IJAER/Nov-Dec 2022/Volume-11/Issue-6 ISSN: 2278-9677

International Journal of Arts & Education Research

MACHINE TRANSLATION WITH COMPARABLE CORPORA

Manpreet Kaur

Research Scholar Tantia University

Dr. Pawan Pareek

Associate Professor Tantia University

Abstract

The development of a statistical machine translation (SMT) system necessitates the use of a parallel corpus for the purpose of educating the translation model as well as monolingual data for the purpose of constructing the target language model. Texts that are written in more than one language make up a parallel corpus, which is also referred to as bitext. Sadly, parallel texts are a scarce resource for a wide variety of language pairings. Using similar corpora that are much more readily accessible is one method for getting around the fact that there are not enough data. In this study, we show the corpus that was constructed for the purpose of automatically extracting parallel data from multimodal comparable corpora, such as those found on the TED and Euronews websites. We present the information that is included inside each corpus as well as the process that was used by our newly developed extraction technique to get the parallel data. In this paper, we analyse the outcomes of the bitext extraction and explain the strategies that were examined for employing multimodal corpora.

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

INTRODUCTION

Word alignment is a crucial step in statistical machine translation's preprocessing phase. Since the IBM models were first presented, a great deal of statistical work has gone into the development of new word alignment techniques.

When there is a sufficient amount of accessible training data, statistical machine translation (Brown et al., 1993) is capable of producing high-quality results, similar to the outcomes of many other statistical natural language processing tasks. This creates a difficulty for language pairings that are considered to have "low density" since they do not have extremely big parallel corpora. For instance, parameter estimates for word-level alignments may be erroneous when terms appear rarely in a parallel corpus. This, in turn, can lead to inaccurate phrase translation. While a limited quantity of training data might lead to a poor coverage issue, this problem can be made worse by the fact that many words that are encountered during runtime are not noticed in the training data, and as a result, their translations will not be learnt.

The use of various heuristics in order to match target and source words is a concept that is repeated often in the work that is linked to this one. In contrast, we approach the issue of parallel phrase extraction as a classification task and employ feature extraction on the training data to train an SVM classifier to differentiate between parallel and non-parallel phrases. This allows us to extract parallel phrases from non-parallel phrases.

Our technique is totally automated and takes more or less the form of a "create and test" methodology. During the generate phases, given source and target language sentences S and T, we first generate all possible phrases of a given length for S and for T. After that, we compute all possible phrase pairings consisting of one phrase from S and one phrase from T. This process continues until all possible phrase pairings have been generated. During the testing phase, we make use of a binary support vector machine (SVM) classifier to assess, for each produced phrase pair, whether it is parallel or not. In order to train the SVM classifier, phrase pairs are collected from parallel data and aligned using Giza++.

After that, we will go to the section on the literature to learn more about the previous study that has been conducted in this specific area. We will investigate a variety of different options in order to extract lexicons, fragments, and whole sentences. A search of the relevant literature will provide us with information on the number of languages and domains that have been the subject of machine translation as well as the translation methods that have been applied to them.

Over the course of the last several decades, statistical methodologies have made considerable strides in advancing the development of machine translation (MT). Unfortunately, the usability of these approaches is directly dependent on the availability of very huge volumes of parallel corpus data. This is a need for the methods to be useful. Recent research has shown that the lack of parallel corpora may be made up for by using a comparable corpus as a substitute.

In this study, we offer an extraction strategy that was applied to similar corpora. This corpus is then used to adapt and enhance machine translation systems that suffer from a lack of re-sources, and it is utilised to do so by using the corpus. In addition to this, we would like to introduce you to the Euronews-LIUM coalition that was established as a result of our work on the French DEPART initiative. The use of data that is both multimodal and multilingual for the sake of machine translation is one of its primary goals.

