Document Alignment Based on Overlapping Fixed-Length Segments

Xiaotian Wang[1] **Takehito Utsuro**[1] **Masaaki Nagata**[2]
[1]Deg. Prog. Sys.&Inf. Eng., Grad. Sch. Sci.&Tech., University of Tsukuba
[2]NTT Communication Science Laboratories, NTT Corporation, Japan
[1]s2320811@_u.tsukuba.ac.jp, [1]utsuro@_iit.tsukuba.ac.jp
[2]masaaki.nagata@_ntt.com

Abstract

Acquiring large-scale parallel corpora is crucial for NLP tasks such as Neural Machine Translation, and web crawling has become a popular methodology for this purpose. Previous studies have been conducted based on sentence-based segmentation (SBS) when aligning documents in various languages which are obtained through web crawling. Among them, the TK-PERT method (Thompson and Koehn, 2020) achieved state-of-the-art results and addressed the boilerplate text in web crawling data well through a down-weighting approach. However, there remains a problem with how to handle long-text encoding better. Thus, we introduce the strategy of **O**verlapping **F**ixed-**L**ength **S**egmentation (OFLS) in place of SBS, and observe a pronounced enhancement when performing the same approach for document alignment. In this paper, we compare the SBS and OFLS using three previous methods, Mean-Pool, TK-PERT (Thompson and Koehn, 2020), and Optimal Transport (Clark et al., 2019; El-Kishky and Guzmán, 2020), on the WMT16 document alignment shared task for French-English, as well as on our self-established Japanese-English dataset MnRN. As a result, for the WMT16 task, various SBS based methods showed an increase in recall by 1% to 10% after reproduction with OFLS. For MnRN data, OFLS demonstrated notable accuracy improvements and exhibited faster document embedding speed.

1 Introduction

During the training phase of tasks such as Neural Machine Translation, a substantial amount of parallel corpora is required. Web crawling has emerged as an efficient approach for gathering large-scale parallel datasets, such as the ParaCrawl Dataset (Bañón et al., 2020), the JParaCrawl Dataset (Morishita et al., 2022), CCAligned Dataset (El-Kishky et al., 2020), Wikimatrix (Schwenk et al., 2021a), and CCMatrix (Schwenk et al., 2021b).

The procedure for developing a parallel web-crawled corpus involves five steps (Bañón et al., 2020): web crawling, text extraction, document alignment, sentence alignment, and sentence pair filtering. Document alignment involves establishing associations between documents that are equivalent translations originating from distinct language collections, and it can be broadly categorized into three strategies, URL matching (El-Kishky et al., 2020; Germann, 2016; Gomes and Pereira Lopes, 2016), methods based on machine translation or bilingual lexicons (Gomes and Pereira Lopes, 2016; Esplà-Gomis, 2009; Dara and Lin, 2016; Shchukin et al., 2016; Marchisio et al., 2021), and leveraging sentence embeddings (Clark et al., 2019; El-Kishky and Guzmán, 2020; El-Kishky et al., 2020; Thompson and Koehn, 2020; Steingrimsson, 2023). The core concept of the last one involves transforming the sentences within documents into a series of feature vectors. These vectors are then used to calculate the similarity between documents from different languages, with pairs exhibiting high similarity selected as alignment results.

However, it should be noted that crawled documents may not have uniform sentence segmentation and contain a lot of boilerplate text, such as headers, dates, and navigation menus. Moreover, for potentially long sentences, critical information may be generalized by other non-essential details when encoding it into embedding. In this case, we explore an alternative approach for subdivision, which involves utilizing a fixed-length sliding window to partition segments, with a specified proportion of overlap between adjacent segments.

In summary, our contributions are as follows:

- We developed a high-quality, small-scale Japanese-English test dataset called MnRN for the document alignment task.
- We replaced SBS with OFLS and conducted reproductions using Mean-Pool, TK-PERT,

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 51–61
August 11-16, 2024 ©2024 Association for Computational Linguistics

and Optimal Transport, three sentence embedding based document alignment methods. The accuracy of each method improved by 1% to 10% on the WMT16 document alignment shared task. Additionally, employing the OFLS strategy on the MnRN dataset achieved comprehensive improvements in both accuracy and speed.

2 Related Work

The concept of mining parallel data from webs has already been proposed in the 20th century (Resnik, 1999). However, in earlier years, the most serious endeavors have been confined to large companies, such as Google (Uszkoreit et al., 2010) and Microsoft (Rarrick et al., 2011). Up to the present, there have been numerous large-scale web crawling datasets obtained through various strategies, including the ParaCrawl Dataset (Bañón et al., 2020) obtained through URL matching, the JParaCrawl Dataset (Morishita et al., 2022) based on machine translation, and both the Wikimatrix (Schwenk et al., 2021a) and the CCmatrix (Schwenk et al., 2021b) derived from multilingual sentence embeddings.

Among the various web crawling methods, Bitextor (Esplà-Gomis, 2009) is one of the most widely adopted tools. Additionally, it incorporates a module known as docalign (Buck and Koehn, 2016b), which employs a TF-IDF strategy to score document pairs within one language through machine translation of documents in other languages.

