Robust Fragment-Based Framework for Cross-lingual Sentence Retrieval

Nattapol Trijakwanich*, Peerat Limkonchotiwat[*], Raheem Sarwar[‡],

Wannaphong Phatthiyaphaibun[*], Ekapol Chuangsuwanich[†], Sarana Nutanong[*]

[*]School of Information Science and Technology, VISTEC, Thailand
[‡]RGCL, University of Wolverhampton, United Kingdom
[†]Department of Computer Engineering, Chulalongkorn University, Thailand
{nattapol.t_s17,peerat.l_s19}@vistec.ac.th
{wannaphong.p_s21,snutanon}@vistec.ac.th
R.Sarwar4@wlv.ac.uk, ekapolc@cp.eng.chula.ac.th

Abstract

Cross-lingual Sentence Retrieval (CLSR) aims at retrieving parallel sentence pairs that are translations of each other from a multilingual set of comparable documents. The retrieved parallel sentence pairs can be used in other downstream NLP tasks such as machine translation and cross-lingual word sense disambiguation. We propose a CLSR framework called Robust Fragment-level Representation (RFR) CLSR framework to address Out-of-Domain (OOD) CLSR problems. In particular, we improve the sentence retrieval robustness by representing each sentence as a collection of fragments. In this way, we change the retrieval granularity from the sentence to the fragment level. We performed CLSR experiments based on three OOD datasets, four language pairs, and three base well-known sentence encoders: m-USE, LASER, and LaBSE. Experimental results show that RFR significantly improves the base encoders' performance for more than 85% of the cases.

1 Introduction

Parallel corpora are essential for many NLP tasks in terms of both quality and quantity (Yang et al., 2019). Tasks like machine translation (Escolano et al., 2021; Zhang et al., 2020), cross-lingual word sense disambiguation (Mahendra et al., 2018; Bevilacqua and Navigli, 2020), and annotation projection (Sluyter-Gäthje et al., 2020) require a substantial amount of high-quality parallel sentences to construct accurate models. Traditionally, creating large-high-quality parallel corpora requires enormous manual effort from human annotators or translators. There are two approaches to reduce such human effort: (i) Using an unsupervised learning method to reduce the reliance on parallel corpora (Artetxe et al., 2018; CONNEAU and Lample, 2019; Kvapilíková et al., 2020). (ii) Using a Cross-lingual Sentence Retrieval (CLSR) method to automate finding parallel sentences. While the first approach may completely avoid using parallel corpora altogether through unsupervised learning, experimental results show that incorporating parallel sentences into the training process improves the model's performance. That is, parallel corpora still play a critical role even when employing unsupervised learning. Consequently, we focus our research attention on the latter approach.

Given a collection Q of query sentences q in one language $L1$ and another collection T of target sentences t in a different language $L2$, CLSR aims to find actual parallel pairs ($q \in Q$, $t \in T$) where q and t are translation sentences of each other. In real-world scenarios, parallel sentences are mined from comparable corpora. Consequently, not every q has a corresponding t and vice versa; we consider such sentences *non-pairing*. An effective CLSR method has to identify parallel pairs (q, t) from many non-pairing sentences. As the number of non-pairing sentences increases, there are more distractors to actual parallel pairs, and the robustness of the method becomes critical.

A popular CLSR approach constructs an embedding space using a *multilingual sentence encoder* (encoder for short) to organize sentences from different languages according to the meanings. Well-known methods utilizing this approach include m-USE (Yang et al., 2020), LASER (Artetxe and Schwenk, 2019b), and LaBSE (Feng et al., 2020). For robustness, CLSR methods generally include a filtering mechanism to avoid including non-pairing sentences into the results.

Using raw scoring from the encoder and hard threshold to filter out non-pairing sentences suffers from globally similarity score inconsistency. To improve the filtering robustness, more sophisticated re-scoring mechanisms have been studied. Artetxe and Schwenk (2019a) proposed a filtering mechanism based on variations of margin-based scorers. Their method considers the margin between a query sentence and its k-nearest neighbor based on a for-

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 935–944
November 7–11, 2021. ©2021 Association for Computational Linguistics

ward and backward search using the cosine similarity function. Yang et al. (2019) found that only a forward search also obtained a comparable performance to that of Artetxe and Schwenk (2019a). They also proposed a BERT-based re-scoring function, which substantially improved the accuracy. The methods mentioned above can robustly filter out non-pairing sentences and can accurately identify sentence pairs in in-domain data. However, their performance significantly drops when applied to Out-of-Domain (OOD) test samples.