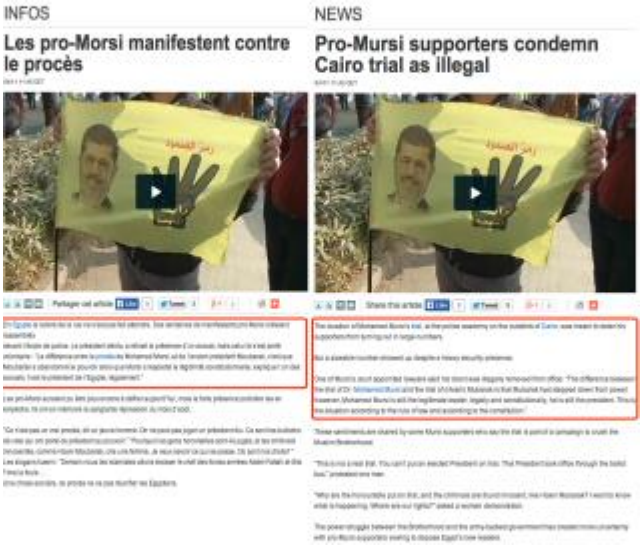
The strategies for enhancing the quality of translation that are offered in this body of work are dependent on multimodal comparable corpora. This refers to several corpora in various modalities that cover the same overarching subjects and occurrences. We evaluate it in comparison to a technique developed by Afli et al. (2013) for the same category of data.

The primary experimental framework that we have developed is intended to investigate two distinct scenarios. The first scenario is one in which we translate data from a new domain, which is distinct from the data used for training. When this occurs, the quality of the translation that is produced is often rather low. The second scenario is one in which we attempt to enhance the performance of an SMT system that has previously been trained on the same variety of data (same domain and or style). The data is taken from the various forms of news (video and text) that are presented on the Euronews website. Both the construction of our TED multimodal comparative corpus and the testing of our extraction algorithms made use of the TED-LIUM corpus that was published by Rousseau et al. in 2012.

The following is the structure of this paper: The new corpora are described in detail in the first two sections. the overarching framework for the generic extraction system.

Multimodal comparable corpora

Euronews



INFOS

Les pro-Morsi manifestent contre le procès

NEWS

Pro-Mursi supporters condemn Cairo trial as illegal

Figure 1: Example of multimodal comparable corpora fromthe *Euronews* website.

The data shown in Figure 1 are examples of multimodal comparable data that were taken from the Euronews website. Together with the similar news in French comes an audio source of a political news story as well as its written form, both of which are accessible in the English language (audio and text modalities). The spoken word in each language may not be translated in the same way, but the films all cover the same ground and discuss the same topics. If this is the case, the data collected from the audio in one language and the written content in the other language may be regarded as equivalent. Extraction of parallel data, both at the sentence and the sub-sentential level, is possible with the help of this corpus.

The news is organised on the Euronews website into a number of distinct categories and subdomains (e.g. Sport, Politics, etc.). These categories are maintained in the unprocessed version of the corpus that was given (but not in extracted versions). The data of our English/French Euronews-LIUM corpus, which was developed from are shown in Table 1.

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

	Audio En		Text	
Sub-Domain	# words	# sentences	# words Fr	# words En
Business	289909	7898	425001	613684
Sport	81768	2369	112736	102923
Culture	388548	16773	262745	274323
Europe	398675	12531	302665	287178
Life Style	28813	1111	18379	19480

Politics	806607	26002	4932055	4666655
Science	231034	9346	147195	141652
Total	2225354	76030	6213995	6127565

Information pertaining to the French and English press during the years 2010 to 2012. This corpus is made up of another corpus that is equivalent to it, and it is made up of transcriptions (which were done using our ASR system, see Section 3.2.) and article content (text found on the webpage). Furthermore supplied are the extracted data that were accomplished using the system that was explained in Section 3.

TED

Over the course of the IWSLT'11 assessment campaign, the TED-LIUM corpus was developed. It has been constructed from many video speeches that were scraped from the website of the TED (Technology, Entertainment, and Design) conference. The corpus contains 773 talks totaling 118 hours of speech and is comprised of those speeches. We created the TED multimodal similar corpus, also known as TED-LIUM, by combining the English audio portion of this corpus with the French text portion of the WIT 3 parallel corpus. Both of these portions are in English. Figure 2

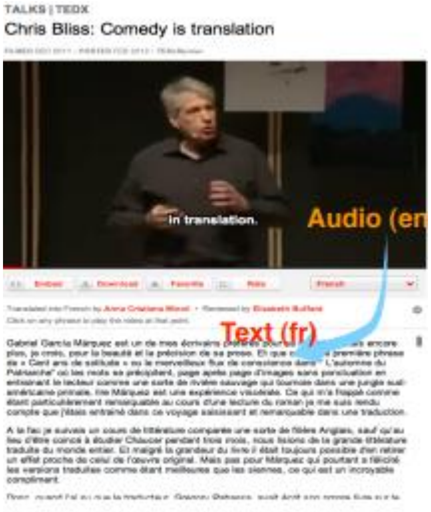


Figure 2: TED's website is one example of multimodal comparable data.