In the WMT16 bilingual document alignment shared task (Buck and Koehn, 2016a), many techniques, systems, and tools were proposed to align cross-lingual document pairs. NOVALINCS (Gomes and Pereira Lopes, 2016) submitted three systems based on a phrase-based statistical machine translation framework, attaining the highest accuracy. In the shared task, there exist numerous alternative methods based on translation systems (Dara and Lin, 2016; Buck and Koehn, 2016b), URL matching (Germann, 2016; Papavassiliou et al., 2016), or bilingual translation lexicon (Azpeitia and Etchegoyhen, 2016; Medved' et al., 2016). However, methods relying on translation systems are contingent upon the availability of a high-quality translator, which is often challenging to obtain in advance.

Since the emergence of Sentence-BERT (SBERT) (Reimers and Gurevych, 2019), which used a Siamese network with cosine similarity for contrastive learning English sentence embedding in 2019, there has been a proliferation of high-precision multilingual pre-trained sentence embedding models to date. In the same year, Artetxe and Schwenk (2019) proposed the LASER model, which employs max-pooling over the output of a stacked LSTM-encoder. Subsequently, Reimers and Gurevych (2020) utilized knowledge distillation to adapt the SBERT for multilingual applications, named multilingual-SBERT (mSBERT). More recently, Feng et al. (2022) (LaBSE) expanded upon the framework of a dual encoder to learn cross-lingual language-agnostic embeddings from a pre-trained language model (Conneau et al., 2020), demonstrating state-of-the-art performance on the bitext mining task.

Just as the application of word embedding in sentence alignment (Kajiwara and Komachi, 2016; Arase et al., 2023) is pertinent, the proposition of introducing sentence embedding in document alignment warrants thorough consideration. In 2020, Thompson and Koehn (2020) proposed a method (TK-PERT) that involves utilizing regionally emphasized windows generated by a modified PERT distribution (Vose, 2000) to assign weights for sentences and then forming the feature vector of the document. Following their steps, Sannigrahi et al. (2023) evaluated the performance of the TK-PERT method using the three currently predominant multilingual sentence embedding models: LASER, mSBERT, and LaBSE.

The application of Optimal Transport in cross-lingual alignment, initially performing sentence-level alignment based on word embeddings, known as Word Movers' Distance (WMD) (Kusner et al., 2015). Analogous to it, Sentence Movers' Distance (Clark et al., 2019; El-Kishky and Guzmán, 2020) based on Optimal Transport (OT) was introduced for document-level alignment.

3 Document Alignment

3.1 Machine Translation based Document Alignment

In this paper, we utilize the docalign module[1] of Bitextor as a baseline to implement TF-IDF based document alignment (Buck and Koehn, 2016b).

It tokenizes the target language documents and machine-translated documents to create a vocabu-

[1] https://github.com/bitextor/bitextor/tree/master/document-aligner

lary, and then calculates the inverse document frequency (IDF) value for each n-gram within it. Next, the feature vectors of both target language documents and translated documents are constructed by individually calculating the term frequency (TF) of their internal n-grams and integrating them with the obtained IDF values to yield TF-IDF representations. Finally, the document pair score is determined by summing the products of the TF-IDF values for matching n-grams in both the target language document and the translated document.

3.2 Sentence Embedding based Document Alignment

Overlapping Fixed-Length Segmentation For any given document, instead of using sentence-based segmentation (SBS), which splits the document into non-overlapping sentences using delimiters such as line breaks or periods, we create segments by tokenizing all the sentences within the document, subsequently splitting it into segments through a fixed-length sliding window, with a proportion of overlap between adjacent segments.

Language-Pair Dependent Overlapping Fixed-Length Segmentation While applying the segmentation strategy as mentioned above, we use the same fixed-length for splitting documents in both the source and target languages. However, it is commonly observed that different languages may require different numbers of tokens to convey the same meaning. For instance, the English sentence "I like dogs" requires only 3 tokens, while the Japanese sentence "私 は 犬 が 好き だ" ("I like dogs") needs 6 tokens. Therefore, it is worth considering whether using distinct fixed-lengths for segmentation in different languages would appear more natural. With this perspective, we propose a language-pair dependent proportion ρ to split the target language document with fixed-length ρL when segmenting the source language document using a fixed-length L.

For any document A, B in the source and target language, a sentence embedding model is used to perform dense sentence-level embedding, resulting in two sets of vectors, $\{e_{A,i}\}$ and $\{e_{B,j}\}$, representing the embeddings in document $*$. We utilized the following three methods to calculate document pair similarity and compare our proposed segmentation strategy OFLS with the use of SBS.

3.2.1 Mean-Pool

Following Thompson and Koehn (2020), we employ the "Mean-Pool" approach as the fundamental sentence embedding based method, which is to use the mean-pooled vectors from the sets $\{e_{A,i}\}$ and $\{e_{B,j}\}$ as the feature vectors for document A and B, using their similarity to score the document pair.

$$e_{A,mean} = \sum_{i=1}^{n} e_{A,i}/n \tag{1}$$

$$e_{B,mean} = \sum_{i=1}^{m} e_{B,i}/m \tag{2}$$

$$Docsim(A,B) = Sim(e_{A,mean}, e_{B,mean}) \tag{3}$$

where $e_{*,mean}$ represents the mean-pooled vector of document $*$, n, m represents the number of vectors in $\{e_{A,i}\}$ and $\{e_{B,j}\}$ respectively, and $Docsim(A,B)$ represents the document similarity score. We use cosine similarity for document similarity scoring.

