Parallel Data Extraction Using Word Embeddings

Pintu Lohar and Andy Way

ADAPT Centre, Dublin City University, Ireland

Abstract. Building a robust MT system requires a sufficiently large parallel corpus to be available as training data. In this paper, we propose to automatically extract parallel sentences from comparable corpora without using any MT system or even any parallel corpus at all. Instead, we use crosslingual information retrieval (CLIR), average word embeddings, text similarity and a bilingual dictionary, thus saving a significant amount of time and effort as no MT system is involved in this process. We conduct experiments on two different kinds of data: (i) formal texts from news domain, and (ii) user-generated content (UGC) from hotel reviews. The automatically extracted sentence pairs are then added to the already available parallel training data and the extended translation models are built from the concatenated data sets. Finally, we compare the performance of our new extended models against the baseline models built from the available data. The experimental evaluation reveals that our proposed approach is capable of improving the translation outputs for both the formal texts and UGC.

Keywords: Machine Translation, parallel data, user-generated content, word embeddings, text similarity, comparable corpora

1 Introduction

A parallel corpus is the main ingredient for building an MT system. Usually, there are two ways of parallel corpus acquisition, namely: (i) manual development, and (ii) automatic extraction. Although manual development is ideal and is produced in most cases by human translators, this process requires a huge amount of time and effort which is considered to be less practical than automatic extraction of parallel data for MT. One of the easiest ways to accomplish this task is to employ an MT system that translates all the source-language texts into the target language and then performs text similarity in the target language. However, using an MT system is not always the best solution mainly due to the following reasons: (i) it requires a significant amount of time to build the MT system itself, especially if this is an NMT system, (ii) it also takes a long time to translate all the source-language documents into the target language especially for large corpora, and (iii) MT systems for all domains and language pairs are not available. These problems demonstrate that finding a suitable alternative to using an MT system for parallel data extraction is an important aim. In this work, we propose to combine the CLIR, text similarity and word embedding-based approach for extracting parallel sentences from the comparable corpora for both formal texts and UGC, without the help of any MT system or any parallel corpus, thereby saving a significant amount of time

David C. Wyld et al. (Eds): ACITY, DPPR, VLSI, WeST, DSA, CNDC, IoTE, AIAA, NLPTA - 2020
pp. 251-267, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101521

and effort. We use the We use the *Euronews* corpus and hotel reviews (discussed in detail in Section 3) as the comparable corpora for parallel data extraction. We conduct our experiments on English and French texts from these corpora. We consider French as the source- and English as the target language in our experiments. As the CLIR-based searching works at document level, we represent each sentence as a document. Initially, we use the CLIR- component of *FaDA* [17] to index all the source and target language documents and then find a set of suitable candidate target-language documents for each source-language document. Afterwards, we translate[1] all the content words (i.e, after removing stopwords) of the French documents using a French-to-English dictionary.[2] Each of the extracted candidate English documents is then compared with the French document using the average word embeddings of the content words of each English document and that of the English translations of the words in the French document. The word embedding-based similarity is also accompanied by text similarity. The English document with the highest similarity score is selected as the parallel counterpart of the French document.

The remainder of this paper is organised as follows. In Section 2, we discuss some of the existing relevant works in this field. The description of the data sets we use in this work is provided in Section 3. In Section 5, we describe the experimental setup which is followed by the results obtained in Section 6. We perform output analysis in Section 7. Finally, we conclude our work and point out some future possibilities in Section 8.

2 Related work

The extraction of parallel sentences/segments plays an important role in improving MT quality [22, 13]. In general, the issue of parallel data extraction is addressed in different ways. For example, [16] propose a crowdsourcing approach for extracting parallel data from tweets. They attempt to find the translations in tweets instead of translating the texts. [8] extract both parallel sentences and fragments from comparable corpora of Chinese–Japanese Wikipedia to improve statistical MT. [12] apply a domain-biased parallel data collection and a structured methodology to obtain English–Hindi parallel data. Deep learning has gained popularity in this task [5, 11] recently. Many work exploits MT for parallel data extraction [7, 19]. As the alternative resources to parallel data, the comparable corpora are considered as valuable resources for MT. For example, [1] use a multimodal comparable corpus

[1] Note that this is merely a word-to-word translation, not a generic MT

[2] The dictionary is available at: `www.seas.upenn.edu/~nlp/resources/TACL-data-release/dictionaries.tar.gz`

of audio and texts built from 'Euronews'[3] and 'TED'[4] web sites for parallel data extraction. [14] propose a bidirectional method to extract parallel sentences from English and Persian document-aligned Wikipedia. They use two MT systems to translate from Persian to English and the reverse after which an IR system is used for measuring the similarity of the translated sentences. Although many parallel data extraction systems employ MT, it is not always a good idea and so we simply discard the requirement of any MT system and any parallel data at all.

