

Semi-Supervised Multilingual Alignment with Lexical Memory for Massively Parallel Text Mining

Weitai Zhang^{*†}, Peiwan Tang[†], Chao Lin[†], Simran Naagar^{*}, Zhongyi Ye^{*†}, Junhua Liu^{*†}

^{*}University of Science and Technology of China, Hefei, China

[†]iFLYTEK Research, Hefei, China

{zwt2021, Simranngr, yzy5630}@mail.ustc.edu.cn, {pwtang, chaolin3, jhliu}@iflytek.com

Abstract—Existing state-of-the-art techniques that employ multilingual sentence embeddings for mining parallel texts predominantly rely on extensive supervision, which often results in sub-optimal performance in the absence of large-scale parallel training datasets. In this study, we introduce a novel method designed to extract high-quality parallel texts from monolingual corpora, particularly targeting zero-and low-resource languages. We learn language-agnostic sentence embeddings with a two-tiered training regimen: an initial phase of lexical knowledge-enhanced pretraining and a subsequent phase of supervised fine-tuning on a minimally sized parallel dataset using contrastive loss. Furthermore, we enhance the model’s performance by adopting an iterative training methodology that leverages both mined data and synthetically augmented data. We illustrate the capability of our method to create high-quality parallel text with various downstream tasks. All results suggest that the proposed method is effective and can surpass previous state-of-the-art supervised methods in zero-and low-resource scenarios.

Index Terms—Parallel corpus mining, Lexical knowledge, Sentence embedding, Contrastive loss, Semi-supervised learning

I. INTRODUCTION

Parallel text is a bilingual form of text that plays an indispensable role in multilingual natural language processing tasks, especially in machine translation (MT). With support from large datasets, MT has made significant progress recently [1]–[4]. However, out of over 7,000 languages worldwide, only a few have large-scale and high-quality parallel datasets [5], severely limiting the deployment of MT in zero-resource and low-resource language pair scenarios [6], [7]. Technologies such as leveraging the structural features of web documents [8], [9], n-gram scoring systems [10], information retrieval techniques [11], and machine translation [12], [13] have proven effective in obtaining high-quality parallel sentences from multilingual websites. However, these methods still face numerous limitations in cross-language and domain scalability.

Some research has explored using language-agnostic sentence embeddings based on large-scale cross-lingual pretraining techniques [14]–[18] to extract parallel sentences from non-parallel corpora. They align translation pairs from two different languages by calculating similarity scores within a cross-lingual embedding space [19]. While language pairs without any explicit parallel training data may benefit from transfer learning, the quality of mined corpus decreases with a notably rapid decline in zero-and low-resource scenarios [20].

Moreover, although unsupervised cross-lingual representation learning on unpaired sentences still underperforms compared to supervised learning [21], [22], these studies indicate that language-agnostic sentence embeddings can benefit from the unsupervised training of monolingual data and supervised training of parallel sentences. By employing intricate augmentations after training to more effectively extract deep linguistic features, these unsupervised models can generate significantly improved multilingual alignment representations [23].

In this study, we introduce a robust framework to aggregate high-quality monolingual and bilingual corpora for application in zero-

resource and low-resource languages. First, in contrast to previous methods that primarily relied on platforms like CommonCrawl¹ [19], [20] to obtain monolingual data, we implemented a more complex monolingual corpus construction pipeline based on a search-oriented web document collection program. Second, to overcome the observed performance decline of language-agnostic sentence embeddings in zero-resource and low-resource languages, we propose a semi-supervised multilingual alignment pretrained cross-lingual language model A that incorporates lexical memory. The model employs a two-phase training process to optimize different training signals separately. In the initial pretraining phase, the model uses cross-lingual dictionaries to facilitate the generation of language-agnostic sentence embeddings. In the fine-tuning phase, the model minimizes sentence representation differences through contrastive loss, even when only using a small amount of supervised training data. Next, we iteratively train on mined and translated corpora, achieving further performance improvement through strategic selection of training samples. We evaluated our model on extensive dual-text retrieval tasks, such as BUCC [24] for resource-rich languages and Tatoeba [19] for resource-limited languages. Our approach achieved remarkable results, surpassing the performance of previous supervised methods for low-resource languages. Finally, experiments including IWSLT2014 and CCMT2019 demonstrated the effectiveness of the parallel texts extracted by our method in enhancing downstream machine translation tasks. Translation models developed using our sentence pair extraction method also outperformed those trained on traditional official datasets across multi-domain test sets.

