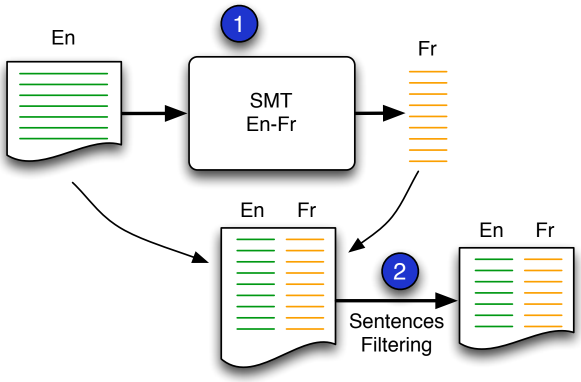


Fig. 10. Principle of the unsupervised training method (*Unsup*).

In our case, all the sentences rejected by the TER filtering are then considered as candidate for unsupervised training. If their translation score is above a certain threshold (determined empirically), then they are kept. The data (called *Unsup*) is added to the generic SMT training data and a new system is trained. The comparative results are presented in Figure 11.

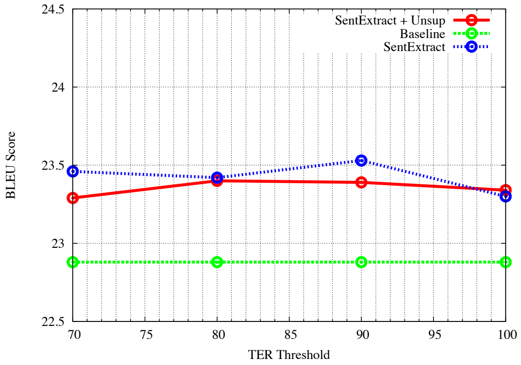


Fig. 11. BLEU score using the Unsupervised training method *Unsup* compared to the baseline system and the proposed extraction method *SentExtract*.

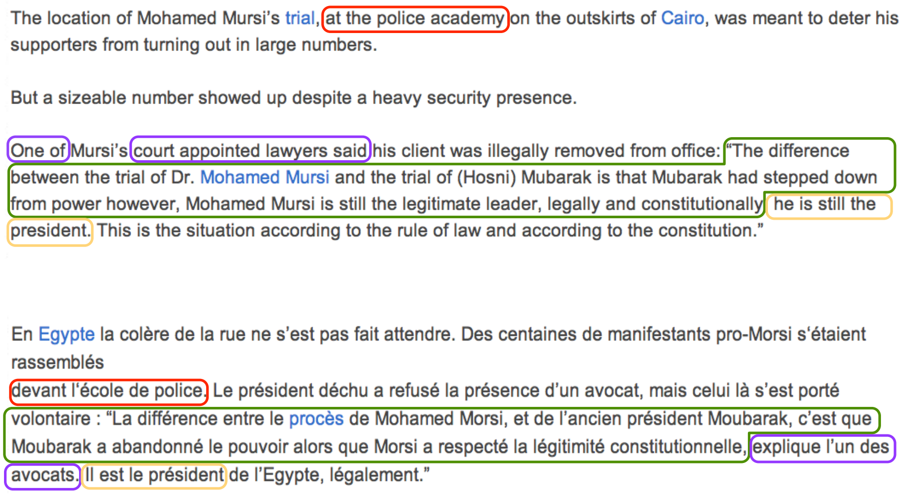
It shows an analysis of the performance of the *Unsup* and *SentExtract* methods measuring by the BLEU score for different TER threshold. We can observe that adding the *Unsup* data does not provide additional improvement compared to using data extracted with our

SentExtract method only. An explanation of this is that the segments are too long for IR to retrieve good parallel ones. As for the *Unsup* method, the baseline system can not provided good translations for specialized sentences, which result in no performance increase.

This clearly suggest to look at the sub-sentential level in order to make the most of the available in-domain data. This is the purpose of the next section.

6 Extracting parallel phrases

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; AbdulRauf 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. 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 12, which presents two paragraphs from news articles from the English and French editions of the Euronews web site[4]. The paragraphs 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.



The location of Mohamed Mursi's trial, at the police academy on the outskirts of Cairo, was meant to deter his supporters from turning out in large numbers.

But a sizeable number showed up despite a heavy security presence.

One of Mursi's court appointed lawyers said his client was illegally removed from office: "The difference between the trial of Dr. Mohamed Mursi and the trial of (Hosni) Mubarak is that Mubarak had stepped down from power however, Mohamed Mursi is still the legitimate leader, legally and constitutionally he is still the president. This is the situation according to the rule of law and according to the constitution."

En Egypte la colère de la rue ne s'est pas fait attendre. Des centaines de manifestants pro-Morsi s'étaient rassemblés
devant l'école de police. Le président déchu a refusé la présence d'un avocat, mais celui là s'est porté volontaire : "La différence entre le procès de Mohamed Morsi, et de l'ancien président Moubarak, c'est que Moubarak a abandonné le pouvoir alors que Morsi a respecté la légitimité constitutionnelle, explique l'un des avocats. Il est le président de l'Egypte, légalement."

