The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 167-181


# Bilingual Parallel Sentence Extraction from Comparable Corpora

Chien-Yu Chien, Chin-Hua Chang, Chih-Ping Wei
Department of Information Management
National Taiwan University
r07725026@ntu.edu.tw, r07725031@ntu.edu.tw, cpwei@ntu.edu.tw

## Abstract

This research aims to develop a parallel sentence extraction method for automatically extracting parallel sentence pairs from bilingual comparable corpora based on cross-lingual word embeddings. Our task is to effectively identify matched sentence pairs from a Chinese-English corpus with the goal of maximizing F1 score. Our method employs pre-trained, task-specific, and hybrid (a combination of pre-trained and task-specific) monolingual word embeddings to construct a cross-lingual transformation matrix respectively to transform the word embeddings between the two languages, and develops two search strategies (sequential and exhaustive) for parallel sentence extraction. Our empirical evaluation results suggest that task-specific word embeddings (directly trained from a task-relevant corpus, i.e., 25,695 Chinese and English abstracts of theses) outperforms their counterparts. With respect to the two search strategies, our evaluation results suggest that the exhaustive search strategy attains a higher recall rate; the sequential search strategy is more efficient in time. Both strategies achieve a promising performance, with an F1 score up to 60.18%.

## 1. Introduction

Recently, the tremendous development of neural network techniques for natural language processing has been introduced. Many studies have demonstrated promising results in many important applications, such as neural machine translation [1] and relation extraction [2]. One of the basic requirements for successful neural network model training is a sufficient number of qualified training data examples. In the case of neural machine translation, a bilingual parallel sentence corpus is a required data set. However, many low-resource language pairs (e.g., Chinese- English) or specific domains (e.g., biomedical research) have only limited bilingual parallel sentence corpora, which are not sufficient to support high-performance model construction. Generating parallel sentences by humans is both time consuming and resource intensive. Hence, recent research has focused on how to automatically extract parallel sentences from comparable corpora [3,4,5].

Comparable corpora include non-aligned sentences, phrases or documents that are not an exact translation of each other but share common features such as domain, genre, sampling period, etc. [6]. Compared to parallel corpora, there are more comparable corpora between languages, such as technical documents with bilingual abstracts. Therefore, once the accuracy of the parallel sentence pairs extracted from the comparable corpus can be realized, the problem of lack of a set of parallel sentences as a training corpus for neural machine translation can be effectively relaxed.

Recent research on parallel sentence extraction from comparable corpora has shifted from the feature-based approach [6] to the word-embedding-based approach [3,4], due to the advances on cross-lingual word embedding. Several methods for building cross-lingual word embeddings have been proposed [7,8,9]; among them, a popular method is through transformation matrix [8,9]. Most of existing word-embedding-based parallel sentence extraction methods are conducted on European language corpora (such as English and French). Prior research pays less attention to European-Oriental language pairs (e.g., English and Chinese), which is the focus of our study. Although some studies have investigated Chinese word embeddings [10], these studies are not for bilingual parallel sentence extraction.

Motivated by this research gap, we attempt to propose a word-embedding-based method for extracting parallel sentence pairs from Chinese-English comparable corpora. Our proposed method consists of three stages. First, we train the word embeddings for each language. Specifically, we obtain pre-trained word embeddings from BERT [11] as well as construct, on the basis of a task-relevant corpus, task-specific word embeddings, using the Word2Vec model [12]. Second, we learn a transformation matrix [8,9] to convert word embeddings from one language to another, thus creating cross-lingual word embeddings to align two different embedding spaces. Finally, with the use of the cross-lingual word embeddings, we compare bilingual sentence pairs by calculating their average word-by-word similarity and then extract parallel sentence pairs with a sequential or exhaustive search strategy. Furthermore, observing the phenomenon that an English sentence (segmented by period or question mark) often corresponds to multiple Chinese sentences (segmented by comma, period, or question mark), our proposed method allows many-to-one alignment, mapping multiple Chinese sentences into a single English sentence.