II. METHODOLOGY

A. Monolingual Dataset Creation and Description

Large monolingual corpora often allow for the extraction of larger parallel corpora, but collecting unlabeled monolingual sentences for low-resource languages on the internet is always very challenging. Therefore, we propose a search-based web document collection program aimed at enriching monolingual corpora for zero-resource and low-resource languages. As shown in Figure 1, this method achieves its goal by preprocessing data obtained from CommonCrawl and compiling query seeds for the target languages. The CommonCrawl corpus covers over 2 billion web pages and includes content from a wide range of domains and in multiple languages. First, we collected approximately 261.5 billion pages from the CommonCrawl corpus, specifically content published from January 2021 to December 2021. Next, we generated up to tens of thousands of query seeds for each target language, which typically consist of various n-grams. Using these query seeds, we conducted searches on search engines such as Google.com and Bing.com to identify large-scale websites related to each language. By counting the number of web pages retrieved for

¹<https://commoncrawl.org>

each query, we were able to systematically evaluate and determine the most relevant large-scale websites. In the final stage, we used open-source web crawlers to deeply scrape these identified websites to collect more web documents. Experiments have shown that this method is relatively efficient in resource utilization.

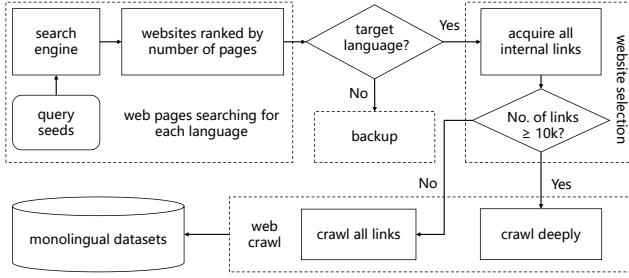


Fig. 1. Search-based web document collection procedure, particularly for zero- and low-resource languages.

B. Cross-lingual Pretraining with Lexical Memory

Several recent studies [21]–[23] utilize pre-trained cross-lingual MLM to derive multilingual sentence embeddings relying only on monolingual data. Although they claim the effectiveness of mining parallel corpora using unsupervised bilingual word embeddings, the performance typically lags far behind supervised methods [19], [20].

Our approach differs from the above in that we introduce cross-lingual lexicons as auxiliary training signals. As illustrated in Figure 2, we train the cross-lingual MLM on a concatenation of two sentences in different languages to learn a joint structure of these two languages together. Random tokens are masked from both sentences and the model is trained to fill in the blanks by attending to any of the tokens of the two sentences. Comparing to the standard cross-lingual transformer architecture, we add one external memory module with an lexical-attention layer in each encoder block. The external memory module aims to encode each bilingual token pair (X_i, Y_j) into a key and value pair (K_i, V_j) . The encoding of bilingual token pairs can be simply implemented using the average of word embeddings in the phrase, such as:

$$K_i = \frac{1}{|X_i|} \sum_{x_i \in X_i} f_{emb}(x_i) \quad (1)$$

$$V_j = \frac{1}{|Y_j|} \sum_{y_j \in Y_j} f_{emb}(y_j)$$

where f_{emb} denotes the embedding function, $|X_i|$ and $|Y_j|$ refer to the length of the bilingual phrases. Practically, using average of word embeddings increases less training parameters and has a higher training efficiency.

Additionally, we retrieve bilingual token pairs in the concatenated sequence as positive samples and randomly replace the target token to generate negative samples. By contrasting the representations of positive and negative samples in a discriminative manner, we train the model with an auxiliary adversarial loss defined as follows:

$$Loss = -\alpha \sum_{i=1}^M \log p_{\theta}(w_i | (W - M)) + \beta \sum_{j,k}^N L_{adv}^{j,k} \quad (2)$$

$$L_{adv}^{j,k} = -\log \left[\frac{\exp(j * k_+)}{\exp(j * k_+) + \sum_{i=1}^N \exp(j * k_-)} \right] \quad (3)$$

where W denotes the whole words, M denotes the masked words, j denotes the source side token, k_+ denotes the positive sample,