We compares results of the base multilingual sentence encoders for in-domain and Out-of-Domain (OOD) scenarios. BUCC (Zweigenbaum et al., 2018) is a standard corpus for CLSR task. In contrast, JW300 (Agić and Vulić, 2019) is constructed from religious-society magazines that are less formal. LASER was trained on formal documents such as Europarl, and United Nation parallel data, while LaBSE used Wikipedia data for training. Thus, we consider JW300 as an OOD dataset for LASER and LaBSE. Note that m-USE did not provide results on BUCC. Results from Table 1 show that both methods perform worse when evaluated on the OOD dataset. For JW300, we used the same settings as described in Section 3.

Dataset	BUCC		JW300	
	FR	DE	FR	DE
LASER (Artetxe and Schwenk, 2019b)	93.9	96.2	75.3	73.7
LaBSE (Feng et al., 2020)	88.7	92.5	70.8	68.9

Table 1: Comparison of retrieval performance for in- and out-of-domain scenarios.

In this paper, we propose a **R**obust **F**ragment-level **R**epresentation (RFR) framework to improve the CLSR robustness when applied to OOD scenarios. The crux of our solution lies in the n-grams sliding window mechanism, which breaks up each sentence into multiple vectors (called fragments) to allow for phrase matching at the subsentence level. To avoid accidental matching, i.e., pairing similar fragments from sentences with different meanings, we also equip each fragment with a traditional sentence encoding. Since different fragments from the same sentence can now be associated with fragments from various sentences, we also propose a process to combine results from multiple fragment matchings to form one single final output for each sentence.

To assess the effectiveness of our solution, we conducted experimental studies on three datasets, which were all OOD with respect to the base and proposed methods. For each dataset, we used two rich-resource language pairs, French-English and German-English, as well as two limited resource language pairs, Arabic-English and Thai-English. We used three well-known encoders as our base encoders, namely m-USE, LASER, and LaBSE. We also implemented a proposed solution on top of each base, namely RFR-m-USE, RFR-LASER, and RFR-LaBSE. The combination of three datasets, four language pairs, and three bases methods formed 36 comparisons in total. Experimental results show that our proposed solution could significantly enhance the performance of the base encoders in 32 out of 36 comparisons. In addition, we also applied our framework to a cross-lingual QA dataset. Our method consistently improves the accuracy of m-USE$_{\mathrm{QA}}$, a well-known encoder for cross-lingual answer retrieval.

The summary of our contributions is as follows: (i) We propose a novel sentence representation model representing each sentence as a collection of fragments. (ii) We propose a novel fragment-level CLSR framework that enhances robustness to base encoders. (iii) We demonstrate significant improvement of our framework on all base encoders via extensive experimental studies.

2 Proposed Framework

We first provide an overview of our RFR framework in Figure 1. It consists of three main components: (i) preprocessing, (ii) similarity search, and (iii) prediction aggregation.

Preprocessing. The preprocessing step transforms each sentence into multiple fragments, where each fragment is represented as a vector. For each sentence s, we first remove all punctuations[1] and represent each word as a token (w_j^s)[2] where j is the word index. Then, a sliding window is applied to generate a collection of n-grams. We call these n-grams sentence fragments (f_i^s) where i indicates the token index. We then encode each fragment using an encoding function $g(\cdot)$ in to a vector (e_i^s). The encoding function can be from any multilingual encoder mentioned earlier. We also append a sentence-level encoding vector (e_s^s) to form a final representation ($[e_i^s : e_s^s]$). We shall refer to this collection of preprocessed fragments as the database.

[1] To clarify, since all encoder used are based on SentencePiece, removing punctuation does not generate UNK tokens.