To evaluate the effectiveness of our proposed method, we conduct several experiments. We collect a Chinese-English comparable corpus that consists of 25,695 abstracts of theses. We then randomly selected 100 pairs (the abstracts of theses in both Chinese and English) in the corpus as the testing set. In this parallel sentence corpus, each

comparable document pair contains at least three matched parallel sentence pairs and a number of unmatched sentences. The other 25,595 pairs then serve as the training set for training monolingual word embeddings and constructing cross-lingual transformation matrices for different monolingual word embedding models. Our evaluation results show that our proposed method with task-specific word embeddings and the exhaustive search strategy achieves the highest effectiveness, reaching up to 60.18% in F1 score. The hybrid word embedding model, which combines pre-trained and task-specific word embeddings, is not as effective as the task-specific embedding model. The exhaustive search strategy attains better performance overall, whereas the sequential search strategy achieves a higher precision rate. We also discover the formation (Cbow or Skipgram [12]) of the monolingual word embeddings are sensitive parameters to this extraction task. The remainder of this paper is organized as follows: In Section 2, we describe the design of our proposed parallel sentence extraction method. Subsequently, we detail our evaluation design and discuss important experimental results in Section 3. Finally, Section 4 provides a summary of this study.

## 2. Our Proposed Method

Our proposed parallel sentence extraction method is to extract bilingual parallel sentence pairs from a comparable corpus. Because aligned documents in a comparable corpus often share similar themes and contents, parallel sentence pairs may exist in these aligned documents [6]. For example, Chinese technical papers or theses typically contain both Chinese and English abstracts, which usually describe the same or highly similar contents in the two languages. A pair of such aligned documents might include sentences that are exact translations or at least share common contents such as subjects, verbs and objectives. These sentence pairs are essential for training a neural machine translation model or for extracting translations for domain-specific terms.

The purpose of this study is to extract all sentence pairs with the same or highly similar content from a set of aligned bilingual documents. The constituent words of a sentence pair are assumed not necessarily consistent with the grammatical order or exact meaning. In order to estimate the similarity of bilingual sentence pairs, we decide to use cross-lingual word embeddings to find out the embedded relations between words and sentences. Accordingly, the research question of this study is formulated as: given a pair of aligned documents written in two different languages, our proposed method is to identify matched sentence pairs in the document pair, with the goal of maximizing the amount of extracted pairs while minimizing the likelihood of extracting wrong pairs.

Our proposed method consists of three stages: monolingual word embedding generation, cross-lingual word embedding generation, and parallel sentence extraction. Figure 1 shows the overall process of the proposed method.

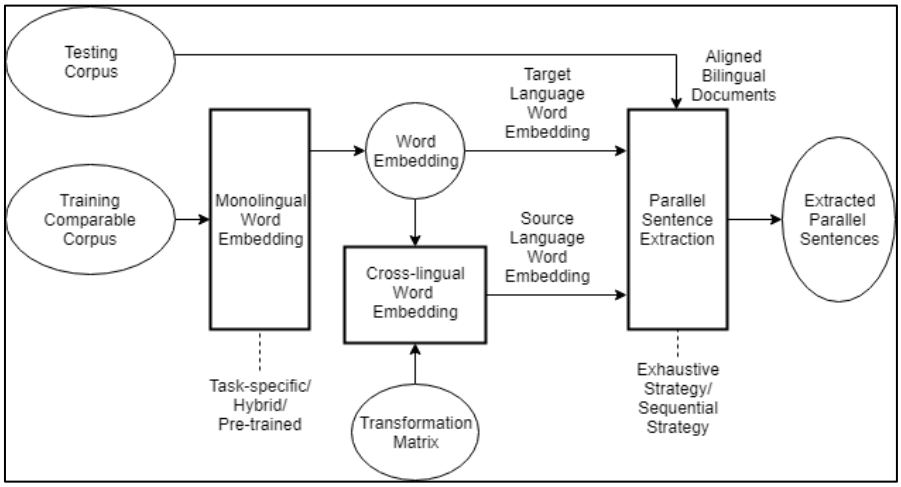


Figure 1. Overall process of our proposed parallel sentence extraction method.

## 2.1. Monolingual Word Embeddings

Before we link the representations of the two target languages (i.e., Chinese and English in our study), we need to create monolingual word embeddings for the two languages. Specifically, we independently train word embeddings for the source language and the target language, respectively. Several pre-processing steps are involved for the English corpus, including case unification, stemming, and stop word removal. For the Chinese corpus, word segmentation and stop word removal are performed.

Because a domain-specific corpus for word embedding training may not contain a sufficient number of paired documents, the quality of the resultant word embeddings may be compromised. In this study, we will incorporate and empirically evaluate BERT pre-trained models [11] in the proposed method. BERT, a language model developed by Google, uses the bidirectional training of Transformer (attention model) to language modelling and has been applied to many natural language tasks. BERT has released language models in more than 100 languages. In our experiments, we will evaluate the following word embedding models:

1. Pre-trained model: monolingual word embeddings directly from BERT pre-trained models.
2. Task-specific model: monolingual word embeddings directly trained from a task-relevant bilingual training corpus (i.e., 25,595 abstracts of theses).
3. Hybrid model: monolingual word embeddings by concatenating pre-trained and task-specific representations, thus doubling the number of dimensions.