3 Data set

We use two different types of data sets in our experiments: (i) formal text corpora from news domain, and (ii) UGC corpora of reviews.

3.1 Formal text corpora from news domain

The formal text corpora consist of the *Euronews* and the *News commentary* corpus.

- **Euronews corpus:** The *Euronews* corpus [2] is a multimodal corpus of comparable documents and their images. In our experiments, we consider only the documents and not the images as this is beyond the scope of this work. Each document in *Euronews* corpus consists of at least one line of text and many of them contain multiple-line texts with multiple sentences.
- **News commentary corpus:** This data set is comprised of the English–French parallel sentence pairs from the 'News-Commentary' corpus.[5] We refer to this data set as *NewsComm* in short.

Data set	**Language**	**# Documents**	**# Sentences**
Euronews	English	40,421	644,226
	French	37,293	614,928
NewsComm	English	/	246,946
	French	/	246,946

Table 1: Data statistics

Table 1 shows the statistics of the *Euronews* and the *NewsComm* data. We already mentioned earlier in Section 1 that we split each document into multiple sentences in this work. We can see in the above table that in the *Euronews* data, $644K$ English and $614K$ French sentences are obtained from $40K$ English and $37K$ French

[3] https://www.euronews.com/

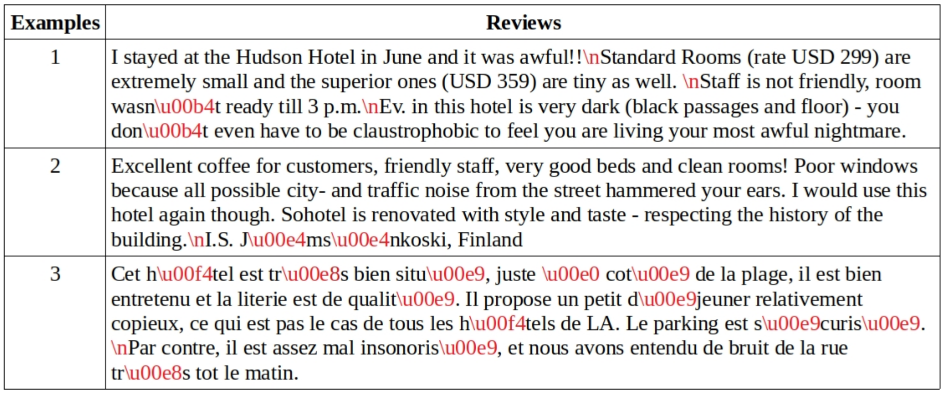
[4] https://www.ted.com/

[5] http://www.casmacat.eu/corpus/news-commentary.html

documents, respectively. Note that the *NewsComm* data set is simply a parallel corpus at sentence level, not any document level, and so the third column entries are replaced by the '/' character which means 'not applicable' in this case.

3.2 UGC corpora of reviews

- **FourSquare parallel corpus:** This data set contains over $11K$ reviews (or $18K$ sentences) from the French–English parallel corpus of Foursquare restaurant reviews[6] [3]. The reviews were originally written in French, which were then translated into English by the professional translators. The authors also provide the official training, development and test splits for this data set.
- **Hotel review corpus:** The *Hotel_Review* corpus[7] consists of $878K$ reviews from $4,333$ hotels crawled from *TripAdvisor*. Although most of the reviews are in English, some of them are also written in French. Table 2 shows randomly selected three example reviews (two English and one French) from this data set. We highlight the special characters such as newlines, unicodes in red.



Examples	Reviews
1	I stayed at the Hudson Hotel in June and it was awful!!\nStandard Rooms (rate USD 299) are extremely small and the superior ones (USD 359) are tiny as well. \nStaff is not friendly, room wasn\u00b4t ready till 3 p.m.\nEv. in this hotel is very dark (black passages and floor) - you don\u00b4t even have to be claustrophobic to feel you are living your most awful nightmare.
2	Excellent coffee for customers, friendly staff, very good beds and clean rooms! Poor windows because all possible city- and traffic noise from the street hammered your ears. I would use this hotel again though. Sohotel is renovated with style and taste - respecting the history of the building.\nI.S. J\u00e4ms\u00e4nkoski, Finland
3	Cet h\u00f4tel est tr\u00e8s bien situ\u00e9, juste \u00e0 cot\u00e9 de la plage, il est bien entretenu et la literie est de qualit\u00e9. Il propose un petit d\u00e9jeuner relativement copieux, ce qui est pas le cas de tous les h\u00f4tels de LA. Le parking est s\u00e9curis\u00e9. \nPar contre, il est assez mal insonoris\u00e9, et nous avons entendu de bruit de la rue tr\u00e8s tot le matin.