3.2.2 TK-PERT

Thompson and Koehn (2020) introduced a windowing approach that incorporates the modified PERT function (Vose, 2000) to assess the significance of each sentence, along with a down-weighting mechanism for boilerplate text. The smoothed overlapping windowing functions embed nuanced positional details into the resultant document vector.

Let $e_{n|n\in\{0,\dots,N-1\}}$ represent the N multilingual sentence embeddings in a given document. The sub-vectors E_j are calculated to emphasize uniformly spaced positions $j \in \{0, ..., J-1\}$ in the document.

$$E_j = \sum_{n=0}^{N-1} e_n H_j(n) B_n \tag{4}$$

where $H_j(n)$ represents a windowing function utilized to accentuate the j^{th} region of the document, B_n serves to diminish the significance of boilerplate text using LIDF.[2]

The final document feature vector E is formed by concatenating normalized position-weighted sub-vectors $E_{j|j\in\{0,\dots,J-1\}}$, and cosine similarity is used to measure the similarity between documents.

[2] We follow the TK-PERT (Thompson and Koehn, 2020) definition of LIDF, which scales sentences based on the inverse of the (linear, rather than logarithmic) number of documents that contain the given sentence.

3.2.3 Optimal Transport based Method

Optimal Transport, also known as Earth Movers' Distance (EMD) (Rubner et al., 2000) and Wasserstein Metric, is a measure of the distance between two probability distributions. For the application in document alignment, known as Sentence Movers' Distance (SMD) (Clark et al., 2019; El-Kishky and Guzmán, 2020), it calculates the minimum cost of transforming the distribution of document A to the distribution of document B. It represents each document as a normalized *bag-of-sentences* (nBOS) where each segment has associated with its some probability mass.

Specifically, all segments from document A, B are utilized to establish a vocabulary of size V, with the sequence of embeddings $\{v_i\}$ for the ith segment. $d_{A,i}$ is defined as the weight of ith segment of vocabulary in document A. We adopt the assumption that gives weight to segments by relative frequencies,[3] which is calculated as follows:

$$d_{A,i} = cnt(i)/|A| \tag{5}$$

where $cnt(i)$ is frequency of ith segment in document A, and $|A|$ is the total number of segments in document A.

We denote $\Delta(i,j)$ as the cosine distance between the ith segment and jth segment, unlike Kusner et al. (2015), who utilized the Euclidean distance to calculate $\Delta(i,j)$. The SMD between document A and B can be calculated as follows:

$$\Delta(i,j) = 1 - Cos(i,j) \tag{6a}$$

$$SMD(A,B) = \min_{T\geq 0} \sum_{i=1}^{V} \sum_{j=1}^{V} T_{ij}\Delta(i,j) \tag{6b}$$

Subject to:

$$\forall i \sum_{j=1}^{V} T_{ij} = d_{A,i} \tag{7a}$$

$$\forall j \sum_{i=1}^{V} T_{ij} = d_{B,j} \tag{7b}$$

and $T \in \mathbb{R}^{V \times V}$ is a nonnegative matrix, where each T_{ij} denotes how much of segment i in document A is assigned to segments j in document B, and constraints ensure the flow of a given segment cannot exceed its allocated mass.

4 Experiment

4.1 Dataset

We manually developed the MnRN dataset by aligning document pairs obtained from four web domains: **M**arubeni, **n**ishi-shinjuku, **R**akuten, and **N**TT Computer Science. The simple introduction to each web domain is provided by Table 1.

Marubeni: `www.marubeni.com` Information about Marubeni Corporation, such as policies, management philosophy, and technical reports.
nishi-shinjuku: `nishishinjuku.co.jp` Information about hotels in nishi-shinjuku.
Rakuten: `corp.rakuten.co.jp___global.rakuten.com` Information about Rakuten Inc., such as employment and stock.
NTT Computer Science: `www.kecl.ntt.co.jp` Information about research presentations, lectures, and reports from the NTT Communication Science Laboratories.

Table 1: The brief introduction of each web domain.

For each web domain, we randomly sampled a set of Japanese documents, and then made a pool of candidates for corresponding English documents on the same web domain using four different document alignment methods:

- Machine Translation + BM25
- Machine Translation + TF-IDF
- URL matching
- CCAligned (El-Kishky et al., 2020)

We then manually selected the correctly corresponding English document for a Japanese document in the pool. Table 2 shows the details of documents in each web domain. Due to the occurrence of different URLs but identical contexts in English web pages, multiple aligned counterparts may exist for a single Japanese document. We consider all of them as gold pairs.

Web Domain	Ja Docs.	Gold Pairs	Candidate En Docs.
Marubeni	73	75	251
Nishi-Shinjuku	16	16	42
Rakuten	75	84	319
NTT CS	68	88	319
All	232	263	931

Table 2: Information of the MnRN dataset.

4.2 Experiment Setting

In this paper, we used the pre-trained JParaCrawl-v3.0-big model[4] (Morishita et al., 2022) based

[3] We refer to the program of OTalign (Arase et al., 2023) for OT calculation, which utilizes the POT Python library (https://pythonot.github.io/).

[4] https://www.kecl.ntt.co.jp/icl/lirg/jparacrawl/

on fairseq toolkit (Ott et al., 2019) for machine translation from Japanese to English on the MnRN dataset.