It is noted that the pre-trained Chinese BERT model is based on characters [11], i.e. its vocabulary consists of single Chinese characters rather than words. It may not be optimized for our parallel sentence extraction task.

## 2.2. Cross-lingual Word Embeddings

Cross-lingual transfer of word embeddings is intended to establish semantic mapping between words in the source and target languages. In this study, we follow the transformation matrix approach to transform the source embedding space (i.e., word embeddings of the source language) to the target embedding one. We use the objective function of cross-lingual word embeddings from [8] to minimize the sum of the loss between $Wx$ and $y$:

$$\min_{W \in R^{d \times d}} \frac{1}{n} \sum_{i=1}^{n} \ell(W_{x_i}, y_i)$$

$\ell$ is the loss function, $W$ is transformation matrix with $d \times d$ dimensions, $x$ and $y$ are seed word pairs, $x_i$ is $i$th $x$'s word embedding, $y_j$ is $j$th $y$'s word embedding, and $n$ is the number of seed word pairs. In this study, we will construct an optimized transformation matrix for each word embedding model (i.e., pre-trained, task-specific, and hybrid model).

## 2.3. Parallel Sentence Extraction

Given a pair of aligned bilingual documents, our goal is to extract all possible semantically equivalent or highly similar sentence pairs in the aligned documents as the extracted parallel sentences. Therefore, we need to estimate the similarity of any pair of bilingual sentences based on the word embeddings of their constituent words.

### 2.3.1 Measuring the Similarity of a Sentence Pair

According to [4], the use of word-by-word similarity can reach a greater effectiveness than the use of sentence embedding similarity when measuring the similarity of two sentences. Thus, in this study, we adopt the word-by-word similarity approach to determine whether two sentences in different languages are similar. Assume that we are estimating the similarity between a source sentence $S$ and a target sentence $T$, where $S = [s_1\ s_2\ \dots\ s_n]$, $s_i$ is the $i$th word in $S$, $n$ is the number of words in $S$, $T = [t_1\ t_2\ \dots\ t_m]$, and $m$ is the number of words in $T$. For each word $s_i$ (denoted as source word) in the source sentence $S$, we calculate the cosine similarity between $s_i$ and every word $t_j$ (denoted as target word) in the target sentence $T$, according to their word embeddings. That is, for source word $s_i$ and target word $t_j$,

we obtain CosSim($Vs_i$, $Vt_j$), where $Vs_i$ is the transformed word embedding of $s_i$ (i.e., $W \times s_i$) and $Vt_j$ are the word embeddings of $t_j$. Among all of the candidates in the target sentence, the word that attains the largest cosine similarity to the source word $s_i$ is identified as the matched word for $s_i$. After every source word has found a matched word from the target sentence, the average of the similarities of all matched word pairs is calculated and used as the similarity of the source and target sentences.

In [4], when a word pair matches, both the source and the matched target word are removed from the corresponding sentences. This process (word removal during the comparison process of two sentences) may be appropriate for sentences that are written in the same language family (e.g., both the source and target languages are European languages). Because our study deals with Chinese-English language pairs, this process may be inappropriate due to the differences between Chinese and English languages. Specifically, in the Chinese-English scenario, especially after word segmentation, there are many cases in which multiple Chinese words link to the same English word. For example, "新創企業" in Chinese means "startup" in English. However, the Chinese term is typically segmented into "新創" and "企業". Assume that we match "新創" with "startup" and then remove "新創" and "startup" from the subsequent comparison process, we may not be able to match "企業" with any other remaining word, because other similar words such as "company" or "firm" may not appear in the focal English sentence. If we keep the word "startup" in the English sentence, it is likely that the word "企業" in the Chinese sentence can match this English word then. Accordingly, in our study, when a word pair matches, we will not remove both the source word and the matched target word from the subsequent comparison process. The similarity between a source sentence $S$ and a target sentence $T$ is then redefined as follows. We will empirically validate whether the redefined similarity method without word removal can achieve better performance in Section 3.5.