Table 2: Review examples

Note that the newline characters are not always explicitly present even if a new sentence starts. For instance, in example 2, there are no newline characters before the sentences such as '*Poor windows....*' and '*I would use this....*'. In addition, a plenty of unicode characters are present in the hexcode format such as '00b4', '00e9', '00e8' etc. most of which are present in the French review in example 3. Considering these observations, we preprocess the data using the following steps.

[6] https://europe.naverlabs.com/research/natural-language-processing/machine-translation-of-restaurant-reviews/

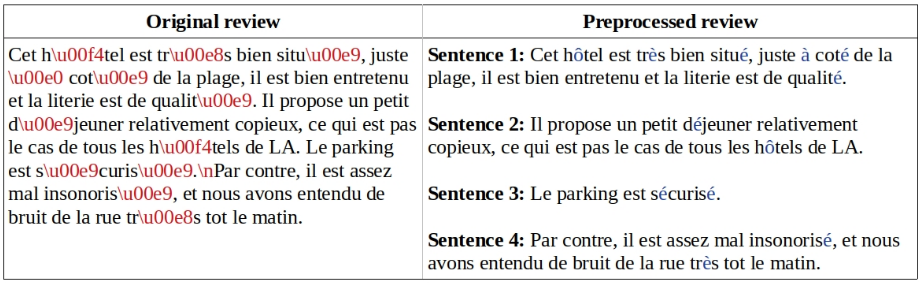
[7] https://www.cs.cmu.edu/~jiweil/html/hotel-review.html

(i) **Language detection:** We perform language detection[8] in order to detect and extract the English and French reviews from this data set.

(ii) **Sentence splitting:** As our parallel data extraction system is implemented at sentence level, we split the multi-sentence reviews into different parts (sentence) and consider each part as a single document.

(iii) **Unicode conversion:** We convert[9] the characters given in unicode format into the Latin characters. For example, the character '00f4' is converted into 'ô'.

Table 3 shows an original French review (example 3 of Table 2) and its preprocessed version. We highlight all the unicodes in the original review in red and the converted characters in the preprocessed review in blue. Note that 4 sentences are generated from this single review after preprocessing.



Original review	**Preprocessed review**
Cet h\u00f4tel est tr\u00e8s bien situ\u00e9, juste \u00e0 cot\u00e9 de la plage, il est bien entretenu et la literie est de qualit\u00e9. Il propose un petit d\u00e9jeuner relativement copieux, ce qui est pas le cas de tous les h\u00f4tels de LA. Le parking est s\u00e9curis\u00e9.\nPar contre, il est assez mal insonoris\u00e9, et nous avons entendu de bruit de la rue tr\u00e8s tot le matin.	**Sentence 1:** Cet hôtel est très bien situé, juste à coté de la plage, il est bien entretenu et la literie est de qualité. **Sentence 2:** Il propose un petit déjeuner relativement copieux, ce qui est pas le cas de tous les hôtels de LA. **Sentence 3:** Le parking est sécurisé. **Sentence 4:** Par contre, il est assez mal insonorisé, et nous avons entendu de bruit de la rue très tot le matin.

Table 3: An example review before and after preprocessing

The statistics of the *FourSquare* and *Hotel_Review* data sets is shown in Table 4.

Data set	**# Reviews**	**# Total sentences**	**# training**	**# Dev**	**# Test**
FourSquare	11,551	17,945	14,864	1,243	1,838
Hotel_Review	878,561	/	/	/	/

Table 4: Statistics of the FourSquare parallel and the Hotel review data sets

[8] https://pypi.org/project/langdetect/

[9] Unicode representation of these characters can be found at: http://www.fileformat.info/info/unicode/char/search.htm

4 System description

Our proposed system is composed of the following components: (i) CLIR-based system, (ii) sentence length-based pruning, (iii) average word embeddings, (iv) text similarity, and (v) score combination.

4.1 CLIR-based system

The CLIR component used in this experiment is a part of the open source bilingual document alignment tool *FaDA* [17]. It works in the following steps:

(i) firstly, the source-language and the target-language documents are indexed,

(ii) each source-language document is used to construct a pseudo-query[10] which is considered as the suitably representative of the document,

(iii) all pseudo-query terms are translated into the target-language by a bilingual dictionary and the translated query terms are then searched in the target-language index, and finally

(iv) the *top-n*[11] target-language documents are retrieved.

4.2 Sentence length-based pruning

Prior to performing the word embedding- and the text-based similarities between the source- and the target-language sentences, we exclude some of the comparisons depending upon the sentence-length ratio. This ratio is calculated in terms of the total number of words in the word translations of the source-language document (sentence) and the total number of words in the target-language document (sentence). We set the threshold for this ratio to 0.5, which means that the shorter of the document pair must be at least the half of the longer document in terms of the total number of words they contain. For example, if a French document contains 5 words and an English document contains 20 words, the ratio is 0.25 which is less than the threshold of 0.5. This document pair, therefore, according to our criteria is less likely to be parallel and so is not considered for comparison. The French document must contain at least 10 words to pass this threshold in order to be considered for further similarity measurements. However, 0.5 is not an empirically determined threshold; we choose this value so that very unlikely candidates can be removed from the comparison, albeit some of the invalid pairs still pass the threshold.