[2] For languages with no explicit word boundaries, such as Thai, we used the word tokenizer provided in Wannaphong Phatthiyaphaibun (2016)

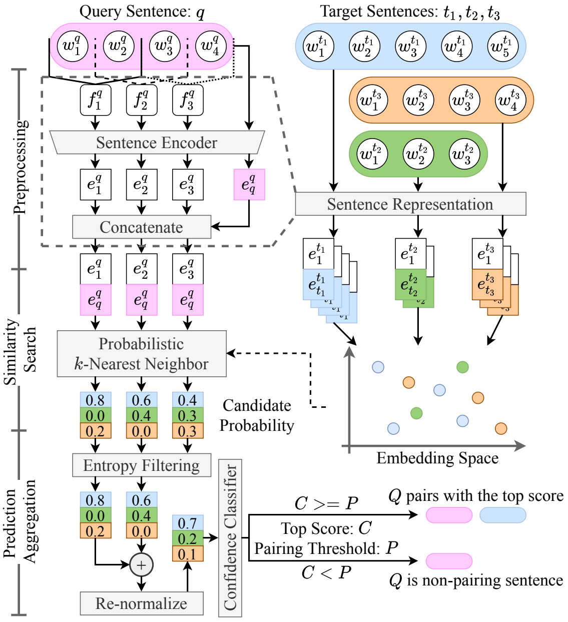


Figure 1: Overview of our method. Query sentence has 4 words, and there are 3 sentences in target corpus which have 5, 3, and 4 words. The sentences are split into fragments and encoded into embeddings for the retrieval process.

Note that in our ablation studies, the addition of sentence-level information significantly improved the performance (Table 5 in Appendix 4.3).

Similarity Search. The next step is to perform a search for similar sentences using Probabilistic k-Nearest Neighbor (PkNN). Given a query sentence q in $L1$, we apply the same sentence fragmentation and representation process as the previously described preprocessing step. In this way, a query sentence q is represented as a collection of fragment vectors. We perform a similarity search on each query fragment independently. Each instance of the similarity search returns a set of k similar target sentence fragments retrieved from the database in $L2$. Treating the sentence id as the class label for each fragment in the database, we use PkNN to compute the probability of each query fragment belonging to each $L2$ sentence. By this means, we effectively transform the problem of target sentence identification into an instance-based learning problem. We choose this learning paradigm due to the following reasons: (i) There is no need to construct a classification model; inference can be conducted by finding similar instances in the database. (ii) As new instances are added to the database, there is no need to reconstruct a classification model. (iii) The PkNN method is non-parametric; hence, we do not need any prior knowledge of the probability distribution.

Prediction Aggregation. To get the final score for the query sentence, we aggregate the probabilities from each query fragment. Since the prediction from each query fragment can be noisy, we first filter uncertain fragments to keep only $p\%$ of the fragments. The filtering is based on the entropy value calculated from the predicted probability mass function. Common n-grams that may be matched to many $L2$ sentences should be discarded in this step. After filtering, we sum all the probability scores together and re-nomalize. To account for the case where there is no actual translation pair, a final filtering is applied by simple thresholding. If the probability value is higher than the pairing threshold P then the query sentence pairs with the top scoring sentence. Otherwise, there is no actual translation pair present.

Hyperparameters	**Range** ([start, stop], step)
n	6
k	([5, 50], 5)
β	([50, 100], 5)
p	([0.1, 1], 0.1)
P	([0, 1], 0.1)

Table 2: Parameter ranges for the parameter tuning process.

3 Experiment Setup and Datasets

This section describes the parameter tuning and two experimental studies: (i) Cross-lingual Sentence Retrieval (CLSR); (ii) Cross-lingual Document Retrieval for Question Answering (CLQA). We use McNemar's test with $p < 0.001$ to establish statistical significance. The competitive methods and datasets used in each study are presented as follow.

Parameter Tuning. We denote n-grams for fragment size (n), PkNN neighbors (k), PkNN spiking coefficient (β), top % min-entropy filter (p), and pairing threshold (P) as parameters to be tuned in all experiments. We set n equal to 6 for all experiments after the preliminary experiment. k and β are tuned for efficient similarity search. The latter parameters are tuned for F1.

All parameters are tuned using a tuning set according to each experiment. The final parameter values depend on the corpora, and the hyperparameter searches were performed using the ranges given in Table 2.