$$Sim(S,T) = \frac{1}{n}\sum_{i=1}^{n} \max_{j \in T} CosSim(V_{s_i}, V_{t_j})$$

### 2.3.2 Matching Sentence Pairs from an Aligned Document Pair

Previous studies focused on one-to-one sentence matching. When dealing with an aligned Chinese-English document pair, we need to address the difference between sentence segmentation for English documents and that for Chinese documents. For English documents, we generally segment sentences by periods and question marks. However, in Chinese writing, people often concatenate many subsentences by commas into a long sentence (ended with a

period symbol or a question mark). Thus, if we follow the sentence segmentation for English documents to segment a Chinese document into sentences, we may create overly lengthy Chinese sentences, each of which may not be semantically coherent. To avoid this problem, in this study, we segment a Chinese document by commas, periods, and question marks, so that a Chinese sentence may correspond to an English sentence or a part of an English sentence. In other words, in our study, a single English sentence may be aligned to one or multiple Chinese sentences. As shown in the first example in Table 1, one Chinese sentence (C1) is mapped to one English sentence (E1). In contrast, in the second example, two Chinese sentences (C2-1 and C2-2) correspond to one English sentence (E2), where C2-1 is the translation of the subsentence before the comma in E2 and C2-2 is for the subsentence after the comma in E2. To deal with many-to-one (multiple Chinese sentences to one single English sentence) sentence matching, we develop two search strategies (sequential vs. exhaustive search strategy), which will be detailed in the following.

Table 1. Examples of parallel sentence pairs in English and Chinese.

| | |
|---|---|
| C1: 藥物 開發 成本 高昂 且 費時 | E1: drug development is costly and time-consuming |
| C2-1: 因此 為了 解決 藥物 開發 的 困難<br>C2-2: 許多 研究 人員 開始 尋求 替代 方法 | E2: as a result to overcome the challenges of drug development, researchers start to explore alternative methods for drug development |

### 2.3.2.1 Sequential Search Strategy

Given an aligned bilingual document pair ($D_S$, $D_T$) and a similarity threshold $\alpha$, the sequential search strategy first compares the first source sentence $S_1$ with the first target sentence $T_1$. In our study, source sentences in $D_S$ are in Chinese language and target sentences in $D_T$ are in English language. When comparing a source sentence (denote as $S_i$) and a target sentence (denoted as $T_j$), the following two cases emerge:

Case 1: If the similarity is lower than a blocking threshold (in this study, we set the blocking threshold = $\alpha/3$, lower than the sentence-similarity threshold $\alpha$), we skip $T_j$ and move to check the next target sentence $T_{j+1}$. The search process continues. When all candidate target sentences have examined (the candidate target sentences for $S_i$ include the target sentences in range of $i \pm (\lambda\text{-}1)$, i.e., from $T_{i-(\lambda-1)}$ to $T_{i+(\lambda-1)}$, where $\lambda$ = the maximum location span to check) and all of the target sentences are dissimilar to the source sentence with respect to the blocking threshold, $S_i$ will be discarded. When this happens, we move to the next source sentence $S_{i+1}$ and start this search process.

Case 2: If the similarity is higher than the blocking threshold, we concatenate with the source sentence ($S_i$) all possible sequential combinations of the following $k$-1 source sentences (i.e.,

$S_{i+1}$ to $S_{i+k-1}$), thus generating $k$ source candidates (including $S_i$, $S_i + S_{i+1}$, $S_i + S_{i+1} + S_{i+2}$, ..., $S_i + \ldots + S_{i+k-1}$). We then compare each of the source candidates with the target sentence (in this case, $T_j$). Specifically, we calculate their similarity discounted by length difference. Sentence pairs with greater length differences (measured by the number of words) are unlikely to be parallel sentences. Therefore, the similarity score should be decreased by length difference between the source candidate and the target sentence. The length-difference-penalized similarity score between a source candidate $S_x$ and a target sentence $T_y$ is defined as follows:

$$Score = Sim(S_x, T_y) * (1 - \frac{|len(S_x) - len(T_y)|}{len(S_x) + len(T_y)})$$

Once we complete the calculation of the similarities of these source candidates with the target sentence, we choose the one with the highest similarity and check if it surpasses the predefined similarity threshold $\alpha$. If the highest similarity does not reach $\alpha$, we then head to the next target sentence $T_{j+1}$. However, if the highest similarity exceeds the predefined $\alpha$, we consider this pair of the specific source candidate and the target sentence as a parallel sentence pair and extract them out of the aligned bilingual document pair ($D_S$, $D_T$).