In general, the average length ratio of English texts over the French translations is near 1.0 [6] but there are many examples that violate this. For example, consider the English sentence '*I like to propose a toast.*' that contains 6 words and

[10] A *pseudo-query* is the modified form a user's original query in order to improve the ranking of retrieval results compared to the original query.

[11] We use the default value of n used in *FaDA*, where $n = 10$, which means the top 10 candidate target-language documents are retrieved.

its equivalent French translation '*J'aime proposer un toast*' that contains 4 words. The sentence-length ratio in this case is below 0.7 which is far less than 1.0. Therefore, setting a high threshold very close to 1.0 can result in discarding many valid sentence pairs like this one.

4.3 Text similarity

We calculate the text similarity using the following steps:

(i) firstly, we remove all the stopwords from both the French and English documents.

(ii) secondly, we translate the remaining content words of the French document into English using a French–English bilingual dictionary.

(iii) some of the word translations contain stopwords such as *to*, *of* etc. We remove these stopwords.

(iv) finally, we calculate the total number of word matches between the words in the English document and the word-level English translation of the French document.

4.4 Average word vector similarity

Consider the Figure 1 that shows a collection of words represented in a two dimensional space. We can observe that the semantically equivalent words are placed in close proximity. For example, the words *electrical*, *electricity*, *electric* etc. are closely grouped together in the same region. However, this figure shows the simplest representation of how the related words are treated. In reality, the words are represented as a vector of real values in much higher dimensions. In pre-trained word embeddings, the semantically related words usually contain similar vector values.

We now discuss how the average word vectors are actually calculated. Let us consider a sentence S with a sequence of n words: $w_1, w_2, w_3.....w_n$. Let the vector embeddings of the words be $u_{w_1}, u_{w_2}, u_{w_3}.....u_{w_n}$. The average word embedding of S is calculated using the Equation (1) as follows:

$$U_s = \frac{1}{n}\sum_{i=1}^{n} u_{w_n} \tag{1}$$

In our experiments, we use the '*fasttext*' pre-trained Wiki word vectors for English which is made available by [4]. In order to obtain the word embeddings for our experiments, we apply the following steps:

(i) All the stopwords in both the word translations of the French document and the English document are removed,

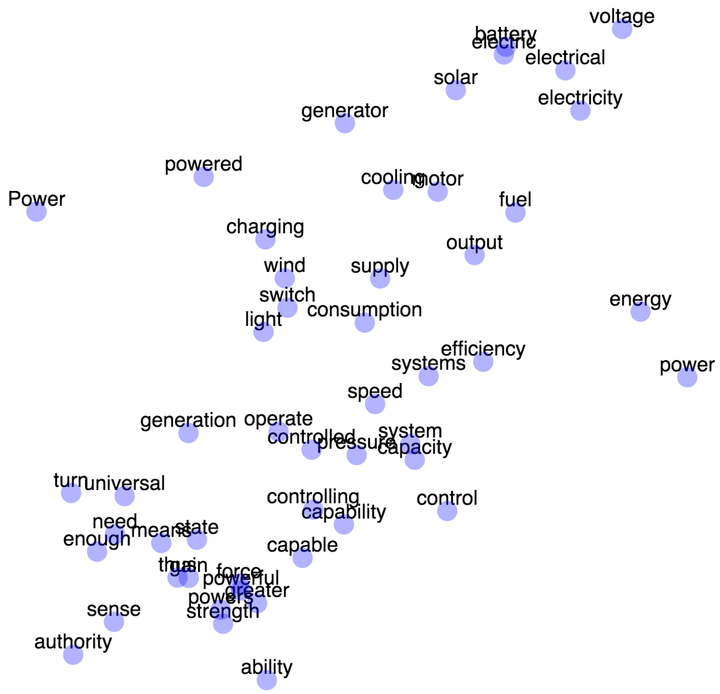


Fig. 1: Example of semantically related words in two dimensional space

(ii) the real word vector values of all the remaining words in the word translations of the French document are retrieved and then the average of all these vector values is calculated,

(iii) the average word vector values for the English document is calculated in a similar manner, and

(iv) the two averages are compared in order to calculate the average word vector similarity.

4.5 Score combination

Once we calculate the text and the average word vector similarities, these scores are then combined to obtain the overall similarity score. The overall similarity score S_{sim} is calculated using the Equation (2) as follows:

$$S_{sim} = w_1 WV_{sim} + w_2 Text_{sim} \tag{2}$$

In the above equation, WV_{sim} and $Text_{sim}$ are the average word vector and the text similarity scores with w_1 and w_2 weight values, respectively.