The size of the fragments can be treated as a hy-

Method	JW300 (F1)				QED (F1)				TED2020 (F1)				Average F1				Improvement (F1 Gap)			
	FR	DE	AR	TH	FR	DE	AR	TH	FR	DE	AR	TH	FR	DE	AR	TH	FR	DE	AR	TH
m-USE	55.5	48.3	13.5	8.5	56.9	52.6	21.9	6.7	64.6	59.2	21.9	3.6	59.0	53.4	19.1	6.3	—	—	—	—
LASER	75.3	73.7	65.1	53.3	68.4	68.6	71.8	71.9	73.3	75.6	74.0	73.3	72.3	72.6	70.3	66.2	—	—	—	—
LaBSE	70.8	68.9	40.6	30.7	65.4	64.7	48.6	44.7	72.8	72.9	57.9	44.8	69.7	68.8	49.0	40.1	—	—	—	—
RFR-m-USE	78.4	84.1	59.7	**63.5**	**79.6**	73.4	71.8	**76.5**	88.8	87.5	81.8	**84.8**	82.3	81.7	71.1	**74.9**	**23.3**	**28.3**	**52.0**	**68.7**
RFR-LASER	81.6	81.0	65.8	61.2	56.2	65.9	71.2	69.7	87.4	83.8	80.8	84.0	75.1	76.9	72.6	71.6	2.7	4.3	2.3	5.5
RFR-LaBSE	**88.2**	**87.9**	**76.8**	47.6	77.5	**76.4**	**78.9**	69.4	**92.6**	**90.4**	**89.9**	59.8	**86.1**	**84.9**	**81.9**	58.9	16.4	16.1	32.8	18.9

Table 3: F1 score for the CLSR task on various language pairs (XX → EN)

perparameter that can be tweaked. When n=1, fragments become sets of single words. From our preliminary experiments, the results were best when n=6. Thus, n=6 were used for all settings. In addition, when n equals the number of words in the sentence, fragments become a full sentence which are the base encoder results in Table 3

CLSR — Competitive Methods. We selected three well-known multilingual sentence encoders as base encoders: m-USE (Yang et al., 2020), LASER (Artetxe and Schwenk, 2019b), and LaBSE (Feng et al., 2020). Using these base encoders, we formulated three competitive methods by applying the margin-based ratio rescoring function from Artetxe and Schwenk (2019a) and fine tuning the threshold for each of them accordingly. For each base encoder, we applied our method and called them RFR-m-USE, RFR-LASER, and RFR-LaBSE, respectively.

CLSR — Datasets. We evaluated our method on a CLSR task with three Out-of-Domain (OOD) datasets: JW300 (Agić and Vulić, 2019), QED (Abdelali et al., 2014), and TED2020 (Reimers and Gurevych, 2020) from Opus (Tiedemann, 2012). For each dataset, we sampled 1,000 sentences for both query and target corpus for a test set. The number of sentences in the test set represents the length of documents in a real-world setting. We additionally sampled 100 sentences in total for tuning hyperparameters as a tuning set. We set the number of actual parallel pairs to 50% of the total number of sentences unless stated otherwise. The non-pairing sentences were randomly selected from the remaining sentences in the corpus for both query and target datasets.

CLQA — Competitive Method. As a base encoder, we used m-USE$_{\text{QA}}$ (Yang et al., 2020), an m-USE variation that supports CLQA. To form a competitive method, we applied the same filtering mechanism as the CLSR competitive methods.

CLQA — Dataset. We choose Xquad (Artetxe et al., 2019), a benchmark dataset for evaluating cross-lingual question answering performance. The Xquad is also considered OOD for all base sentence encoders. Question sentences were used as query sentences to retrieve documents or paragraphs that contain the answer. Either target paragraphs or documents functioned as a target collection. We split each target paragraph/document into fragments disregarding sentence boundaries. Thus, the target documents or paragraphs differ greatly in length. We used the entire Xquad as the test set with no non-pairing questions. To tune the parameters of the retrieval methods, we used TED2020.

Method	Doc-level(F1)			Para-level(F1)		
	DE	AR	TH	DE	AR	TH
m-USE$_{\text{QA}}$	85.3	74.2	80.0	71.0	59.5	64.5
RFR-m-USE$_{\text{QA}}$	**85.4**	**80.9**	**86.3**	**75.3**	**71.9**	**73.0**

Table 4: Performance on the Xquad dataset

4 Experimental Results

4.1 CLSR Results

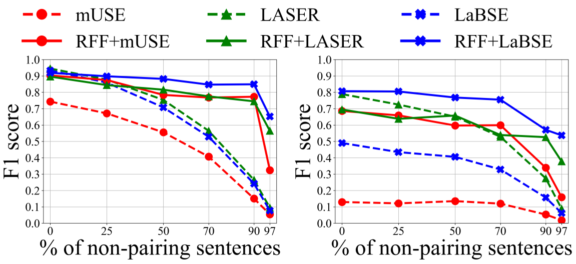
Table 3 presents results from CLSR experiments on the three datasets. For each dataset, there are four language pairs (XX→ English) where XX denotes the query language which can be French (FR), German (DE), Arabic (AR), and Thai (TH). The first two represent rich-resource language pairs, and the rest represent limited resource ones.