After successfully extracting a parallel sentence pair, we move to the next source sentence and the next target sentence. For example, suppose we find a successful parallel sentence pair ($S_1$ + $S_2$, $T_1$) and we will restart the search process using $S_3$ as the source sentence and $T_2$ as the target sentence. If we discard any source sentence, we will anchor the search process from the next source sentence (e.g., $S_{i+1}$) and the range of the target sentences to check is from $T_{i-\lambda}$ to $T_{i+\lambda}$. For example, suppose $S_1$ fails and is then discarded. We will start the search process by letting the source sentence as $S_2$ and the target sentence as $T_1$ (because $T1$ has not been aligned with any source sentence and is within the range of the target sentences to check for $S_2$). However, if we discard too many source sentences, we will start discarding target sentences. For example, let $\lambda$ = 5. Assume that $S_1$ to $S_5$ are all failed and discarded. Instead of keeping the target location at $T_1$, we will start the search process for $S_6$ with the target sentence $T_2$ (not $T_1$, because $T_1$ is not within the range of the target sentences to check for $S_6$) as the beginning search point.

### 2.3.2.2 Exhaustive Search Strategy

The exhaustive search strategy is to compare each source candidate (one or at most $k$ consecutive Chinese sentences in $D_S$) with every target sentence (an English sentence in $D_T$) in an aligned bilingual document pair. Then, we select the sentence pair (consisting of a source candidate and a target sentence) with the highest similarity. If the similarity of the

selected sentence pair is equal to or higher than the predefined threshold $\alpha$, it is extracted as a parallel sentence pair and the corresponding source candidate and target sentence are removed from $D_S$ and $D_T$, respectively. Subsequently, the sentence pair with the next highest similarity is selected and checks against $\alpha$ to see whether it can be extracted as a parallel sentence pair. The process repeats until the selected sentence pair's similarity is less than $\alpha$.

The differences between the exhaustive search strategy and the sequential search strategy are twofold. First, the search process of the sequential search strategy is sequential, from the beginning of each document, whereas the search process of the exhaustive search strategy compares all possible sentence pairs. As a result, the sequential search strategy is more efficient than the exhaustive search strategy, especially when source and target documents are large in their length. Second, the sequential search strategy imposes a blocking threshold (i.e., $\alpha/3$ in our study) and a maximum location span ($\lambda$) during the search process, while the exhaustive search strategy does not. As a result, the sequential search strategy may result in a suboptimal solution, possibly leading to inferior extraction effectiveness. We will report our evaluation of the two search strategies in Section 3.

# 3. The Experiments

## 3.1 Dataset

Our dataset was a corpus containing 25,695 bilingual (Chinese and English) abstracts of theses from the science, engineering, management, and medical colleges in National Taiwan University, Taiwan. We randomly selected 100 bilingual abstracts in this corpus as the testing set. The remaining 25,595 abstracts are the training set for generating monolingual word embeddings and a cross-lingual transformation matrix. Seven coders (graduate students of National Taiwan University) helped manually identify parallel sentence pairs from the testing set as the ground truth for our experiments. Each matched parallel sentence pair contains one English sentence and multiple (one to five) Chinese sentences, and there are at least three matched parallel sentence pairs for each pair of abstracts. 66.7% of the testing English sentences have matched Chinese sentences, and the average number of Chinese sentences in each parallel sentence pair is 1.81. Table 2 lists the statistics of our data set.

Table 2. Statistics of our data set including training and testing sets. 1044 out of 1462 Chinese sentences, and 576 out of 864 English sentences are matched pairs.

| | **# of Documents** | **Word Count** | **Sentence Count** | **# of Sentences per Doc** |
|---|---|---|---|---|
| Training Set (Chinese, Zh) | 25,595 | 2,388,729 | 346,591 | 13.54 |
| Training Set (English, En) | 25,595 | 1,981,510 | 195,910 | 7.65 |
| Testing Set (Zh) | 100 | 14,618 | 1,462 | 14.62 |
| Testing Set (En) | 100 | 15,000 | 864 | 8.64 |

## 3.2 Comparative Evaluation Results

As mentioned previously, our proposed parallel sentence extraction method can use one of the following monolingual word embeddings: 1) task-specific word embeddings (denoted as **TS**) directly trained from the training set, using Cbow or Skipgram from [12], 2) pre-trained word embeddings (denoted as **PRE**) extracted from BERT [11], and 3) hybrid word embeddings (denoted as **HB**) that concatenate TS and PRE word vectors. In the following experiment, we first employed Cbow to build task-specific word embeddings. We will compare the performance differential when using Cbow or Skipgram to build task-specific word embeddings in Section 3.4. Furthermore, for the TS model and the PRE model, the number of dimensions for word embedding was set to 200, and for the HB model, it was 400. The number of dimensions of the PRE model was originally 768 and was reduced from 768 to 200 via dimension reduction using principal component analysis.