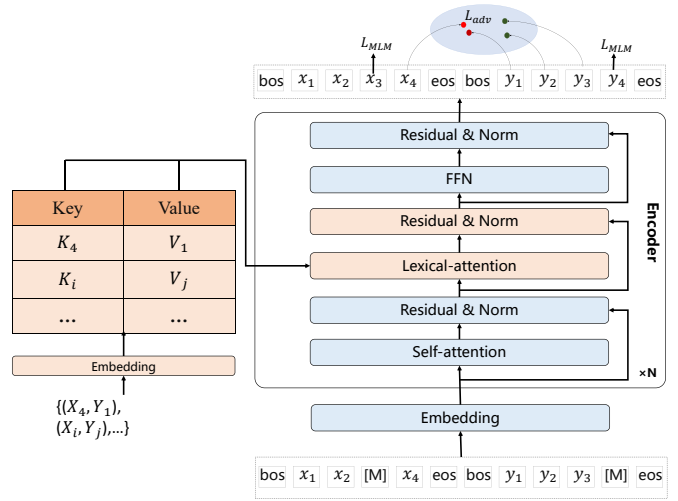


Fig. 2. Cross-lingual language model architecture of multilingual alignment with lexical memory. [M] denotes masked words.

and k_- denotes negative samples. The parameters are shared by all languages involved and updated on the concatenation of all training corpora. We further finetune the model with training data of desired language pairs which can encourage similarity between these two languages.

C. Finetuning with Sentence-level Contrastive Loss

Sentence embeddings can be composed of token representations by simple element-wise averaging. Even though mean-pooling of the encoded output is a naive approach, we find it's effective in our scenario and yields satisfying results. Inspired by the effectiveness of contrastive objectives in multilingual MT [25], we further derive language-agnostic sentence embeddings with a contrastive loss based on bilingual word embeddings. In our exploration focusing on zero- and low-resource language pairs, we only use this objective to finetune the pre-trained model obtained in Section II-B on extremely small parallel dataset. Formally, given a positive parallel sample (x_i, y_j) , we randomly choose a sentence y_k to form a negative sample (x_i, y_k) . The objective of contrastive learning is to minimize the following loss:

$$L_{ctr} = - \sum_{(x_i, y_j)} \log \left[\frac{e^{sim^+(R(x_i), R(y_j))}}{\sum_{y_k} e^{sim^-(R(x_i), R(y_k))}} \right] \quad (4)$$

where $sim(\cdot)$ calculates the similarity of different sentences and $R(\cdot)$ denotes the mean-pooling of encoded output. To obtain optimal cross-lingual sentence embeddings, the model will be optimized by jointly training with lexical signals in the finetuning stage.

D. Iterative Training with Data Augmentation

To enhance the performance of cross-lingual pretraining models, we adopt an iterative augmentation method using synthetic parallel sentences. First, we leverage cross-linguistic sentence embedding techniques to identify sentence pairs in multiple languages and use these pairs to train a multilingual machine translation system. Next, we translate monolingual sentences into other languages to generate synthetic parallel sentence pairs, and utilize these pairs to further train the cross-lingual masked language model (MLM). With each iteration, the training improves the translation capabilities of

TABLE I

LIST OF 15 LOW-RESOURCE LANGUAGES ALONG WITH THEIR TRAINING SIZE, TEST SIZE AND THE RESULTING SIMILARITY ACCURACY ON TATOEBE.

	ang	arq	arz	awa	csb	dsb	gsw	hsb	max	nov	orv	pms	swg	tzl	war	avg.
LASER	58.96	58.62	31.24	63.20	54.55	48.64	52.99	42.44	48.24	33.07	68.26	50.86	50.00	54.81	84.20	53.34
LaBSE	60.43	49.92	73.45	76.46	58.76	68.45	60.45	74.35	73.65	78.44	49.77	71.23	65.95	60.84	78.65	66.72
Ours	65.82	62.56	76.42	75.20	67.69	80.23	65.09	79.45	80.50	86.42	75.45	79.33	74.21	67.22	88.66	74.95
Testset	134	911	477	231	253	479	117	483	284	257	835	525	112	104	1000	286
Trainset	221	1426	658	22	694	616	351	1030	143	169	472	298	1860	241	1026	409

Highest scores in all languages are in bold. Language codes are based on ISO 639.3.

TABLE II
F1 SCORES ON THE BUCC MINING TASK.

Lang	fr→en	de→en	ru→en	zh→en	en→fr	en→de	en→ru	en→zh
LASER	84.2	87.9	87.1	86.2	83.6	87.5	86.3	88.2
LaBSE	88.7	92.3	88.9	89.5	87.8	92.1	90.4	91.0
Ours	89.4	92.5	88.2	88.9	88.4	92.4	89.9	89.0

the model, significantly enhancing the accuracy of the final model, especially for low-resource languages.