	WMT16 test data	MnRN
English Docs.	682k	931
French Docs.	522k	-
Japanese Docs.	-	232
Web domains	203	4
Gold Pairs	2402	263
Search direction	Fr-En	Ja-En
Search strategy	each domain	all domains
J	16	8
γ	20	16

Table 3: Counts and experiment settings for WMT16 test data and MnRN dataset.

LaBSE tokenizer and model[5] (Feng et al., 2022) was utilized for tokenizing and sentence embedding. As shown in Table 3, we used the test data provided by the WMT16 document alignment shared task (WMT16 test data) to conduct alignment for each web domain from French to English. However, for the MnRN dataset, we performed alignment without distinguishing domains from Japanese to English. J is used to determine the number of windows produced in the TK-PERT method, where for each document, modified PERT distributions (Vose, 2000) with modes of $(\frac{j+0.5}{J})N$ are generated for j over $[0, J-1]$, with N being the number of segments in the document, and γ is a hyperparameter to control the peakedness of the distribution.

Due to the abundance of documents within the web domain of the WMT16 test data, we utilized Faiss (Johnson et al., 2019) search to retrieve the top 32 similar documents for alignment candidates. As for the MnRN dataset, we only retrieved the top 20 candidates using "Mean-Pool" or "TK-PERT" for the OT method due to its smaller scale.

For "TK-PERT", following Thompson and Koehn (2020) and Sannigrahi et al. (2023) setting for the modified PERT distribution,[6] we use $J = 16$ and set its shape parameter to $\gamma = 20$ for the WMT16 test data, while we designate $J = 8$, $\gamma = 16$ for our self-established MnRN dataset.

However, it should be noted that the language-pair dependent proportion ρ is akin to the prior information. Nevertheless, we have not exploited the validation data for the MnRN dataset. Consequently, in our experiment, we used the bootstrap sampling strategy to extract 30 pairs of aligned document pairs and calculate the average ratio of the token counts between them during each iteration, repeating this 10 times. Finally, the mean value 0.63 of average ratios is adopted as the value for ρ.

The final result enforces the 1-1 rule: Each document should be aligned only once. We evaluate the final result on the MnRN dataset using the F1 Score,[7] which is contingent upon both precision and recall, where precision represents the ratio of Japanese documents in the correct pairs within the final result, and recall denotes the proportion of Japanese documents in the correct pairs out of the total Japanese documents. Meanwhile, we adhere to Buck and Koehn (2016a)[8] to evaluate the document pairs for the WMT16 bilingual document alignment shared task.

All the experiments are conducted on two NVIDIA RTX A6000 GPUs.

4.3 Result of MnRN dataset

As the result shown in Table 4, we measured the F1 Scores and the execution time consumed by all the document alignment methods.

For "MT + docalign", we recorded the time cost for translation and the time utilized for alignment using the docalign tool. For sentence embedding based methods, we calculated the time spent on generating embeddings or feature vectors based on those embeddings, as well as the time required for computing similarity between documents.

4.3.1 Accuracy

According to the results on the MnRN dataset, all sentence embedding based methods achieved F1 scores surpassing MT based docalign. Furthermore, utilizing overlapping fixed-length segments (OFLS) for document alignment comprehensively outperforms the approach relying on SBS. However, it is also noted that when using fixed-length segmentation without overlapping (FLS), all the methods exhibit slight improvements or even

[5] `https://huggingface.co/setu4993/LaBSE`

[6] However, in contrast to their research, we opt to utilize the "mc2d" library in Recovery Component (R) for generating modified PERT distributions, and abstaining from employing Principal Component Analysis (PCA) for dimensionality reduction of sentence embeddings.

[7] Due to adherence to the 1-1 rule, even if multiple gold pairs exist for a single Japanese document, there can be at most one in the final result. Therefore, when calculating the F1 Score, we rely on the number of Japanese documents in the correct pairs to determine precision and recall.

[8] We use a "soft" recall metric, wherein credit is assigned to pairs of documents where either the English or French document (but not both) deviates from a reference document pair by less than 5%, as measured by text edit distance.

Alignment Method	Segment Strategy	FL	OR	ρ	F1 Score	Time (sec.) (Translation\Embedding)	Time (sec.) (Similarity)
MT + docalign	SBS	-	-	-	0.7880	158.02s	3.93s
Mean-Pool	SBS	-	-	-	0.8276	277.29s	0.36s
	FLS	150	0.0	-	0.8147	71.17s	0.28s
	OFLS	150	0.5	-	**0.8621**	123.96s	0.33s
	OFLS	150	0.5	0.63	0.8491	120.07s	0.31s
TK-PERT	SBS	-	-	-	0.8448	352.50s	0.29s
	FLS	150	0.0	-	0.8578	124.78s	0.26s
	OFLS	150	0.5	-	**0.9052**	220.57s	0.27s
	OFLS	150	0.5	0.63	0.9009	288.41s	0.27s
OT w/Mean-Pool	SBS	-	-	-	0.8448	276.61s	25.92s
	FLS	100	0.0	-	0.8534	69.30s	15.07s
	OFLS	100	0.5	-	0.8966	119.44s	16.07s
	OFLS	100	0.5	0.63	**0.9267**	121.28s	16.55s
OT w/TK-PERT	SBS	-	-	-	0.8319	353.29s	25.45s
	FLS	100	0.0	-	0.8362	154.84s	14.85s
	OFLS	100	0.5	-	0.8966	280.49s	15.80s
	OFLS	100	0.5	0.63	**0.9267**	367.19s	16.30s