Before we conduct our experiment, the first test is to decide the transformation direction, i.e., whether the transformation from Chinese (Zh) words to English (En) is better than the opposite direction (the transformation from English words to Chinese. In our test on 576 matched pairs in the testing set and another 576 randomly selected, non-matched pairs, the sentence similarities calculated by the Zh-En transformation attained higher average similarity on the matched pairs, lower average similarity on the random pairs, and greater difference between true and false pairs, as compared to those of the En-Zh transformation, as Table 3 illustrates. As a result, we decided the transformation direction is from Chinese to English, and set the Chinese corpus as source documents and the English corpus as target documents for subsequent experiments.

Table 3. Comparison of sentence similarity conducted by different transformation directions (Zh-En and En-Zh transformation).

| **Word Embedding** | **Source Language** | **Target Language** | **Avg Sim (Matched Pairs)** | **Avg Sim (Random False Pairs)** | **Difference** |
|---|---|---|---|---|---|
| TS | Zh | En | 0.3142 | 0.1504 | **0.1648** |
| TS | En | Zh | 0.3021 | 0.1537 | 0.1484 |
| PRE | Zh | En | 0.5556 | 0.4238 | **0.1321** |
| PRE | En | Zh | 0.6726 | 0.561 | 0.1116 |

We built each word embedding model's transformation matrix by linear regression with stochastic gradient descent, using 2,000 most commonly used English words (stop words and words without Chinese translation have been removed) in the training set as seed words. We then evaluated our proposed method using the metrics of precision, recall, and F1.

Table 4 shows the comparative evaluations results, across the three word embedding models and two search strategies, where SEQ denotes the sequential search strategy and EX is the exhaustive search strategy. To determine the similarity threshold $\alpha$, both strategies took the multiplication of the average sentence similarity of 1,000 prepared parallel sentences and a coefficient (an optimal coefficient was empirically determined). Furthermore, for the sequential search strategy, we set the blocking threshold as one third of $\alpha$. For each English sentence candidate, we set $k$ = 5 (up to 5 Chinese sentences to be concatenated), and $\lambda$ (maximum location span) = 5.

Table 4. Performance comparison of our proposed parallel sentence extraction method using different search strategies and word embedding models.

| **Search Strategy** | **Word Embedding** | **Threshold Coefficient** | **Recall** | **Precision** | **F1** |
|---|---|---|---|---|---|
| SEQ | TS | 0.9 | 36.54% | **68.55%** | 47.67% |
| SEQ | PRE | 0.8 | 31.17% | 49.60% | 38.28% |
| SEQ | HB | 0.95 | 31.53% | 59.74% | 41.28% |
| EX | TS | 0.7 | **56.78%** | 64.00% | **60.18%** |
| EX | PRE | 0.7 | 21.94% | 22.77% | 22.35% |
| EX | HB | 0.8 | 54.23% | 64.67% | 58.99% |

As Table 4 shows, the exhaustive search strategy using the task-specific embedding model achieved the best performance in recall and F1 measure, while the sequential search strategy using the task-specific embedding model attained the highest precision rate. With either the sequential search strategy or the exhaustive search strategy, the task-specific embedding model generally outperformed its counterparts, whereas the pre-trained word embedding model performed worst. This finding suggests that the compatibility of the corpus used to generate monolingual word embeddings and the testing corpus (i.e., parallel sentence extraction task) significantly affects the effectiveness of parallel sentence extraction.

We also observed that when using the exhaustive search strategy, the F1 score attained by the pre-trained embedding model was significantly lower than when using the sequential search strategy. The sequential search strategy compares only sentences with similar positions in the two aligned bilingual documents, while the exhaustive search strategy compares all possible sentence pairs. Because of the low compatibility of the pre-trained embedding model with the testing corpus, the exhaustive search strategy identified more false pairs than the sequential search strategy, highlighting the limitation of the pre-trained embedding model.

On the other hand, we expect that the hybrid embedding model could combine the advantages of the pre-trained embedding model and the task-specific embedding model and could achieve a better performance than the other two models. However, according to Table 4, the performance of the hybrid embedding model was in between that of the task-specific embedding model and the pre-trained embedding model. This is because the unacceptable performance attained by the pre-trained embedding model implicates the performance of the hybrid model.

