Parallel texts extraction from multimodal comparable corpora

Haithem Afli, Loïc Barrault, and Holger Schwenk

LIUM, University of Le Mans

Abstract. Statistical machine translation (SMT) systems depend on the availability of domain-specific bilingual parallel text. However parallel corpora are a limited resource and they are often not available for some domains or language pairs. We analyze the feasibility of extracting parallel sentences from multimodal comparable corpora. This work extends the use of comparable corpora by using audio sources instead of texts on the source side. The audio is transcribed by an automatic speech recognition system and translated with a baseline SMT system. We then use information retrieval in a large text corpus of the target language to extract parallel sentences. We have performed a series of experiments on data of the IWSLT'11 speech translation task that shows the feasibility of our approach.

Keywords: statistical machine translation, automatic speech recognition, multimodal comparable corpora, extraction of parallel sentences

1 Introduction

The construction of a statistical machine translation (SMT) 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 aligned at the sentence level.

Unfortunately, parallel text are a sparse resource for many language pairs with exception of English, French, Spanish, Arabic, Chinese and some European languages [6]. Furthermore, these kind of corpus is mainly derived from parliamentary proceedings, some news wire or the United Nations. 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 [9]. A comparable corpus is a collection of texts composed independently in the respective languages and combined on the basis of similarity of content. These are documents in one to many languages, 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 [13]. The degree of

parallelism can vary considerably, from noisy parallel texts, to quasi parallel texts [3]. 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. What we need is to explore other kind of sources like audio to generate parallel texts for each domain.

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

- Feasibility: the multimodal comparable corpora can be useful to generate a parallel text.
- Good quality: the quality of the parallel text generated from multimodal corpora should be comparable with bitext extracted from text comparable corpora.
- Effectiveness: since one of our motivations for exploiting comparable corpora is to adapt a SMT system for a specific domain, extracted bitext needs to be useful to improve SMT performance.

In the following sections, we will first describe the related work in parallel text extraction from comparable corpora for SMT. In section 3, we will describe our method. Section 4 describes our experiments and results.

2 Related work

In the machine translation community, there is a long-standing belief that "there are no better data than more data". Following this idea, many works have been undertaken for mining large amounts of data 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 don't have any information regarding document pair similarity. Generally, there exist many documents in one language which don't 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 [19], 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 [18], an alignment method at different level (title, word and character) based on dynamic programming is presented. The goal is to identify the one-to-one title pairs in the English/Chinese corpus collected from the web,

They applied longest common sub-sequence (LCS) to find the most reliable Chinese translation of an English word. [13] propose a web-mining based system called STRAND and show that their approach is able to find large numbers of similar document pairs.

[17] 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.

[9] 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. [1] 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, [12] 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 other language.

3 Extracting parallel texts from multimodal comparable corpora

3.1 Basic Idea

Our main experimental framework is designed to address the situation when we translate data from a domain different than the training data. In such condition, the translation quality is rather poor.

In this proposed scenario of machine translation in specific domains, we seek to improve SMT systems in domains that suffer from resource deficiency, by automatically extracting bitexts from an audio and text comparable corpora. The solution we propose, based on an extension of the methods of [1], is described in figure 1 that shows our system architecture. The overall system consists of tree 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 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 generated bitext will benefit to the system.

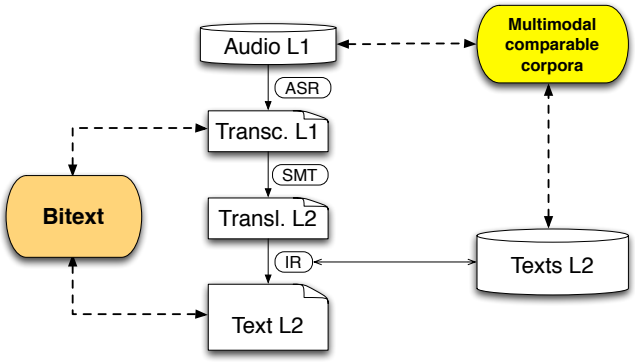

Fig. 1. Extracting parallel texts from multimodal comparable corpora

3.2 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 2. 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 of the SMT system. In this case, the errors comes only from the SMT system. Finally, the third experiment (*Exp 3*) represents the complete proposed framework, described in section 3.1. It corresponds to a real scenario.