The best performer for each language-dataset combination was either RFR-m-USE or RFR-LaBSE. On average the proposed RFR framework improves over the baseline embedding methods. Although all methods were optimized for F1, the RFR framework greatly improves precision while sacrificing some recall (see Appendix A.6). This is preferable for mining high-quality sentence pairs. Matching fragments helps to increase precision because every fragment has to have a matching pair. Two long sentences with very similar overall content can have small differences in some clauses (see the third set of examples in Figure 4) Note that on QED, RFR-LASER took a large hit to recall lowering the F1 score compared to the baseline.

Method	Concat	Entropy Filter	JW300 (F1)				QED (F1)				TED2020 (F1)				Average F1
			FR	DE	AR	TH	FR	DE	AR	TH	FR	DE	AR	TH	
m-USE	-	-	55.5	48.3	13.5	8.5	56.9	52.6	21.9	6.7	64.6	59.2	21.9	3.6	34.4 ± 23.6
RFR-m-USE	No	Yes	56.5	64.3	51.4	37.9	68.7	61.7	57.7	59.0	83.8	69.9	74.6	58.1	62.0 ± 11.7
RFR-m-USE	Yes	No	84.1	83.2	59.7	**63.5**	79.3	74.4	69.9	73.9	89.7	77.0	88.1	**86.2**	77.4 ± 9.6
RFR-m-USE	Yes	Yes	78.4	84.1	59.7	**63.5**	**79.6**	73.4	71.8	**76.5**	88.8	87.5	81.8	84.8	77.5 ± 9.1
LASER	-	-	75.3	73.7	65.1	53.3	68.4	68.6	71.8	71.9	73.3	75.6	74.0	73.3	70.4 ± 6.2
RFR-LASER	No	Yes	49.4	26.1	29.6	40.0	40.8	58.9	52.2	35.7	69.3	71.7	64.7	48.4	48.9 ± 15.1
RFR-LASER	Yes	No	82.7	79.9	65.4	59.1	56.3	65.9	70.0	69.3	87.6	83.9	79.4	84.7	73.7 ± 10.7
RFR-LASER	Yes	Yes	81.6	81.0	65.8	61.2	56.2	65.9	71.2	69.7	87.4	83.8	80.8	84.0	74.1 ± 10.3
LaBSE	-	-	70.8	68.9	40.6	30.7	65.4	64.7	48.6	44.7	72.8	72.9	57.9	44.8	56.9 ± 14.4
RFR-LaBSE	No	Yes	74.0	74.0	69.6	44.6	72.9	72.8	58.3	60.5	89.1	79.2	76.9	70.3	70.2 ± 11.4
RFR-LaBSE	Yes	No	83.5	87.4	**80.5**	54.5	77.0	**81.2**	78.2	51.3	**92.9**	**90.8**	**89.9**	57.7	77.1 ± 14.6
RFR-LaBSE	Yes	Yes	**88.2**	**87.9**	76.8	47.6	77.5	76.4	**78.9**	69.4	92.6	90.4	**89.9**	59.8	**78.0** ± **13.6**

Table 5: Performance comparisons in our ablation experiment.

Effect of the Non-pairing Sentences Percentage. We also varied the percentage of non-pairing sentences over all sentences to evaluate the RFR framework's robustness against an increasing number of non-pairing sentences. We started from 0% of non-pairing sentences, and then we replaced some actual parallel pairs with non-pairing sentences in both query and target corpus. The F1 scores were measured for each step of replacement. Figure 2 confirms our framework's robustness against high amounts of non-pairing sentences. More analysis details are provided in Appendix A.



(a) JW300: FR → EN

(b) JW300: AR → EN

Figure 2: F1 score as the number of non-pairing sentences increases

4.2 CLQA Results

This study aims to show the flexibility of our framework in other query-based tasks, namely cross-lingual document/paragraph retrieval for QA. We used m-USE$_{\text{QA}}$ as a base encoder and used paragraphs and documents as input and context for the m-USE$_{\text{QA}}$ respectively. Our framework has to retrieve a document or paragraph that contains an answer to the query question sentence in the Xquad dataset in this task.