5 Experiments

5.1 MT configuration

The MT models are built using the freely available open source NMT toolkit 'OpenNMT'[12] [15]. We consider French as the source and English as the target language. In our experiments, we use all the default parameter settings: *RNN* as the default type of encoder and decoder, $word\_vec\_size = 500$, $rnn\_size = 500$, $rnn\_type = LSTM$, $global\_attention\_function = softmax$, $save\_checkpoint\_steps = 5000$, $training\_steps = 100,000$ etc. We evaluate the translation quality using BLEU [18].

5.2 Sentence-level document alignment

We store each sentence of the *Euronews* corpus in a single document which results in creating more than $600K$ documents per language. These documents are then fed as input to the CLIR component of *FaDA*. Once the top n English documents are obtained for a French document, we aim to find its closest semantically equivalent English document. It is, therefore, expected that the total number of extracted sentence pairs is over $600K$. However, it is impractical to consider all these sentence pairs as parallel data because many of them are not semantically equivalent. We, therefore, extract only those pairs that have the similarity score greater than a threshold (discussed in detail in Section 5.4). Table 5 shows the data size of the existing parallel corpora and the extracted sentence pairs from the *Euronews* and *Hotel_Review* data sets.

Text type	**Data set**	**# Sentence**
Formal text	NewsComm parallel Extracted sentence pairs (Euronews)	246,946 31,860
UGC text	FourSquare parallel Extracted sentence pairs (Hotel_Review)	14,864 6,188

Table 5: Existing parallel corpora vs extracted sentence pairs

5.3 Translation models

Note that we show data combinations for two different types of data sets: (i) formal text, and (ii) UGC text. We built a baseline translation model and an extended translation model for each of the above types of texts.

[12] https://github.com/OpenNMT/OpenNMT-py

Models for formal text corpora The extracted parallel sentences from the *Euronews* corpus are used as the additional data set for MT training for formal texts. We build two translation models: one is the baseline model and another is the extended model. The baseline model is built using only the *NewsComm* data whereas the extended model is built using the concatenated data. We held out $1,000$ sentence pairs for development and another $1,000$ sentence pairs for tuning purposes from the *NewsComm* data. We refer to this baseline model as $Base_{FT}$ and the extended model as Ext_{FT}, where 'FT' stands for 'formal text'. Table 6 shows the data distribution. Each translation model is tuned and tested on the same development and test data sets, respectively.

Model	**Data set**	**# training**	**# Dev**	**# Test**
$Base_{FT}$	News	$226,946$	$1,000$	$1,000$
Ext_{FT}	News + Euronews	$253,592$	$1,000$	$1,000$

Table 6: Data distribution for two different MT models for formal text corpora

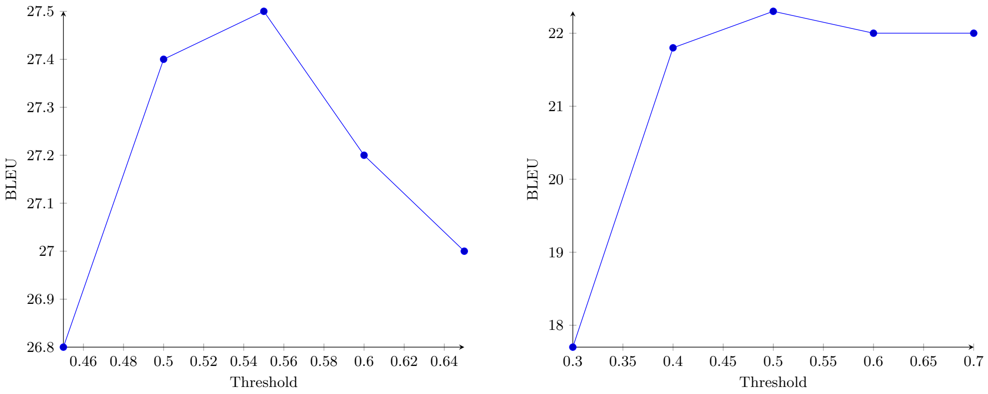
Models for UGC corpora Once the sentence pairs are extracted from the *Hotel_Review* data set, we consider them as the additional parallel resource and concatenate with the parallel training sentences of the *FourSquare* corpus. We build following MT models: (i) a baseline model, which is built from the $14,864$ parallel training sentences of the *FourSquare* corpus, and (ii) an extended model, which is built from the concatenation of the *FourSquare* data and the sentence pairs extracted from the *Hotel_Review* data set. The baseline model is referred to as '$Base_{UGC}$' and the extended model is referred to as 'Ext_{UGC}'. Table 7 shows the data distribution. Both translation models are tuned and tested on the same development and test data sets, respectively.