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. 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 the TER [15] 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 which is used to translate the test data. Finally, the BLEU [11] score is computed for the SMT evaluation.

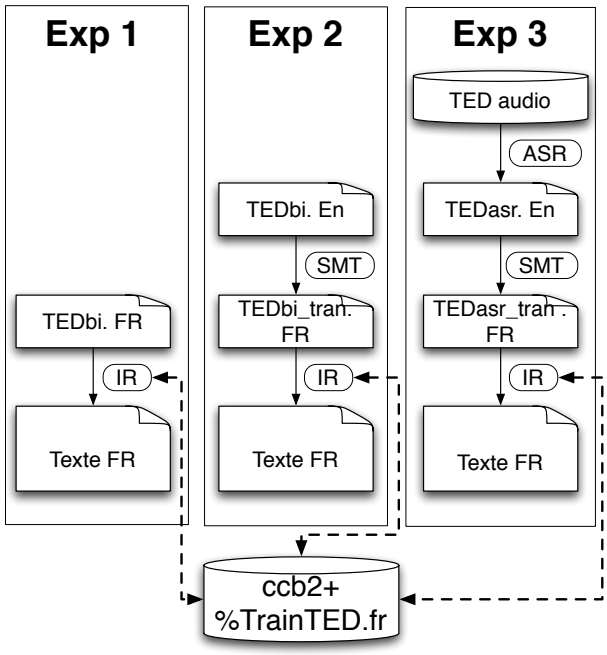

Fig. 2. Diffrent experiments to analyze the impact of the errors of each module

4 Experimental setup

4.1 Data description

Our comparable corpus consist of two monolingual corpora, one spoken in English and one written in French. In our experiments we use all available data from IWSLT'11 evaluation campaign. The goal of this task, detailed in [14], is to translate spoken presentations from TED[1] from English into French.

For MT training, we considered the following corpora among those available: the latest versions of News-Commentary (nc7) and Europarl (eparl7) corpus, the TED corpus provided by IWSLT'11 (TEDbi) and a subset of the French–English 10^9 Gigaword corpus (ccb2). The Gigaword corpus was filtered with the same techniques described in [14]. We name it ccb2_px70. We decoded all the TED audio data with the ASR system described in section 4.2 and name it TEDasr. Table 1 summarizes the characteristics of those different corpora. Each corpus is labeled whether it is in- or out-of domain with respect to our task.

The development corpus (dev) consists of 19 talks and represents a total of 4 hours and 13 minutes of speech. Among these, male speech counts for 3 hours and 14 minutes, while female speech represents 59 minutes. We use the same test data as provided by IWSLT'11 for the speech translation task. dev.outASR

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

bitexts	# words	in-domain ?
nc7	3.7M	no
eparl7	56.4M	no
ccb2_px70	1.3M	no
TEDbi	1.9M	yes
TEDasr	1.8M	yes

Table 1. MT training data.

and test.outASR are the automatic transcriptions of respectively the development and test corpus. The reference translations are named dev.refSMT and tst.refSMT respectively. Table 2 summarizes the characteristics of the different corpora used in our experiments.

Dev	# words	Test	# words
dev.outASR	36k	tst.outASR	8.7k
dev.refSMT	38k	tst.refSMT	9.1 k

Table 2. Dev and test data.

4.2 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 [2]. 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 [5]. Table 3 shows performances of ASR system on the dev and test corpora. The SRILM toolkit [16] was used for language modeling (LM).

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

Table 3. Performances of the ASR system on dev and test data (% WER).

4.3 SMT system description

Our system is a phrase-based system [8] 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

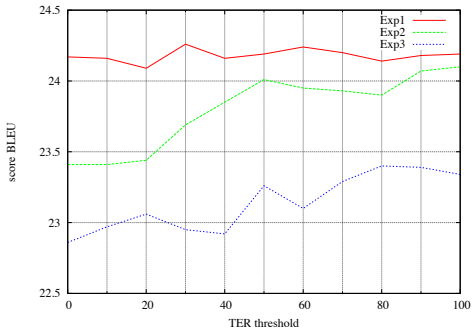

Fig. 3. BLEU score on dev using SMT systems adapted with bitexts extracted from *ccb2 + 100% TEDbi* index corpus.