Table 4: The final results of Ja-En document alignment on MnRN dataset incorporating hyper-parameter settings, where "SBS" represents for sentence-based segmentation, "FLS" represents for fixed-length segments without overlapping, "OFLS" represents for overlapping fixed-length segments, "FL" represents for fixed-length of Japanese documents, "OR" represents for overlapping rate, "ρ" represents the language-pair dependent proportion as mentioned in Section 3, "Time (sec.) (Translation \ Embedding)" represents time consumption for Translation, which combines data preprocessing and translation process, or Embedding, which combines sentence embedding generation, feature vector development, and candidate search, "OT w/$*$" represents rescoring the top 20 candidates found based on the "$*$" method using Optimal Transport, where the sequence of sentence embeddings used for "OT" is as same as the "$*$" method, and "-" represents for not-used hyper-parameter.

declines. Hence, we discuss the impact of overlapping rates in Section 5.1.

"Mean-Pool" is considered as the most fundamental approach among sentence embedding based methods, yet every other method performs better than it in the F1 Score. Nevertheless, comparing different methods using only a single fixed-length may introduce bias into the experimental conclusion. Therefore, in Section 5.2, we conduct an evaluative analysis of the performance of each method across fixed-lengths from 10 to 300.

The "Language-Pair Dependent Overlapping Fixed-Length Segmentation" (LD-OFLS) leads to a slight decrease in performance for "Mean-Pool" and "TK-PERT", possibly due to the reliance on averaging or weighted averaging to derive the final feature vectors for distinguishing between documents, thereby attenuating the individual influence of each segment. However, this strategy has a positive impact on "OT w/$*$", as it considers the influence of each segment when calculating distances between documents, ultimately achieving the highest accuracy on the MnRN dataset. We also analyzed the overall performance of LD-OFLS in Section 5.3.

4.3.2 Calculation Speed

As the time cost recorded in Table 4, using OFLS noticeably reduces the time required for embedding compared to SBS.

Despite having the lowest accuracy among various sentence embedding based methods, "Mean-Pool" exhibits the fastest speed, suggesting its potential as a candidate-finding approach with fault tolerance. Although "TK-PERT" demonstrates high accuracy, due to the generation of LIDF and the modified PERT distribution, it requires additional time to generate feature vectors.

As for "OT w/$*$", since the search for candidates can be rapidly accomplished under Faiss retrieval, the time required for its embedding is essentially equivalent to the time needed to generate feature vectors. However, due to the limitations imposed by the "ot" function of the POT Python library, which can only operate on a pairwise basis, computing OT becomes computationally disadvantageous when the data size is enormous.

However, it is observed that "OT w/TK-PERT" and "OT w/Mean-Pool" exhibit minor differences on the MnRN dataset. This may be attributable to the small data size, where both “Mean-Pool” and “TK-PERT” can retrieve the ground truth into the candidates. In this case, the performance of “OT w/$*$” may rely more on its intrinsic accuracy rather than the candidates’ accuracy.

4.4 Result of WMT16 test data

We also conducted experiments on the WMT16 document alignment shared task. However, constrained to the substantial resource and time consumption brought about by the vast size of the dataset, we merely employed the OFLS segment strategy with a simple setting of fixed-length $FL = 100$ and overlapping rate $OR = 0.5$ without language-pair dependent proportion ρ for comparison against the SBS strategy. Additionally, we compared our results with the best-reported previous works, which are presented in Table 5.

Method	Segment Strategy	Recall
Previous work		
Dara and Lin (2016)	SBS	96.0%
Buck and Koehn (2016b)	SBS	96.2%
TK-PERT (LASER) (Thompson and Koehn, 2020)	SBS	**97.1%**
TK-PERT (LASER) (Sannigrahi et al., 2023)	SBS	96.4%
TK-PERT (LaBSE) (Sannigrahi et al., 2023)	SBS	94.2%
This work		
Mean-Pool	SBS	82.6%
Mean-Pool	OFLS	**92.6%**
TK-PERT (LaBSE)	SBS	95.2%
TK-PERT (LaBSE)	OFLS	**96.3%**
OT w/Mean-Pool	SBS	90.6%
OT w/Mean-Pool	OFLS	**93.7%**
OT w/TK-PERT	SBS	95.6%
OT w/TK-PERT	OFLS	**96.8%**

Table 5: Document recall on WMT16 test data, compared to previous best-reported results, where fixed-length FL is 100, overlapping rate OR is 0.5 for OFLS, and language-pair dependent proportion ρ is not used.

As mentioned in Section 4.2, due to the distinct configuration of “TK-PERT” as compared to previous works (Thompson and Koehn, 2020; Sannigrahi et al., 2023), we reproduced it using the LaBSE model under SBS. Upon contrasting SBS of this work with OFLS, it is observed that the recall of all document alignment methods improved by varying degrees from 1.1% to 10.0%, with “Mean-Pool” achieving the greatest enhancement.