Model	**Data set**	**# training**	**# Dev**	**# Test**
$Base_{UGC}$	FourSquare	$14,864$	$1,243$	$1,838$
Ext_{UGC}	FourSquare + Hotel review	$21,052$	$1,243$	$1,838$

Table 7: Data distribution for two different MT models for UGC text corpora

5.4 System tuning

As we discussed earlier in Section 4, we calculate the overall similarity score of a sentence pair using Equation (2). However, it is required to obtain a threshold for the the similarity score above which all the sentence pair can be considered as parallel sentences. We explored different threshold values for both the *Euronews* and the *Hotel_Review* data sets.



(a) Formal text: *Euronews* + *NewsComm*

(b) UGC text: *Hotel_Review* + *FourSquare*

Fig. 2: Tuning threshold value with BLEU scores

We extract a set of parallel sentence pairs for each threshold value. Aftwerwards, each set is added to the existing parallel resources to build extended translation models. For example, if we set the threshold for similarity score to 0.5, all the sentence pairs whose similarity score are higher than 0.5 can be considered as parallel sentences and would be added to the existing training data. Using this method, different sets of such concatenated data are obtained using different threshold values. We then build different translation models using each data set separately. The model for which we obtain the highest BLEU score is considered as the extended model and the corresponding threshold is considered as the optimal threshold. Figure 2a and 2b show the BLEU score comparison with different threshold values used for extracting sentence pairs from the *Euronews* and *Hotel_Review* data sets, respectively and adding them to the parallel sentences from *NewsComm* and *FourSquare* data sets, respectively. Note that the former combination belongs to formal text and the later one belongs to UGC text. It is obvious from Figure 2a that the BLEU score decreases as the threshold is reduced or increased from 0.55

for the formal text corpus. The highest BLEU score of 27.5 is obtained at this threshold. In contrast, we can observe from Figure 2b that the highest BLEU score of 22.3 is obtained for the UGC text corpus using the similarity threshold of 0.5. We, therefore, set the optimal similarity thresholds for the formal and the UGC text corpora to 0.55 and 0.5, respectively.

6 Results

We show the BLEU scores obtained by the Baselines and the *Extended* models for both the formal text and UGC corpora in Table 8. Note that the formal text corpus is comprised of the parallel sentences from *NewsComm* and extracted sentence pairs from *Euronews* data, whereas the UGC text corpus consists of the parallel sentences from *FourSquare* and extracted sentence pairs from *Hotel_Review* corpus.

Corpus type	Translation model	BLEU score
Formal text	$Base_{FT}$	27.1
	Ext_{FT}	**27.5**
UGC text	$Base_{UGC}$	22.1
	Ext_{UGC}	**22.3**

Table 8: BLEU score comparison

For formal text corpus, we can observe in Table 8 that the addition of parallel sentences extracted from the *Euronews* corpus using our proposed system improves the BLEU score, i.e. the Baseline ($Base_{FT}$) is outperformed by the extended model (Ext_{FT}) by 0.4 BLEU points. On the comparison, we notice a slight improvement in BLEU score for UGC text corpus, i.e. the Baseline ($Base_{UGC}$) is outperformed by the extended model (Ext_{UGC}) by 0.2 BLEU points. We also perform the statistical significance test of the outputs using MultEval [9]. However, we found that these improvements in BLEU score are not statistically significant as $p > 0.1$.

We notice that the improvement in BLEU score for UGC text corpus (i.e. 0.2 points) is less than that for the formal text corpus (i.e. 0.4 points). One probable reason for this degradation is that the *Euronews* data actually contains some parallel texts. Therefore, extracting and adding them to the existing *NewsComm* parallel training data helps improve the BLEU score to some extent. In contrast, the *Hotel_Review* data set does not contain parallel texts. In fact, the reviews are generated randomly by different users without any translation usage in mind. However, there exists some texts that are very close in meaning even though they are not parallel. Due to this reason, such partially semantically similar texts help improve the BLEU score very slightly.

7 Output Analysis

The improvement in BLEU score shows that our automatic parallel data extraction system helps improve MT quality by supplying additional training data. However, this is the beginning phase of our experiment and further plans are made to extend this work. As of now, we illustrate some example outputs where the *Baseline* models are outperformed by our *Extended* models.

Example	Reference	Baseline model	Extended model
1	But, equally important, workers organized themselves to defend their interests.	But, as important, workers have been organized to defend their interests.	But, equally important, workers have been organized to defend their interests.
2	Overall, however, the inequality gaps are large and, in many cases, growing.	Overall, however, the inequality gap remains acute and in some cases even expansion.	Overall, however, the inequality gap remains deep, and in some cases it expands.
3	Countries that import currently subsidized food will be worse off.	Countries that imports currently will suffer.	Countries that import products currently subsidized will suffer.
4	He ate chocolate and watched NBA games.	He ate chocolate and watched from the NBA games.	He ate chocolate and watched the NBA games.