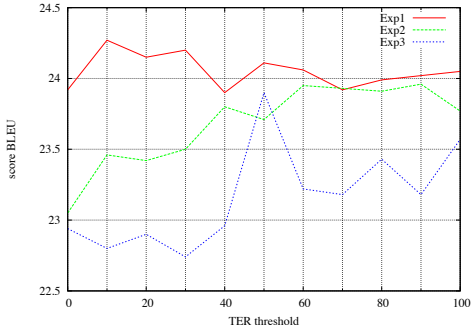

Fig. 4. BLEU score on dev using SMT systems adapted with bitexts extracted from *ccb2 + 75% TEDbi* index corpus.

target language model. It is based on the Moses SMT toolkit [7] and is constructed as follows. First, word alignments in both directions are calculated. We used the multi-threaded version of the GIZA++ tool [4]. 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 [16], on all the French data distributed in IWSLT 2012 evaluation campaign without TED data. The baseline system is trained with eparl7 and nc7 bitexts.

4.4 IR system

We used the Lemur IR toolkit [10] for the sentence extraction procedure. We first index all French text data into a database using *Indri Index*. This feature enabled 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 these means we can retrieve the best matching sentences from the French side of the comparable corpus. The index data consist

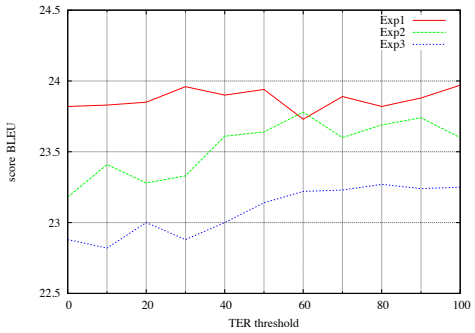

Fig. 5. BLEU score on dev using SMT systems adapted with bitexts extracted from *ccb2 + 50% TEDbi* index corpus.

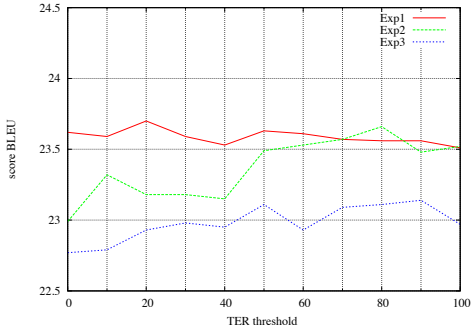

Fig. 6. BLEU score on dev using SMT systems adapted with bitexts extracted from *ccb2 + 25% TEDbi* index corpus.

of the french part of ccb2_px70 and different percentage of the French side of TEDbi as described in section 3.2.

4.5 Experimental Results

As mentioned in section 3.1, the TER score is used as a metric for filtering the result of IR. We keep only the sentences which have a TER score below a certain threshold determined empirically. Thus, we filtered the selected sentences in each condition with different TER thresholds from 0 to 100. The extracted bitexts were added to our generic training data in order to adapt the baseline system. Figure 3, 4, 5 and 6 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. [1] 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

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

Table 4. BLEU scores on dev and test after adaptation of a baseline system with bitexts extracted in conditions *Exp1, Exp2 and Exp3* (100% TEDbi).

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 70. 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 [1].

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 IR should in theory find all the sentences in the large corpus with a TER of zero. It can happen that our generic ccb2 corpus also contains some similar sentences which are "accidentally" retrieved. The four figures show that 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 increases from 22.93 to 24.14 when 100% is injected while we only achieve 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 our 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. One must note that these automatic transcriptions represent the source side of our extracted parallel corpus. Error eventually contained in the transcriptions would less affect the translation system since the data to translate would rarely contain such errors. Nevertheless, we observed in our experiments that these extracted sentences do improve the SMT system. The performance is actually only 0.5 BLEU points below those obtained in *Exp1* or *Exp2*.