Fig. 12. Exemples of two comparable paragraphs with highlight on bilingual segments.

There has been considerable amount of work on comparable corpora for discovering parallel phrases.

[4] www.euronews.com/

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 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 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 the 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.

We propose a method based on a combination of the the translation approach and the alignment one for the extraction of parallel phrases.

6.1 System architecture

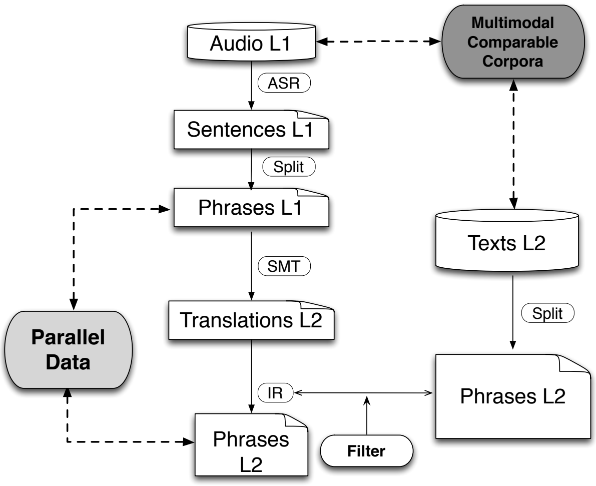


Fig. 13. Principle of the *PhrExtract* parallel phrase extraction system from multimodal comparable corpora.

The system architecture is described in Figure 13. As in the *SentExtract* method, 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 previously split into phrases as well. 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.

We report an extension of the work of the section 4 by splitting the transcribed sentences and the text parts of the multimodal corpus into phrases with a length (empirically chosen) between two to ten tokens. All combinations of two to ten word sequences are extracted from each sentence of the corpus.

However, the extracted phrases are of different level of quality, and a filtering step is required in order not to degrade the performance of the baseline system.

One of the drawbacks of TER filtering method is that it can remove a large number of phrases, which often results in a lower impact on the baseline system.

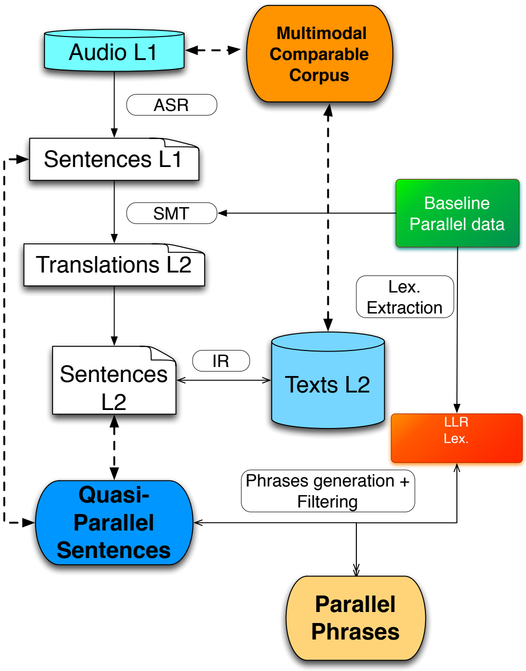


Fig. 14. Principle of the *PhrExtract_LLR* parallel data extraction system from multimodal comparable corpora.

In order to resolve this problem, we proposed a new parallel phrase extraction system presented in Figure 14. We begin by extracting comparable sentences with the same method as presented in section 4 called *SentExtract*. Then, we apply 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). This technique is similar to that of the *PhrEx-*

Corpus	# words En	# words Fr
devEuronews	74k	84k
tstEuronews	61k	70k
devTED	36k	38k
tstTED	8.7k	9.1k

Table 8. MT development and test data.

tract system, but we bypass the need of the TER filtering by using an LLR lexicon. We call this new extended system *PhrExtract_LLR*.

6.2 Data description

To train, optimize and test our baseline MT system, we used the data presented in Table 3. 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 WMT'10 and WMT'11 evaluation campaigns, respectively. *devTED* and *tstTED* are the official development and test corpora from the IWSLT'11 international evaluation campaign.

6.3 Results

For the sake of comparison, we ran several experiments with the two methods. The first one, is *PhrExtract_LLR* (presented in section 4), and the second one corresponds to the method described in the section 6 (called *PhrExtract*). 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.

Tables 9 and 10 show the statistics of the bitexts extracted from Euronews-LIUM and TED-LIUM. These bitexts are injected into our generic training data in order to adapt the baseline MT system.

Tables 11 and 12 present the BLEU scores obtained with the best bitext extracted from each multimodal corpus with *PhrExtract* and *PhrExtract_LLR* methods. The TER threshold is set to 50 for Euronews-LIUM and 60 for TED-LIUM.

In the experiment with TED data, we seek to adapt our baseline SMT system to a new domain. We can see in table 11 that our new system obtains similar results as the *PhrExtract* method. This means that the extracted texts are useful for adaptation purpose.