Table 9: Example outputs: Baseline vs Extended model (formal text corpora)

Let us first show some example outputs produced by the translation models built from the formal text corpora (*NewsComm* and *Euronews*) in Table 9 and explain how the *Baseline* model is outperformed by the *Extended* model. In example 1, the word '*equally*' is missing in the *Baseline* output. The second example shows that the ending phrase '*in some cases even expansion*' of the output produced by the *Baseline* model is grammatically incorrect whereas the *Extended* model produces the phrase '*in some cases it expands*' which is grammatically correct and semantically equivalent to the phrase '*in many cases, growing*' in the reference translation.

In example 3, both translation outputs are erroneous but the output produced by the *Extended* model is better as it includes the word '*products*' which although is not equivalent to the word '*food*' in the reference translation but at least conveys a little bit of similar meaning. Finally, example 4 shows the case where both translation outputs are mostly correct except the presence of extra prepositions. The phrase '*watched the NBA games*' that is produced by the *Extended* model is better than the phrase '*watched from the NBA games*' produced by the *Baseline* when compared with the reference translation.

Table 10 illustrates some example outputs produced by the translation models built from the UGC text corpora (*FourSquare* and *Hotel_Review* data sets) and shows how the *Baseline* model is outperformed by the *Extended* model.

Example	Reference	Baseline model	Extended model
1	Cozy little teahouse, amazing sweets and teas.	Disgusting room, very good cakes and teas.	Small tea room, very good cakes and teas.
2	A nice atmosphere to hang out with friends.	Friendly atmosphere for a relaxed dinner.	Friendly atmosphere for a walk with friends.
3	The sales assistants are super friendly.	The sales assistants are really welcoming.	The sales assistants are super welcoming.
4	Their famous hot chocolate, one of the best in the world, is worth the wait!	Its suggestion hot chocolate, one of the best in the world is worth the wait!	Its legendary chocolate, one of the best in the world is worth the wait!
5	They do really good burgers.	They serve very good burgers.	They do very good burgers.

Table 10: Example outputs: Baseline vs Extended model (UGC text corpora)

We can notice in the table that although the phrase '*Small tea room*' (in example 1) produced by the *Extended* model is not a proper translation, it is still much better than the completely wrong translation output '*Disgusting room*' produced by the *Baseline* model. In example 2, the phrase '*hang out with friends*' in the reference translation is semantically closer to the phrase '*walk with friends*' (produced by the *Extended* model) than to the phrase '*relaxed dinner*' (produced by the *Baseline* model). Moreover, '*super friendly*' is more synonymous to '*super welcoming*' than '*really welcoming*' in example 3. Furthermore, the word '*suggestion*' (see example 4) is completely meaningless when used before '*hot chocolate*' that is produced by the *Baseline* model. In contrast, although '*legendary chocolate*' is not a proper translation (produced by the *Extended* model), it is partially similar to '*famous hot chocolate*' in the reference. Finally, both of the translation outputs in example 5 are sensible but the output produced by the *Extended* model is closer to the reference.

8 Conclusions and Future work

In this paper, we proposed a parallel data extraction technique from comparable corpora of both formal texts and UGC in order to generate additional parallel training data for MT. Many research works employ MT itself to ease this task. However, it is not always a practical solution because in addition to building the MT system in the first place, it also requires a huge amount of time to translate all the source-language documents of the comparable corpus into the target-language in order to be able to perform the text similarity in the target language. To overcome this situation, we implemented a parallel data extraction system without any help from MT or even any parallel corpus. We initially used the CLIR component of *FaDA* tool to extract the candidate target-language sentences for a source-language sentence. We then used the average word-embeddings and text similarity with the help of a bilingual dictionary in order to obtain parallel sentences from the *Euronews* and *FourSquare* corpus. These extracted sentence pairs were then concatenated with the existing parallel training data to build the extended translation models which outperformed the baseline systems that are built from only the existing parallel training data.

We noticed that extracting parallel texts from the *Euronews* corpus obtains a slightly higher BLEU score improvement than for the *Hotel_Review* data set. One probable reason is that the *Euronews* data actually contains some parallel texts and so extracting and adding them to the existing *NewsComm* parallel training data helps improve the BLEU score to some extent. In contrast, the hotel reviews are extremely unlikely to contain parallel texts as they are randomly generated by different users without translation usage foreseen. Although not being strictly parallel, some of them are semantically equivalent, and adding them as extra training data improves the BLEU score very slightly over the *Baseline* model. It is, therefore, expected that the BLEU score can be improved further if there exists a considerable amount of parallel texts in a comparable corpus of UGC. As we did not use any MT system or any parallel corpus for this task, our proposed system is very simple and can be easily applied to a large comparable corpus. Our findings in this research are encouraging as our system relies on only the text similarity, word embeddings and a bilingual dictionary, for which the required resources are easily available online. We believe that our proposed model has the potential to benefit further research in this field.