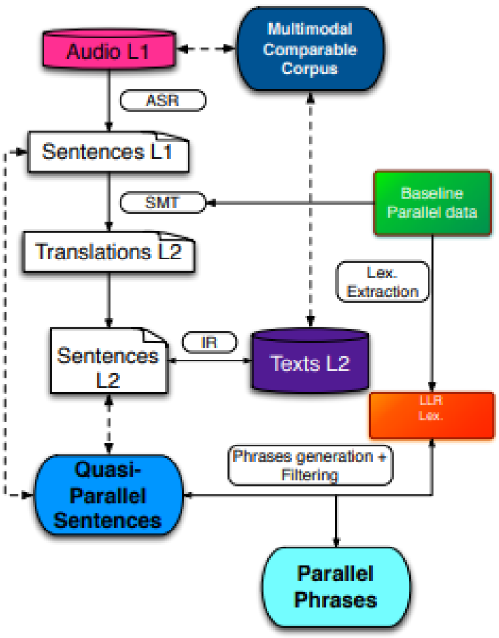


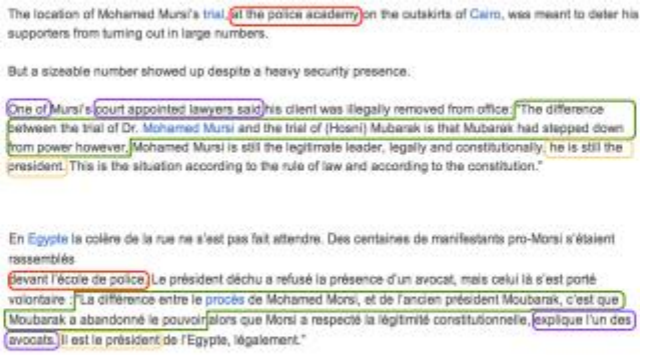
Figure 3: Parallel data extraction from multimodally similar corpora: an overview of the underlying principles.

Parallel data extraction

System Architecture

The primary components of the system architecture are shown in Figure 3. At the beginning, we begin by using the SentExtract methodology that was outlined in (Afli et al., 2012) in order to extract terms that are comparable to those that they currently have. This system is comprised of three separate operations: automatic speech recognition (also known as ASR), statistical machine translation (also known as SMT), and data retrieval (IR). The audio information is spoken in L1, and the system produces text in that language. The automatic speech recognition (ASR) system. After this, a conventional SMT system will do the translation into L2. After this, the translated sentences are entered as queries into an IR system in order to locate the piece of the multimodal comparable corpus's indexed text that is the most similar to the ones that were translated. The transcribed text in L1 and the IR result in L2 are combined to form the sentences that are comparable to one another.

Nevertheless, the quality of the extracted words varies widely, which makes it necessary to undertake a filtering phase in order to guarantee that the functionality of the baseline system is not jeopardised. In the past, many different kinds of strategies have been used to filter the extracted sentences, such as the Term Equivalence Rule (TER) between the IR query and the returned phrase (Abdul-Rauf and Schwenk, 2011) and the bilingual lexicon log-likelihood ratio (LLR). These are just two examples of the many different kinds of strategies that have been used (Snover et al., 2006). (Munteanu and Marcu, 2006). There is a possibility that filtering won't include numerous sentences, which in most cases won't have much of an impact on the original system. This is one of the possible drawbacks of filtering. It is also quite interesting to try to salvage useful parallel sections from sentences that have been eliminated.



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."

Figure 4 an example from the Euronews website with similar sentences with parallel wording

Consider, for instance, figure 4, which has two excerpts from the news stories shown in figure 1. These excerpts are offered here as an example. Despite the fact that both pieces report on the same occurrence and contain ideas that are similar to one another, they cannot be deemed to be parallel writings. They don't have any pairs of sentences that are exactly parallel to one another, but as the boxes in the image demonstrate, they do have some parallel phrases at the sub-sentential level.