## 3.3 Performance of Sequential and Exhaustive Search Strategies

Figure 2 shows the performance differences using the sequential or exhaustive search strategy. Since the sequential search strategy does not compare a source sentence with target sentences located far from the corresponding location of the source sentence. This strategy is likely to miss some matched pairs. Thus, the recall rate of the sequential search strategy is expected to be lower than that of the exhaustive search strategy. In contrast, because these relative distant sentence pairs are mainly false positives, the sequential search strategy can reach a higher accuracy than the exhaustive search strategy. Overall, the F1 score attained by the exhaustive search strategy is higher than that by the sequential search strategy, but the exhaustive search strategy is more time consuming. Figure 3 and Figure 4 show that the exhaustive search strategy performed better in recall rate, while the sequential search strategy performed slightly better in precision rate.

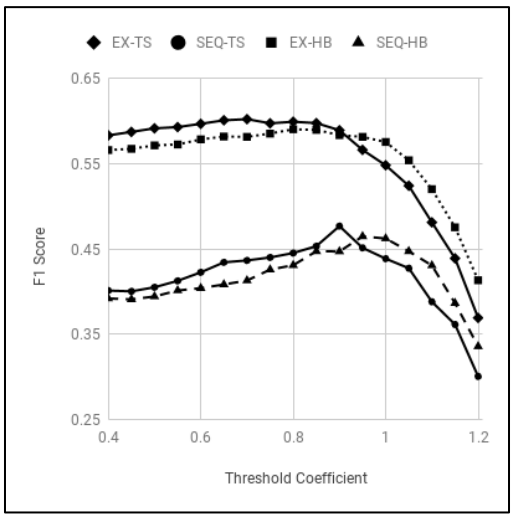


Figure 2. F1 measures obtained by the sequential and exhaustive search strategies. X dimension represents threshold coefficient and Y dimension represents F1 score.

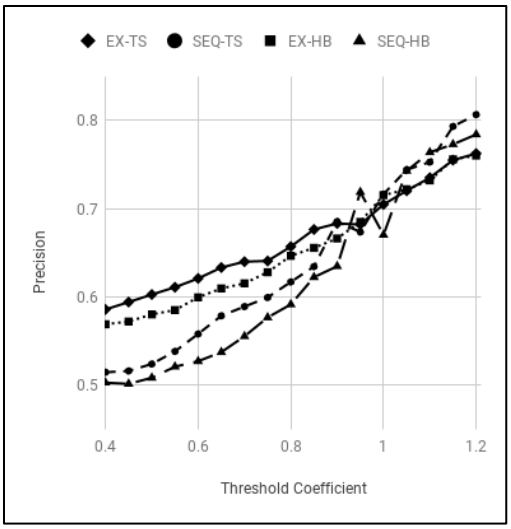


Figure 3. Precision rates obtained by using the sequential and exhaustive search strategies. X dimension represents threshold coefficient and Y dimension represents precision rate.

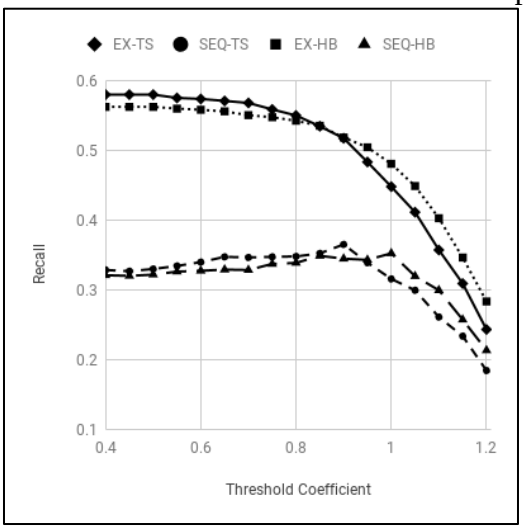


Figure 4. Recall rates obtained by using the sequential and exhaustive search strategies. X dimension represents threshold coefficient and Y dimension represents recall rate.

## 3.4 Effect of Cbow and Skipgram

We also analyzed the effect of Cbow and Skipgram on the effectiveness of parallel sentence extraction. Table 5 shows that the word embedding models (task-specific and hybrid) constructed by Cbow performed better than the models constructed by Skipgram, similar to the results reported in [13]. According to [12], if Cbow is trained on a large corpus, it would perform better than Skipgram. It seems that our corpus is relatively sufficient for Cbow.