Results from Table 4 show that our framework improves m-USE$_{\text{QA}}$'s performance in all cases with 4.4% and 8.4% improvement on average for document- and paragraph-level, respectively.

4.3 Ablation Studies

We performed ablation studies to determine the importance of each step in our proposed framework. The results are summarized in Table 5.

Whole Sentence Embedding Concatenation. As discussed in Section 2, the fragment embedding is concatenated with the whole sentence embedding. We compared the results with and without the sentence embedding. The results show that the sentence embedding improves the performance for all cases.

Entropy Filter. An entropy filter is used to filter unpromising fragment candidates out from the aggregation step. We compared the results with and without our filtering mechanism to validate the importance of the entropy filter. The overall results show a slight improvement in the average performance with lower standard deviations.

5 Conclusion and Future Work

We propose a novel sentence representation model representing each sentence as a collection of fragments for query-related tasks. Our CLSR framework can enhance the robustness of any pretrained multilingual sentence encoder. Extensive experiments on four pairs of rich- and low-resource languages show that our method significantly improves over the base encoders. We also demonstrated the usefulness of our framework on document retrieval for question-answering in three languages and obtained improvements in all cases. For future work, we would like to explore the possibility of returning sub-sentence matching in order to improve the recall of our framework.

References

Ahmed Abdelali, Francisco Guzman, Hassan Sajjad, and Stephan Vogel. 2014. The AMARA corpus: Building parallel language resources for the educational domain. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1856–1862, Reykjavik, Iceland. European Language Resources Association (ELRA).

Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *International Conference on Learning Representations*.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. *CoRR*, abs/1910.11856.

Mikel Artetxe and Holger Schwenk. 2019a. Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, Florence, Italy. Association for Computational Linguistics.

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

Michele Bevilacqua and Roberto Navigli. 2020. Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2854–2864, Online. Association for Computational Linguistics.

Alexis CONNEAU and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Carlos Escolano, Marta R. Costa-jussà, José A. R. Fonollosa, and Mikel Artetxe. 2021. Multilingual machine translation: Closing the gap between shared and language-specific encoder-decoders. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 944–948, Online. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding.

Ivana Kvapilíková, Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Ondřej Bojar. 2020. Unsupervised multilingual sentence embeddings for parallel corpus mining. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 255–262, Online. Association for Computational Linguistics.

Rahmad Mahendra, Heninggar Septiantri, Haryo Akbarianto Wibowo, Ruli Manurung, and Mirna Adriani. 2018. Cross-lingual and supervised learning approach for Indonesian word sense disambiguation task. In *Proceedings of the 9th Global Wordnet Conference*, pages 245–250, Nanyang Technological University (NTU), Singapore. Global Wordnet Association.

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*. Association for Computational Linguistics.

Henny Sluyter-Gäthje, Peter Bourgonje, and Manfred Stede. 2020. Shallow discourse parsing for under-resourced languages: Combining machine translation and annotation projection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1044–1050, Marseille, France. European Language Resources Association.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Charin Polpanumas Arthit Suriyawongkul Lalita Lowphansirikul Pattarawat Chormai Wannaphong Phatthiyaphaibun, Korakot Chaovavanich. 2016. PyThaiNLP: Thai Natural Language Processing in Python.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2020. Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 87–94, Online. Association for Computational Linguistics.

Yinfei Yang, Gustavo Hernandez Abrego, Steve Yuan, Mandy Guo, Qinlan Shen, Daniel Cer, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Improving multilingual sentence embedding using bidirectional dual encoder with additive margin softmax. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5370–5378. International Joint Conferences on Artificial Intelligence Organization.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2018. A Multilingual Dataset for Evaluating Parallel Sentence Extraction from Comparable Corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

A Appendices

A.1 Runtime

The experiments were conducted on Intel Xeon Gold 5222 CPU @ 3.80GHz running on Ubuntu 18.04.03 and 188 GB RAM. All the methods were implemented in Python, and their running time are provided in Table 6. Note that the tuning time was included.