A parallel phrase extraction method that runs in two phases was created by our company. Initially, candidates for parallel phrase pairs are identified with the help of the IBM1 model (Brown et al., 1993). After that, the candidates are filtered using a probabilistic translation lexicon (which was learnt on the baseline SMT system training data), and the log-likelihood ratio (LLR) approach is used to construct parallel phrases. Our approach is comparable to that of a programme known as PhrExtract; however, we eliminate the need for TER filtering by use an LLR lexicon instead. PhrExtract LLR is the name that we have given to this new enhanced system.

Baseline systems

The automatic speech recognition (ASR) system that was used in our studies was an in-house five-pass system that was modelled after the LIUM'08 French ASR system described in. This system was based on the open-source CMU Sphinx system (versions 3 and 4). The audio models were trained in the same way, with the exception that we included a multi-layer perceptron (MLP) by using the Bottle-Neck feature extraction method.

Table 2: MT training and development data.

Corpus	# words En	# words Fr
nc7	3.1M	3.7M
eparl7	51.2M	56.4M
devEuronews	74k	84k
tstEuronews	61k	70k
devTED	36k	38k

tstTED	8.7k	9.1k

We made use of the SRILM tools in order to train the language models (LM). We trained a 4-gram LM on all of our corpus of text in a single language.

The Moses SMT toolkit is the foundation for the phrase-based SMT system, which was developed by Moses. The typical set of fourteen feature functions, including 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, are employed in this translation. The following is how it is put together: Initially, the multi-threaded version of the GIZA++ programme is used to do calculations on word alignments in both directions (Gao and Vogel, 2008). The Moses toolkit's default parameters are used in the process of extracting phrases as well as lexical reorderings. Our system's parameters were optimised by utilising the MERT tool on a development corpus to make adjustments (Och, 2003). Table 2 contains the information that we utilised to train, optimise, and evaluate our baseline machine translation system.

We picked the most relevant development and test corpora for each comparable corpus, including Euronews-LIUM and TED-LIUM. The news corpora that were used in the WMT'10 and WMT'11 assessment campaigns, respectively, are known as devEuronews and tstEuronews. The official development and testing corpora for the IWSLT'11 worldwide evaluation campaign are known as devTED and tstTED respectively.

Ogilvie and Callan's Lemur IR toolbox (2001) is what we utilise for the operation of phrase extraction, and we leave the parameters as they were originally designed. Then, we index the whole of the French text, treating each phrase as a separate document. Because of this, it is possible to utilise the translated phrases in the IR toolkit as queries. The information retrieval system uses a bag of words to represent each phrase and then returns the sentences that are most comparable to the query. The English question phrase is then linked with this sentence to complete the translation. By using these methods, we are able to get the French phrases from the similar corpus that have the best possible matches.

Results

For the purpose of making a comparison, we conducted a number of studies using two different approaches. The first one is named PhrExtract LLR, and it was presented in section 3. The second one is a technique that was used by (Afli et al., 2013). (called PhrExtract as in their paper). The challenges of English to French TED and Euronews translation were the subjects of certain experiments.

PhrExtract makes use of TER in order to filter the results that are produced by IR. It does this by retaining only those phrases that have a TER score that is lower than an experimentally set threshold. As a result, we apply various TER thresholds to the chosen phrases in each condition, ranging from 0 to 100 in 10-step increments, and then we filter the results. The BLEU score is used in order to make judgements on the different SMT systems.

Table 3: Total number of TED-LIUMcorpus words and phrases

Methods	# words (en)	# words (fr)

PhrExtract (TER 60)	16.61M	13.82M
PhrExtract_LLR	1.68M	2.27M

Our goal in the experiment using TED data is to modify our baseline SMT system so that it may be used to a new field. As can be seen in table 5, the new system that we have developed produces results that are comparable to those produced by the PhrExtract approach. This indicates that the extracted texts may be used effectively for the aim of adaption.

The similar pattern of behaviour may be seen on the Euronews job (Table 6). An existing SMT system that has previously been trained on the same sort of data may be improved with the text that was extracted from it.