While the best result of this work “OT w/TK-PERT” does not surpass the best-reported recall of 97.1% achieved by Thompson and Koehn (2020) in the WMT16 document alignment shared task, the replication of “TK-PERT” by Sannigrahi et al. (2023), utilizing different multilingual sentence embedding models, indicates that the LaBSE model performs less effectively on the WMT16 test data compared to the LASER model. Nevertheless, we achieved the best result in experiments based on the LaBSE model, surpassing the research based on machine translation by Dara and Lin (2016) and Buck and Koehn (2016b).

5 Ablation Analysis

In this section, we conducted an ablation analysis on three factors of OFLS: overlapping rate, fixed-length, and language-pair dependent proportion ρ. However, due to the substantial size of the WMT16 test data, our analysis was limited to the smaller-scale MnRN dataset.

5.1 Overlapping Rate

According to the results in Table 6, there are apparent discrepancies regarding the utilization of overlapping, and most F1 scores reach maximum values at the rate of 0.5, while “OT w/TK-PERT” achieves superior performance at the rate of 0.8.

Overlapping Rate	0.0	0.3	0.5	0.8
Mean-Pool	0.8147	0.0129↑	**0.0474**↑	0.0258↑
TK-PERT	0.8578	0.0172↑	**0.0474**↑	0.0086↑
OT w/Mean-Pool	0.8534	0.0388↑	**0.0432**↑	0.0216↑
OT w/TK-PERT	0.8362	0.0560↑	0.0604↑	**0.0690**↑

Table 6: The F1 Scores of different overlapping rates on the MnRN dataset, where fixed-length $FL = 150$ for “Mean-Pool” and “TK-PERT”, $FL = 100$ for “OT w/$*$”, and language-pair dependent proportion ρ is not used. The results of each method represent the relative differences from the case of the overlapping rate 0.0.

Conclusively, the judicious selection of the overlapping rate, with a suggested universally applicable value of 0.5, holds the potential for substantial improvement across diverse methods under the OFLS segmentation strategy.

5.2 Fixed-Length

In this section, we discuss the impact of fixed-length on the four methods. However, since the accuracy of “OT w/$*$” depends partly on the accuracy of candidates, and we only aim to compare the performance of OT, we standardize the candidates

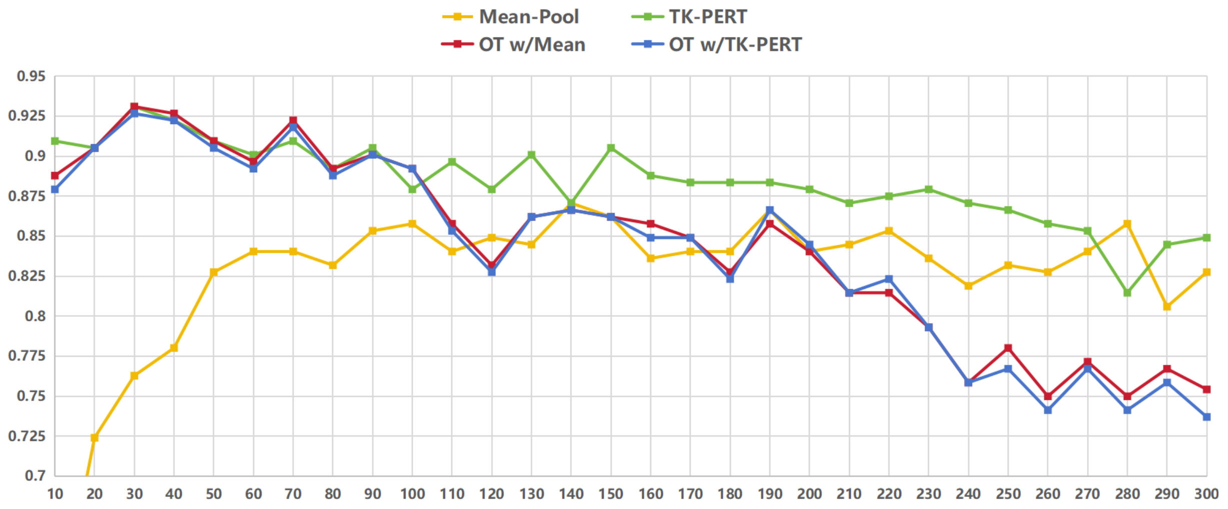

Figure 1: The F1 Scores of different fixed-lengths on the MnRN dataset. All the overlapping rates are 0.5, and language-pair dependent proportion ρ is not used.

retrieved for "OT w/$*$" in this section to $FL = 150$, $OR = 0.5$ without using ρ.

Based on the results depicted in Figure 1, it is observed that "Mean-Pool" exhibits poor performance when the fixed-length is less than 50. On the contrary, concurrently, the other three methods demonstrate commendable performance. As the fixed-length increases, the accuracy of "Mean-Pool" stabilizes without significant variation. Conversely, "TK-PERT" shows a slow declining trend, while "OT w/$*$" displays an obvious decrease, even becoming substantially weaker than "Mean-Pool" after reaching a fixed-length of 200.

On the one hand, the fixed-length determines the structure of segments, which may lead to variations in accuracy across methods, not displaying a strictly monotonic trend. On the other hand, it determines the number of segments: a small fixed-length results in numerous segments.