Method	JW300 (seconds)			
	FR	DE	AR	TH
m-USE	31	30	32	32
LASER	16	16	36	24
LaBSE	155	150	152	199
RFR-m-USE	1172	1145	1106	1198
RFR-LASER	1126	1207	1272	1401
RFR-LaBSE	2572	2465	2761	3266

Table 6: Running time, including tuning and testing, in seconds

A.2 Additional Results Non-pairing Sentences Experiments

Here we provide additional results for Figure 2 in the main text. Figure 3 shows results for all language pairs on JW300. The same trend occurs where every method performs worse when the number of non-pairing sentences increases. However, our method outperforms the baselines especially when the number of non-pairing sentence is high.

A.3 Effect of Size of Tuning Set

In this experiment, we want to study how the size of the tuning data affects the performance. The training size is set to 50, 100, 200 sentences with 50% actual translations available with the rest of the setup are same. We selected m-USE as the base in this experiment.

Results from Table 7 show that our result improves as the tuning size increases.

A.4 Robustness to Different Tuning Sets

In this experiment, we consider how different tuning sets can affect the tuning and the final results. We created 10 different tuning sets to perform our experiments. The average F1 scores and standard deviations are shown in Table 8.

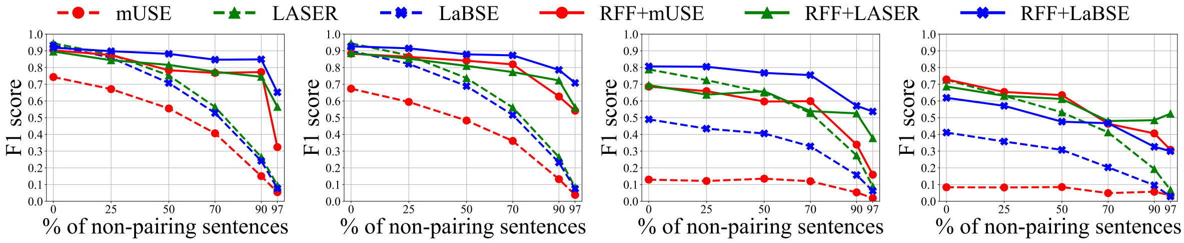
A.5 Error Analysis

To better understand our framework, various types of failure cases are shown in Figure 4. False positives are aligned pairs that are not in the gold pairs. False negatives are gold pairs not identified by our framework. All items in Figure 4 are picked from Thai to English pairs with LASER base embeddings.

The false positives identified by our framework can be caused by the filtered out fragment. The address portion of the sentence was filtered out causing an incorrect match. This, however, opens up the possibility of clause level matching with this framework. For false negatives, some of them are from incorrect ground truth pairs presented in the dataset. Our method does not perform well on shorter sentences because they tend to have lower pairing probability values, and thus filtered out by the pairing threshold. Some normalization based on the sentence length might be required to alleviate this effect. The false positives from non-fragment-based methods can be from sentence pairs that are very similar except for a few words. This is due to the limitation of sentence embedding that only broadly captures the meaning of the entire sentence.

A.6 Precision and Recall Breakdown

Table 9 shows the precision and recall on QED which is the only dataset where our method did not improve the base LASER embeddings. While our method greatly improved precision, our recall also dropped significantly. This, however, is not always the case in other embeddings.



(a) JW300: FR → EN (b) JW300: DE → EN (c) JW300: AR → EN (d) JW300: TH → EN

Figure 3: F1 score as the number of non-pairing sentences increases

Tuning Set Size	JW300 (F1)				QED (F1)				TED2020 (F1)				Average F1
	FR	DE	AR	TH	FR	DE	AR	TH	FR	DE	AR	TH	
50 sentences	80.9	82.9	53.6	57.5	71.6	70.7	74.1	71.5	92.5	76.2	**88.6**	86.7	75.6 ± 11.8
100 sentences	78.4	84.1	59.7	63.5	**79.6**	73.4	71.8	**76.5**	88.8	87.5	81.8	84.8	77.5 ± 9.1
200 sentences	**83.9**	**84.5**	**65.3**	**68.3**	71.4	**75.7**	**76.3**	71.1	**92.7**	**89.9**	88.5	**86.9**	**79.5 ± 9.3**

Table 7: The F1 performance for different size tuning set (XX→EN).