This new extraction approach eliminates the need to apply the TER filtering, which necessitated a great deal of trial and error in order to determine the optimal threshold for each specific activity.

In addition, by examining the widths of the extracted text in Tables 3 and 4, we can see that the LLR technique generates a great deal less data while yet achieving the same level of speed. This suggests that just the data that is the most relevant to the question at hand are retrieved using this method.

As the example in Table 7 demonstrates for us, the incorporation of the extracted phrases may have a beneficial impact on the overall quality of the translation.

Table 4: The total number of words and phrases retrieved using the PhrExtract and PhrEx- tract LLR techniques from the Euronews-LIUMcorpus.

Methods	# words (en)	# words (fr)
PhrExtract (TER 50)	2.39M	1.95M
PhrExtract_LLR	636.8k	224.1k

The statistics of the bitexts that were taken from Euronews-LIUMand TED-LIUMare shown in Tables 3 and 4, respectively. It is possible to see that the sizes of the two sides of the multilingual text taken from Euronews-LIUM are very different from one another (English side is al- most three times larger than French size). This behaviour is not seen on the TED data, and we do not yet have an explanation for this finding, which needs a more fine grain investigation of the bitexts that were collected. In order to customise the default MT system, we have inserted these bitexts into our generic training data.

The BLEU scores that were achieved with the best bitext retrieved from each multimodal corpus using the PhrExtract and PhrExtract LLR algorithms are shown in Tables 5 and 6, respectively. The TER threshold for Euronews-LIUM is set at 50, while the TER barrier for TED-LIUM is set at 60.

Table 5: BLEU scores on devTED and tstTED after adapta-tion of a baseline system with bitexts

extracted from TED-LIUMcorpus.

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

Table 6: BLEU scores on devEuronews and tstEuronews after adaptation of a baseline system with bitexts extractedfrom Euronews-LIUMcorpus.

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

CONCLUSION

We have introduced a whole new multimodal corpus that was developed with the intention of extracting parallel data for SMT systems. In addition to this, we presented a novel method for the extraction of parallel fragments from a multimodal comparable corpus. Experiments that were run 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. This was demonstrated by the fact that our method significantly outperformed the existing approaches (Euronews). This is a promising conclusion, which, in contrast to the TER filtering approach, does not need any threshold to be experimentally set. Our strategy has room for development in a number of key areas. The LLR lex- icon that is used for filtering is generated with the utilisation of a parallel corpus. Constructing a sizable bilingual vocabulary from similar content would be another approach that might be used. This dictionary would then be used by the filtering module. In this scenario, the lexicon would be more useful if it had terms that were particular to the activity that was being targeted (in the case of adaptation). The careful selection of similar data to be utilised in the extraction framework is yet another interesting improvement that might be included. This decision might be made using a similarity measure that was calculated before to the extraction procedure. Doing so would contribute to an improvement in the system's performance.

REFERENCES

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

[2]. H. Afli, L. Barrault, and H. Schwenk. 2012. Parallel texts extraction from multimodal comparable corpora. In *Jap-TAL*, volume 7614 of *Lecture Notes in Computer Sci- ence*, pages 40–51. Springer.

[3]. Haithem Afli, Loïc Barrault, and Holger Schwenk. 2013. Multimodal comparable corpora as resources for extract-ing parallel data: Parallel phrases extraction. *Interna- tional Joint Conference on Natural Language Process- ing*, October.

[4]. Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Comput. Linguist.*, 19:263–311, June.

[5]. Mauro Cettolo, Marcello Federico, and Nicola Bertoldi. 2010. Mining parallel fragments from comparable texts.*Proceedings of the 7th International Workshop on Spo- ken Language Translation*.

[6]. P. Deléglise, Y. Estève, S. Meignier, and T. Merlin. 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 (UnitedKingdom), 6-10 september.

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

[8]. F. Grézl and P. Fousek. 2008. Optimizing bottle-neck fea- tures for LVCSR. In *2008 IEEE International Confer- ence on Acoustics, Speech, and Signal Processing*, pages 4729–4732. IEEE Signal Processing Society.

[9]. P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Fed-erico, N. Bertoldi, B. Cowan, W. Shen, C. Moran,

[10]. R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: open source toolkit for statistical machinetranslation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Ses-sions*, ACL '07, pages 177–180.