Table 4 lists the adaptation results of the baseline system in different conditions for the development and test set. It shows that starting with a baseline BLEU of 23.96% on the test set, adaptation with automatically extracted in-

domain bitext resulted in an improvement 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 corpus amounts to about 1.8M words. We were able to extract automatically 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.

We measure the performance of the extraction process by computing the precision and recall. Precision is computed as the ratio of sentence pairs correctly identified as parallel considering the chosen threshold to the total number of sentence pairs extracted. Recall is computed as the ratio of parallel sentence pairs extracted by the extraction system to the total number of sentences i.e., in-domain injected (TEDbi) and out-of-domain (ccb2). Both are expressed as percentages. Then:

$$Precision = \frac{100 * nb\ parallel\ sentences\ retrieved}{total\ nb\ sentences\ extracted} \tag{1}$$

$$Recall = \frac{100 * nb\ parallel\ sentences\ extracted}{total\ nb\ sentences\ bitext} \tag{2}$$

The combination of the two measures with an equal weight gives the F1 measure, presented by the following expression:

$$F1 = \frac{2.Recall.Precision}{Recall + Precision} \tag{3}$$

However as we can see in figure 6, that the values of the Recall is stable because we extract the same number of sentences in all of our experiments. So we can consider the values of the Recall in equation 3 by a constant α as in the following expression:

$$F1 = \frac{2.\alpha.Precision}{\alpha + Precision} \tag{4}$$

We can see clearly in figure 7, the performance in terms of F1 measure of our system extraction depend of the degree of parallelism of the comparable corpus. This curve validate the previous result in terms of BLEU.

Other performance metrics could measure the *incorrectness of the system extraction. We define False Acceptation Rate (FAR) as the probability that the system incorrectly accepts the non-parallel sentences, and the False Rejection Rate (FER) as the probability that the system incorrectly rejects the parallel sentences. Hence, FAR an FFR are given by the following expressions:*

$$FAR = \frac{nb\ no\_parallel\ sentences\ extracted}{total\ nb\ parallelsentences\ extracted} \tag{5}$$

$$FRR = \frac{nb\ parallel\ sentences\ no\_extracted}{total\ nb\ parallel\ sentences\ injected} \tag{6}$$

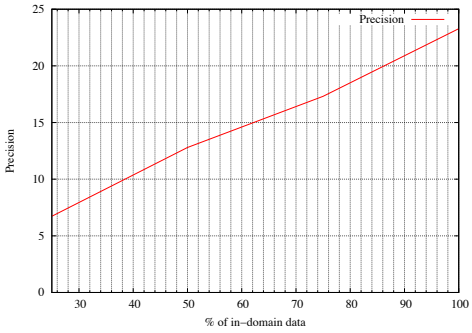

Table 5. Precision of the extraction system

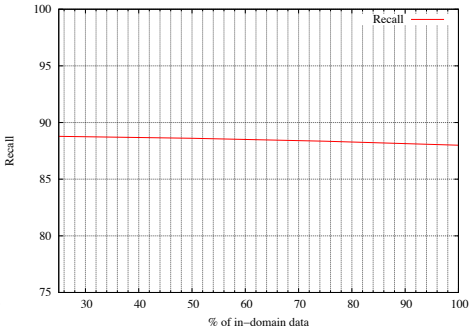

Table 6. Recall of the extraction system

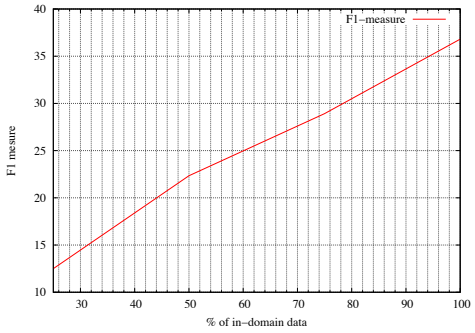

Fig. 7. F1 measure of the system extraction

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 can decide that the system extraction has a good performance if both FAR and FRR have jointly minimal values.