Table 5. Performance comparison across different word embedding structures.

| **Word Embedding** | **Structure** | **Threshold Coefficient** | **F1** |
|---|---|---|---|
| TS | Skipgram | 0.7 | 43.03% |
| TS | Cbow | 0.7 | **60.18%** |
| HB | Skipgram | 0.8 | 30.33% |
| HB | Cbow | 0.8 | **58.99%** |

## 3.5 Effect of Word Removal When Measuring the Similarity of Two Sentences

To understand the effect of word removal when measuring the similarity of two sentences, Table 6 shows the performance obtained with or without word removal using the exhaustive search strategy. In general, our proposed parallel sentence extraction method without word removal achieved a lower precision rate, but a higher recall rate, as compared to our proposed method with word removal. With respect to F1 score, our proposed method without word removal outperformed that with word removal, across the two word-embedding models (task-specific and hybrid)

Table 6. Performance comparison with or without word removal during sentence extraction.

| **Word Embedding** | **Threshold Coefficient** | **Word Removal** | **Recall** | **Precision** | **F1** |
|---|---|---|---|---|---|
| HB | 0.8 | Yes | 50.44% | 68.52% | 58.10% |
| HB | 0.8 | No | 54.23% | 64.67% | **58.99%** |
| TS | 0.7 | Yes | 54.51% | 66.25% | 59.81% |
| TS | 0.7 | No | 56.78% | 64.00% | **60.18%** |

## 4. Concluding Remarks

In this work we have proposed and implemented an effective method for extracting parallel sentence pairs from bilingual comparable corpora. The effects of differences in word embedding model (task-specific/pre-trained/hybrid), search strategy (sequential/exhaustive), word vector formation (Cbow/Skipgram), and word removal or not have been empirically evaluated. By using the task-specific word embedding with the exhaustive search strategy, our proposed method can achieve the best performance in F1 score.

## 5. References

[1] M. Artetxe, G. Labaka, E. Agirre, and K. Cho (2018). "Unsupervised Neural Machine Translation." In *Proceedings of the Sixth International Conference on Learning Representations*.

[2] Y. Su, H. Liu, S. Yavuz, I. Gur, H. Sun, and X. Yan (2018). "Global Relation Embedding for Relation Extraction." In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 820-830.

[3] H. Bouamor and H. Sajjad (2018). "H2@BUCC18: Parallel Sentence Extraction from Comparable Corpora Using Multilingual Sentence Embeddings." In *Proceedings of the*

*Eleventh Workshop on Building and Using Comparable Corpora* at *International Conference on Language Resources and Evaluation*, pp. 43-47.
[4] V. Hangya , F. Braune, Y. Kalasouskaya, and A. Fraser (2018). "Unsupervised Parallel Sentence Extraction from Comparable Corpora." In *Proceedings of the International Workshop on Spoken Language Translation*, pp. 7-13.
[5] J. Smith, C. Quirk, and K. Toutanova (2010). "Extracting Parallel Sentences from Comparable Corpora Using Document Level Alignment." In *Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 403-411.
[6] D. Wu, and P. Fung (2005). "Inversion Transduction Grammar Constraints for Mining Parallel Sentences from Quasi-comparable Corpora." In *Proceedings of the International Conference on Natural Language Processing*, pp. 257-268.
[7] W. Yang, W. Lu, and V. Zheng (2017). "A Simple Regularization-based Algorithm for Learning Cross-Domain Word Embeddings." In *Proceedings of the International Conference on Empirical Methods in Natural Language Processing*, pp. 2898-2904.
[8] A. Joulin, P. Bojanowski, T. Mikolov, H. Jegou, and E. Grave (2018). "Loss in Translation: Learning Bilingual Word Mapping with A Retrieval Criterion." In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2979-2984.
[9] M. Artetxe, G. Labaka, and E. Agirre (2017). "Learning Bilingual Word Embeddings with (Almost) No Bilingual Data." In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pp. 451-462.
[10] R. Yin, Q. Wang, P. Li, R. Li, and B. Wang (2016). "Multi-granularity Chinese Word Embedding." In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 981-986.
[11] J. Devlin, M. Chang, K. Lee, and K. Toutanova (2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". arXiv preprint arXiv:1810.04805.
[12] T. Mikolov, Q. Le, and I. Sutskever (2013). "Exploiting Similarities among Languages for Machine Translation." arXiv preprint arXiv:1309.4168.
[13] L. Jin and W. Schuler (2015). "A Comparison of Word Similarity Performance Using Explanatory and Non-explanatory Texts." In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 990-994.