One of the drawbacks in our approach is that we have not compared our system with some of the most popular existing sentence alignment systems. Some examples of well-known works in this field are [20], [10] and [21]. In future, we would like to explore these approaches and apply their sentence alignment systems on the data sets we used in this work. Our main objective is to combine the best performing system with our system. Another possibility is to apply all of them separately and select the sentence alignments that are common outputs generated by all or most of these alignment systems. In addition, we also plan to apply our system to other types of UGC such as tweets, customer feedback, movie reviews etc.

Acknowledgements

This research has been supported by the ADAPT Centre for Digital Content Technology which is funded under the SFI Research Centres Programme (Grant 13/RC/2106).

Bibliography

[1] Haithem Afli, Loïc Barrault, and Holger Schwenk. Building and using multimodal comparable corpora for machine translation. *Natural Language Engineering*, 22, 11 2015.

[2] Haithem Afli, Pintu Lohar, and Andy Way. MultiNews: A web collection of an aligned multimodal and multilingual corpus. In *Proceedings of the First Workshop on Curation and Applications of Parallel and Comparable Corpora*, pages 11–15, Taipei, Taiwan, November 2017.

[3] Alexandre Berard, Ioan Calapodescu, Marc Dymetman, Claude Roux, Jean-Luc Meunier, and Vassilina Nikoulina. Machine translation of restaurant reviews: New corpus for domain adaptation and robustness. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 168–176, Hong Kong, 2019.

[4] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[5] Houda Bouamor and Hassan Sajjad. H2@bucc18: Parallel sentence extraction from comparable corpora using multilingual sentence embeddings. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, Miyazaki, Japan, 2018.

[6] Aitao Chen. Cross-language retrieval experiments at clef 2002. *Advances in Cross-Language Information Retrieval*, pages 28–48, 2003.

[7] Chenhui Chu. *Integrated Parallel Data Extraction from Comparable Corpora for Statistical Machine Translation.* PhD dissertation, Kyoto University, 2015.

[8] Chenhui Chu, Toshiaki Nakazawa, and Sadao Kurohashi. Integrated parallel sentence and fragment extraction from comparable corpora: A case study on chinese–japanese wikipedia. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 15(2), 2015.

[9] Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies:*, pages 176–181, Portland, Oregon, USA, 2011.

[10] Luís Gomes and Gabriel Pereira Lopes. First steps towards coverage-based sentence alignment. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2228–2231, Portorož, Slovenia, May 2016.

[11] Francis Grégoire and Philippe Langlais. Extracting parallel sentences with bidirectional recurrent neural networks to improve machine translation. In

Proceedings of the 27th International Conference on Computational Linguistics, pages 1442–1453, Santa Fe, New Mexico, USA, 2018.

[12] Deepa Gupta, Vani Raveendran, and Rahul Yadav. Domain biased bilingual parallel data extraction and its sentence level alignment for english-hindi pair. *Research Journal of Applied Sciences, Engineering and Technology*, 7:1001–1012, 02 2014.

[13] Viktor Hangya and Alexander Fraser. Unsupervised parallel sentence extraction with parallel segment detection helps machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1224–1234, Florence, Italy, 2019.

[14] Akbar Karimi, Ebrahim Ansari, and Bahram Sadeghi Bigham. Extracting an English-Persian parallel corpus from comparable corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, pages 1–5, Miyazaki, Japan, May 2018.

[15] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada, 2017.

[16] Wang Ling, Luís Marujo, Chris Dyer, Alan W. Black, and Isabel Trancoso. Crowdsourcing high-quality parallel data extraction from twitter. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 426–436, Baltimore, Maryland, USA, June 2014.

[17] Pintu Lohar, Debasis Ganguly, Haithem Afli, Andy Way, and Gareth J. F. Jones. Fada: Fast document aligner using word embedding. *The Prague Bulletin of Mathematical Linguistics*, 106:169–179, 2016.

[18] 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*, pages 311–318, Philadelphia, Pennsylvania, USA, 2002.

[19] Dana Ruiter. Online parallel data extraction with neural machine translation. Masters thesis, Saarland University, 2019.

[20] Rico Sennrich and Martin Volk. MT-based sentence alignment for OCR-generated parallel texts. In *The Ninth Conference of the Association for Machine Translation in the Americas*, pages 1–11, Denver, Colorado, USA, 2010.

[21] Brian Thompson and Philipp Koehn. Vecalign: Improved sentence alignment in linear time and space. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 1342–1348, Hong Kong, 2019.

[22] Krzysztof Wolk, Emilia Rejmund, and Krzysztof Marasek. Multi-domain machine translation enhancements by parallel data extraction from comparable corpora. *Computing Research Repository*, abs/1603.06785, 2016.