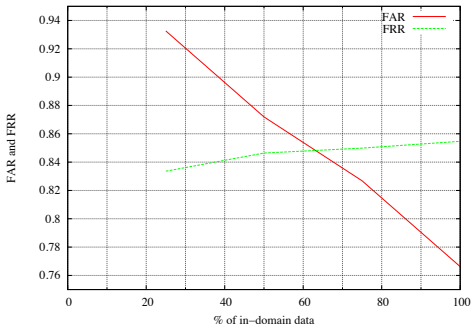

Fig. 8. Curves of the performance of the system extraction in term of FAR and FRR

Figure 8 shows that the degree of incorrectness increase when the degree of parallelism of the comparable corpus decrease.

The idea behind using these metrics is to have a look at the evaluation of our system from many different angles. We came to the conclusion that these different methods confirms the same results about the importance of degree 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 like news.

5 Conclusion

Domain specific parallel data is crucial to train well performing SMT systems, but it is often not easily and freely available. During the last years, there are

several works that propose to exploit comparable corpora for this purpose and many algorithms were proposed to extract bitexts from a comparable corpus.

In this paper, we have proposed to extend this concept 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 an information retrieval toolkit. We validate the feasibility of our approach by a set of experiments to analyze the impact of the errors committed by each module. We were able to improve a generic SMT system to the task of lecture translation by 0.7 BLEU point by extracting parallel data from a multimodal comparable corpus composed of 118 hours of untranscribed speeches in the source language and 1.8M words of translations injected into a large generic corpus.

References

1. *Sadaf Abdul-Rauf and Holger Schwenk. Parallel sentence generation from comparable corpora for improved smt.* Machine Translation*, 2011.*
2. *P. Deléglise, Y. Estève, S. Meignier, and T. Merlin. 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), 6-10 september 2009.*
3. *Pascale Fung and Percy Cheung. Multi-level bootstrapping for extracting parallel sentences from a quasi-comparable corpus. In* Proceedings of the 20th international conference on Computational Linguistics*, COLING '04, 2004.*
4. *Qin Gao and Stephan Vogel. Parallel implementations of word alignment tool. In* Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, SETQA-NLP '08, pages 49–57, 2008.*
5. *F. Grézl and P. Fousek. Optimizing bottle-neck features for LVCSR. In* 2008 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 4729–4732. IEEE Signal Processing Society, 2008.*
6. *Sanjika Hewavitharana and Stephan Vogel. 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, 2011.*
7. *Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 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, 2007.*
8. *Philipp Koehn, Franz Josef Och, and Daniel Marcu. 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, 2003.*
9. *Dragos Stefan Munteanu and Daniel Marcu. Improving Machine Translation Performance by Exploiting Non-Parallel Corpora.* Computational Linguistics*, 31(4):477–504, 2005.*
10. *Paul Ogilvie and Jamie Callan. Experiments using the lemur toolkit.* Procedding of the Trenth Text Retrieval Conference (TREC-10)*, 2001.*

11. *Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, 2002.*
12. *Matthias Paulik and Alex Waibel. Automatic translation from parallel speech: Simultaneous interpretation as mt training data.* ASRU, Merano, Italy*, Decembre 2009.*
13. *Philip Resnik and Noah A. Smith. The web as a parallel corpus.* Comput. Linguist.*, 29:349–380, September 2003.*
14. *Anthony Rousseau, Fethi Bougares, Paul Deléglise, Holger Schwenk, and Yannick Estève. LIUM's systems for the IWSLT 2011 speech translation tasks.* International Workshop on Spoken Language Translation 2011*, 2011.*
15. *Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation.* Proceedings of Association for Machine Translation in the Americas*, pages 223–231, 2006.*
16. *A. Stolcke. SRILM - an extensible language modeling toolkit. In* International Conference on Spoken Language Processing*, pages 257–286, November 2002.*
17. *Masao Utiyama and Hitoshi Isahara. 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, 2003.*
18. *Christopher C. Yang and Kar Wing Li. Automatic construction of english/chinese parallel corpora.* J. Am. Soc. Inf. Sci. Technol.*, 54:730–742, June 2003.*
19. *Bing Zhao and Stephan Vogel. Adaptive parallel sentences mining from web bilingual news collection. In* Proceedings of the 2002 IEEE International Conference on Data Mining*, ICDM '02, pages 745–, Washington, DC, USA, 2002. IEEE Computer Society.*