Method	JW300 (F1)			
	FR	DE	AR	TH
RFR-m-USE	83.18 ± 3.80	80.09 ± 3.25	62.63 ± 2.48	66.51 ± 3.33
RFR-LASER	79.67 ± 2.66	79.60 ± 3.05	61.90 ± 3.60	61.03 ± 2.19
RFR-LaBSE	87.35 ± 3.35	87.84 ± 2.08	77.13 ± 2.94	52.80 ± 7.24

Table 8: Performance statistics for different tuning subsets.

Method	QED (Precision, Recall, F1)											
	FR			DE			AR			TH		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
m-USE	48.3	69.2	56.9	45.1	63.2	52.6	23.0	21.0	21.9	8.5	5.6	6.7
LASER	56.1	**87.6**	68.4	56.6	**87.0**	68.6	60.5	**88.2**	71.8	61.0	**87.6**	71.9
LaBSE	53.5	84.0	65.4	53.4	82.0	64.7	42.9	56.0	48.6	42.7	46.8	44.7
RFR-m-USE	92.3	70.0	**79.6**	92.9	60.6	73.4	93.3	58.4	71.8	**86.2**	68.8	**76.5**
RFR-LASER	**94.3**	40.0	56.2	90.0	52.0	65.9	93.2	57.6	71.2	75.6	64.6	69.7
RFR-LaBSE	**94.3**	65.8	77.5	**95.2**	63.8	**76.4**	**99.1**	65.6	**78.9**	79.4	61.6	69.4

Table 9: Performance breakdown on the QED dataset.

RFR-LASER False Positives Cases	
Query Sentence	หากคุณยินดีจะรับข้อมูลเพิ่มเติมหรือยินดีให้ใครสักคนมาเยี่ยมเพื่อนำการศึกษาคัมภีร์ไบเบิลกับคุณที่บ้าน โดยไม่คิดมูลค่า โปรดเขียนถึงพยานพระยะ โฮ วา 69 / 1 สุขุมวิท ซอย 2 กรุงเทพฯ 10110 หรือตามที่อยู่ที่เหมาะสม ในหน้า 2.
Translation	If you are willing to get more information or are happy to visit someone to bring the Bible study with you at home without thinking, please write to Jehovah's Witnesses 691 Sukhumvit Soi 2, Bangkok 10110 or according to the appropriate address inPage 2
False Pair	If you would welcome further information or would like to have someone call at your home to conduct a free Bible study with you , please write to Watchtower , 25 Columbia Heights , Brooklyn , NY 11201 - 2483 , or to the appropriate address listed on page 2 .
Query sentence	▫ พระเยซูทรงหมายความอย่างไรเมื่อพระองค์ตรัสว่า “ เราเป็น . . .
Translation	▫ What did Jesus mean when he said, "I am.
False Pair	(b) What does the name Jesus mean , and how did God’s Son live up to his name ?
RFR-LASER False Negatives Cases	
Query Sentence	ผลจึงเป็นเช่นเดียวกับคนตัวสูงอยู่ในห้องที่มีเพดานต่ำ — คือความเจ็บปวด .
Translation	The result is the same as a tall person in a room with low ceilings pain.
Incorrect Ground Truth	Trigger points can refer pain anywhere in the body ; one in the shoulder can cause severe pain on the side of the head , mimicking migraines . . . .
Query Sentence	ฉันควรทำศัลยกรรมเสริมสวยไหม?
Translation	Should I have cosmetic surgery?
Missed Pair	Should I Have Cosmetic Surgery ?
LASER False Positive Cases	
Query Sentence	ยิ่งคุณพยายามบังคับผู้ป่วยมากเท่าใดการต่อสู้ก็จะยิ่งยืดเยื้อมากเท่านั้น.
Translation	The more you try to force the patient, the more prolonged the battle will be.
False Pair	But the more I had sexual relations , the more insecure I felt . ”
Query Sentence	ความผูกพันรักใคร่อาจเกิดขึ้นได้เช่นกันเมื่อหนุ่มสาวเขียนจดหมายติดต่อกับคนเหล่านั้นซึ่งเป็นที่รู้จักกันว่าไม่เป็นตัวอย่างที่ดีในฐานะคริสเตียน.
Translation	Affectionate attachment can also arise when a youth writes correspondence with those who are known to be impractical as Christians.
False Pair	A Christian may begin to have romantic feelings for someone who does not love Jehovah , thinking that a suitable mate cannot be found among true Christians .

Figure 4: Example failure cases chosen from the Thai-English pair.