"Mean-Pool" can be viewed as an averaged representation of information within document segments. Excessive segmentation may dilute the features of each information component, ultimately failing to represent the document meaningfully. This may be a reason for its subpar performance at small fixed-lengths. However, it is noteworthy that the other methods perform well at small fixed-lengths. In the case of "OT w/$*$", compared to "Mean-Pool", it considers each segment without pooling the information, potentially making its performance superior with more segments. As for "TK-PERT", like "OT w/$*$", it utilizes multiple feature vectors to represent the document and achieves a similar trend but is more stable.

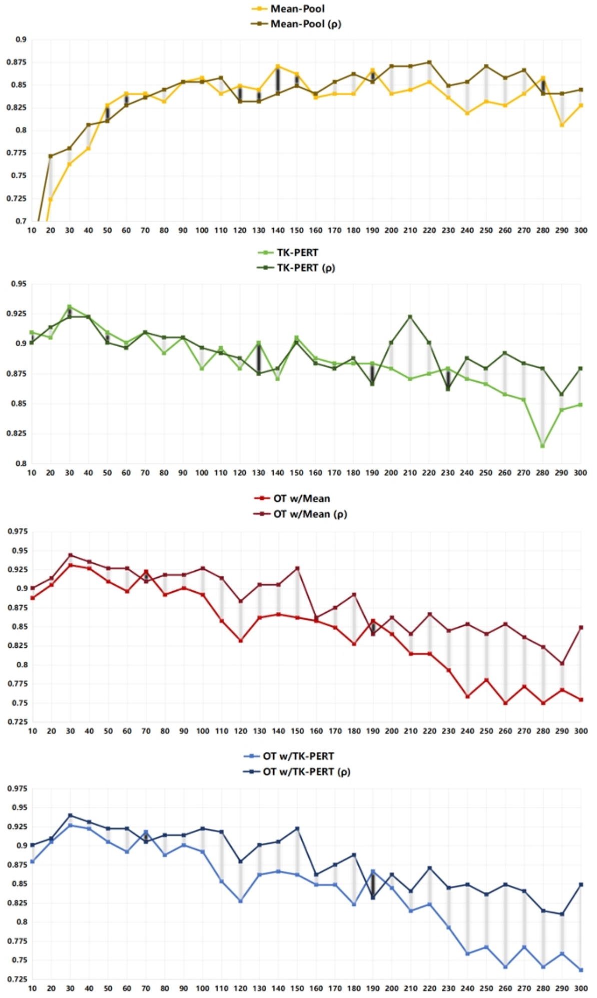

Figure 2: The F1 Scores of different fixed-lengths on the MnRN dataset with $\rho = 0.63$. All the overlapping rates are 0.5. The cases where the accuracy improved with the utilization of ρ are marked by gray bars between the two broken lines, whereas black bars are employed to denote the contrary scenario.

5.3 Language-Pair Dependent Proportion

We investigate the impact of $\rho = 0.63$ on different fixed-length settings while still fixing the candidates for "OT w/$*$" as in Section 5.2.[9]

Based on the results depicted in Figure 2, it is observed that for "OT w/$*$", the implementation of ρ leads to a comprehensive improvement in accuracy across various fixed-lengths. Furthermore, it mitigates the rapid decline in accuracy that typically accompanies an increase in fixed length.

The influence of ρ for "Mean-Pool" and "TK-PERT" is non-obvious prior to a fixed-length of 200. However, after the threshold of 200, a pronounced enhancement in performance is evident.

6 Conclusion

This paper presents the OFLS strategy designed for splitting documents into overlapping fixed-length segments for the document alignment task. Building upon the previous sentence embedding based methods, compared to SBS, OFLS yields better results on the WMT16 document alignment shared task. Specifically, the OFLS based "TK-PERT" and "OT w/TK-PERT" surpass the two best-recorded machine translation based methods, achieving the highest recall among LaBSE based approaches.

Simultaneously, we observed the same results on the MnRN dataset. Furthermore, according to the ablation analysis in Section 5, a smaller fixed-length can further improve accuracy for "TK-PERT" and "OT w/$*$", though it also results in longer embedding time and higher storage cost. Appropriate hyperparameters can enable OFLS to surpass SBS in both accuracy and speed.

Limitations

In Section 4, we conducted speed measurements exclusively on the MnRN dataset. However, the speed is constrained by the algorithm and computational memory. We can only compare various methods under relatively fair conditions, such as setting similar hyperparameters. Additionally, while we achieved better results than machine translation based methods across the two datasets, the resource consumption for storing sentence embeddings is higher than that for storing translated documents. Moreover, we only performed experiments on two language directions, which are relatively high-resourced. Lastly, this study focused solely on the document alignment task and did not discuss its subsequent impact on downstream work, like constructing machine translation datasets.

Ethical statement

The models used in this paper, LaBSE (Feng et al., 2022), and the JParaCrawl-v3.0-big model (Morishita et al., 2022), are publicly available for research. The WMT16 test data used in this study is provided by the WMT16 document alignment shared task (Buck and Koehn, 2016a).

References

Yuki Arase, Han Bao, and Sho Yokoi. 2023. Unbalanced optimal transport for unbalanced word alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3966–3986, Toronto, Canada. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Andoni Azpeitia and Thierry Etchegoyhen. 2016. DOCAL - vicomtech's participation in the WMT16 shared task on bilingual document alignment. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 666–671, Berlin, Germany. Association for Computational Linguistics.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Christian Buck and Philipp Koehn. 2016a. Findings of the WMT 2016 bilingual document alignment shared task. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 554–563, Berlin, Germany. Association for Computational Linguistics.