Methods	# words (en)	# words (fr)
PhrExtract (TER 60)	16.61M	13.82M
PhrExtract_LLR	1.68M	2.27M

Table 9. Number of words and sentences extracted from TED-LIUMcorpus with *PhrExtract* and *PhrExtract_LLR* methods.

Methods	# words (en)	# words (fr)
PhrExtract (TER 50)	2.39M	1.95M
PhrExtract_LLR	236.8k	224.1k

eu

Table 10. Number of words and sentences extracted from Euronews-LIUM corpus with *PhrExtract* and *PhrExtract_LLR* methods.

Systems	devTED	tstTED
Baseline	22.93	23.96
PhrExtract (TER 60)	23.70	24.84
PhrExtract_LLR	23.63	24.88

Table 11. BLEU scores on devTED and tstTED after adaptation of a baseline system with bitexts extracted from TED-LIUMcorpus.

Systems	devEuronews	tstEuronews
Baseline	25.19	22.12
PhrExtract (TER 50)	30.04	27.59
PhrExtract_LLR	30.00	27.47

Table 12. BLEU scores on devEuronews and tstEuronews after adaptation of a baseline system with bitexts extracted from Euronews-LIUM corpus.

The same behavior is observed on Euronews task (Table 12). The extracted text can be used to improve an existing SMT system already trained on the same kind of data.

This new extraction method bypass the use of the TER filtering. This avoid the need of many experiments to determine the best threshold for each task. Moreover, looking at the extracted text sizes in Tables 9 and 10, we can observe that the LLR method generate less data while obtaining equivalent performance. This suggests that only the most relevant data is extracted by this technique.

Source EN (ASR output)	for me it's a necessity to greece stays in the euro zone and that greece gets the chance to get back on track the problem
Baseline FR	**pour moi une nécessité pour la grèce** reste dans la zone euro et que la **grèce** aura la chance **de revenir sur la piste problème**
Adapted FR	**Je vois la nécessité que la Grèce** reste dans la zone euro et que la **Grèce** aura la chance **de se remettre sur pieds .**

Table 13. Example of translation quality improvements of the baseline MT system after adding parallel data extracted from Euronews corpus.

We can see in the example in Table 13, that adding the extracted phrases can have a positive effect on translation quality.

7 Conclusion

In this paper, we presented a new system to extract parallel sentences and phrases from a multimodal comparable corpus. We have proposed to extend the exploitation of the comparable corpora to multimodal comparable corpora, i.e. the source side is available as audio and the target side as text. This is achieved by combining a large vocabulary speech recognition system, a statistical machine translation system and information retrieval.

We validate the feasibility of our approach by a set of experiments to analyze the impact of the errors committed by each module. Experiments conducted on TED and Euronews data showed that our method significantly outperforms the existing approaches and improves MT performance both in two different situations. The first is domain adaptation; for the TED task, the generated data help a baseline system unadapted to this particular domain to improve its performance. For the Euronews task, additional in-domain data are injected in an already adapted system to the news domain.

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).

8 Acknowledgements

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

References

AbdulRauf, S. and Schwenk, H. (2011). Parallel sentence generation from comparable corpora for improved smt. *Machine Translation*.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation: parameter estimation. *Comput. Linguist.*, 19:263–311.

Cettolo, M., Federico, M., and Bertoldi, N. (2010). Mining parallel fragments from comparable texts. *Proceedings of the 7th International Workshop on Spoken Language Translation*.

Deléglise, P., Estève, Y., Meignier, S., and Merlin, T. (2009). Improvements to the LIUM french ASR system based on CMU Sphinx: what helps to significantly reduce the word error rate? In *Interspeech 2009*, Brighton (United Kingdom).

Fung, P. and Cheung, P. (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.

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

Grézl, F. and Fousek, P. (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.

Hewavitharana, S. and Vogel, S. (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.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (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.

Koehn, P., Och, F. J., and Marcu, D. (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.

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

Munteanu, D. S. and Marcu, D. (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.

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

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (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.

Paulik, M. and Waibel, A. (2009). Automatic translation from parallel speech: Simultaneous interpretation as mt training data. *ASRU, Merano, Italy*.

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

Resnik, P. and Smith, N. A. (2003). The web as a parallel corpus. *Comput. Linguist.*, 29:349–380.

Riesa, J. and Marcu, D. (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.

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

Rousseau, A., Deléglise, P., and Estève, Y. (2012). Ted-lium: an automatic speech recognition dedicated corpus. *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12).*

Schwenk, H. (2008). Investigations on large-scale lightly-supervised training for statistical machine translation. *In International Workshop on Spoken Language Translation*, pages 182–189.

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

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

Utiyama, M. and Isahara, H. (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.

Yang, C. C. and Li, K. W. (2003). Automatic construction of english/chinese parallel corpora. *J. Am. Soc. Inf. Sci. Technol.*, 54:730–742.

Zhao, B. and Vogel, S. (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.