[9] Under the conditions of $FL = 150$ and $OR = 0.5$, we also experimented with various values of ρ for the three alignment methods to simulate the scenario of optimizing ρ by a validation dataset. The results indicate that changes in ρ have little impact on "Mean-Pool" and "TK-PERT", although an appropriate ρ value can still maximize the accuracy of "TK-PERT". Meanwhile, the choice of ρ has a more obvious effect on the accuracy of OT, with the experiment achieving the highest accuracy at the value of approximately 0.63.

Christian Buck and Philipp Koehn. 2016b. Quick and reliable document alignment via TF/IDF-weighted cosine distance. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 672–678, Berlin, Germany. Association for Computational Linguistics.

Elizabeth Clark, Asli Celikyilmaz, and Noah A. Smith. 2019. Sentence mover's similarity: Automatic evaluation for multi-sentence texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2748–2760, Florence, Italy. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Aswarth Abhilash Dara and Yiu-Chang Lin. 2016. YODA system for WMT16 shared task: Bilingual document alignment. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 679–684, Berlin, Germany. Association for Computational Linguistics.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5969, Online. Association for Computational Linguistics.

Ahmed El-Kishky and Francisco Guzmán. 2020. Massively multilingual document alignment with cross-lingual sentence-mover's distance. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 616–625, Suzhou, China. Association for Computational Linguistics.

Miquel Esplà-Gomis. 2009. Bitextor: a free/open-source software to harvest translation memories from multilingual websites. In *Beyond Translation Memories: New Tools for Translators Workshop*, Ottawa, Canada.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Ulrich Germann. 2016. Bilingual document alignment with latent semantic indexing. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 692–696, Berlin, Germany. Association for Computational Linguistics.

Luís Gomes and Gabriel Pereira Lopes. 2016. First steps towards coverage-based document alignment. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 697–702, Berlin, Germany. Association for Computational Linguistics.

J. Johnson, M. Douze, and H. Jégou. 2019. Billion-scale similarity search with GPUs. *Journal 2019 IEEE*, pages 535–547.

Tomoyuki Kajiwara and Mamoru Komachi. 2016. Building a monolingual parallel corpus for text simplification using sentence similarity based on alignment between word embeddings. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1147–1158, Osaka, Japan. The COLING 2016 Organizing Committee.

M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger. 2015. From word embeddings to document distances. In *Proc 32nd PRML*, pages 957–966.

Kelly Marchisio, Philipp Koehn, and Conghao Xiong. 2021. An alignment-based approach to semi-supervised bilingual lexicon induction with small parallel corpora. In *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 293–304, Virtual. Association for Machine Translation in the Americas.

Marek Medved', Miloš Jakubíček, and Vojtech Kovář. 2016. English-French document alignment based on keywords and statistical translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 728–732, Berlin, Germany. Association for Computational Linguistics.

Makoto Morishita, Katsuki Chousa, Jun Suzuki, and Masaaki Nagata. 2022. JParaCrawl v3.0: A large-scale English-Japanese parallel corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6704–6710, Marseille, France. European Language Resources Association.

M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proc NAACL 2019*, pages 48–53.

Vassilis Papavassiliou, Prokopis Prokopidis, and Stelios Piperidis. 2016. The ILSP/ARC submission to the WMT 2016 bilingual document alignment shared task. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 733–739, Berlin, Germany. Association for Computational Linguistics.

Spencer Rarrick, Chris Quirk, and Will Lewis. 2011. MT detection in web-scraped parallel corpora. In *Proceedings of Machine Translation Summit XIII: Papers*, Xiamen, China.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Philip Resnik. 1999. Mining the web for bilingual text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 527–534, College Park, Maryland, USA. Association for Computational Linguistics.

Y. Rubner, C. Tomasi, and L. Guibas. 2000. The earth mover's distance as a metric for image retrieval. In *Journal 2000 IJCV*, pages 99–121.

Sonal Sannigrahi, Josef van Genabith, and Cristina España-Bonet. 2023. Are the best multilingual document embeddings simply based on sentence embeddings? In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2306–2316, Dubrovnik, Croatia. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021a. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021b. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6490–6500, Online. Association for Computational Linguistics.

Vadim Shchukin, Dmitry Khristich, and Irina Galinskaya. 2016. Word clustering approach to bilingual document alignment (WMT 2016 shared task). In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 740–744, Berlin, Germany. Association for Computational Linguistics.

S. Steingrimsson. 2023. A sentence alignment approach to document alignment and multi-faceted filtering for curating parallel sentence pairs from web-crawled data. In *Proc 8th WMT*, pages 366–374.

Brian Thompson and Philipp Koehn. 2020. Exploiting sentence order in document alignment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5997–6007, Online. Association for Computational Linguistics.

Jakob Uszkoreit, Jay Ponte, Ashok Popat, and Moshe Dubiner. 2010. Large scale parallel document mining for machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1101–1109, Beijing, China. Coling 2010 Organizing Committee.

D Vose. 2000. Risk analysis: a quantitative guide. John Wiley & Sons.