To ensure the efficacy of the entire process, we focus on two key points: first, we use the COMET tool [26] to evaluate the quality of the sentences generated by the machine translation system, filtering out high-quality samples to reduce noise [27]. Second, we use these high-quality parallel sentence pairs in training rather than randomly selected sentence pairs, thereby providing a stronger supervisory signal. During the pretraining phase, we also employ a contrastive loss function to obtain language-independent sentence embeddings, and then fine-tune the cross-lingual MLM model using these high-quality sentence pairs.

III. EXPERIMENTS

A. Evaluation on Bitext Retrieval Tasks

Our goal is to find the best translation of a source language sentence in bilingual text retrieval tasks. We used the latest methods to create a bilingual dictionary [28] and tested our model on the Tatoeba and BUCC datasets.

For the Tatoeba corpus, we focused on 15 low-resource languages, excluded the test set, and fine-tuned our model using the remaining sentence pairs to evaluate the matching accuracy of each sentence [19]. Table I show that our model achieved an average accuracy of 73.95%, surpassing the existing LASER and LaBSE models. This is due to the combination of high-quality sentences and aligned lexical knowledge, as well as iterative semi-supervised fine-tuning.

In the BUCC task, we handled languages such as French, German, Russian, and Chinese and their English translations. The goal was to find translation pairs from a monolingual corpus and evaluate using the F1 score. We used cosine similarity between sentence embeddings to determine translation pairs. Table II indicate that our model outperformed LASER in the French and German tasks and was comparable to LaBSE. However, in the Russian and Chinese tasks, we were slightly outperformed by LaBSE. This could be because our model performs better when handling related languages, whereas supervised models like LaBSE perform better in tasks with larger language differences.

B. Evaluation on Machine Translation Tasks

To evaluate the quality of our mined datasets, we conducted machine translation experiments on two well-known benchmarks. Given the resource-intensive nature of mining parallel sentences, our study focuses primarily on six low-resource language pairs: three English-centered pairs (evaluated on the IWSLT2014 benchmark) and three Chinese-centered pairs (evaluated on the CCMT-2019 benchmark). As a baseline, we used several datasets to train the MT systems, including the CCMatrix [29] mined using the LASER framework, which is currently one of the largest publicly available datasets, as well as the official datasets provided by IWSLT and CCMT.

We utilized the Transformer-base architecture (512/2048, 8 attention heads, 6 layers) for our experiments, using the fairseq toolkit [30] as the default configuration. Byte Pair Encoding (BPE) [31] was employed to generate subwords, performing 30,000 merge operations for each language pair. Translation quality was assessed using the case-sensitive BLEU score [32] with the SacreBLEU toolkit². For the IWSLT benchmark, tst2013 and tst2014 were used as development and test sets, respectively, ensuring that the sentence pairs in the test set were deliberately excluded from the training dataset.

TABLE III
STATISTICS OF MINED CORPORA FOR THE MACHINE TRANSLATION EXPERIMENTS.

Lang		tr	fa	he	ug	bo	MN
Monolingual	CC	740m	258m	192m	2.6m	2.3m	1.2m
	Ours	1.1b	705m	458m	18.4m	9.0m	7.3m
Bilingual	CCMatrix	47m	24.6m	25.2m	-	-	-
	Ours	20.4m	15.2m	14.1m	1.9m	0.8m	0.7m

Language codes are based on ISO 639.1 except for Traditional Mongolian which we use MN. CC denotes number of monolingual sentences mined from CommonCrawl, **CCMatrix** denotes parallel sentences mined by LASER.

1) *Mined Corpora*: Table III presents the data we collected for the machine translation task. We improved our monolingual data acquisition method by identifying more large-scale websites and collecting substantial data in low-resource languages. In our experiments, we gathered 5.43 billion Chinese sentences and 9.94 billion English sentences. We then used cross-lingual sentence embedding techniques to align sentences in low-resource languages with similar Chinese or English sentences, filtering out sentences with a similarity score below 0.8 to ensure quality. It is worth noting that CCMatrix uses a lower similarity threshold, resulting in larger data volumes but potentially lower quality.

²Signature: BLEU+case.mixed+numrefs.1+smooth.exp+tok.13a+version.1.4.1

TABLE IV
BLEU SCORES ON TST2014 TESTSETS OF IWSLT2014 BENCHMARK.

Lang		en→tr	en→fa	en→he	tr→en	fa→en	he→en	avg
Constrained	Official	16.63	15.25	21.87	24.49	25.77	32.08	22.68
	CCMatrix	7.43	7.22	13.51	15.13	12.33	21.40	12.84
	Ours	9.02	9.95	15.55	18.00	14.25	24.88	15.28
	Training Size	156k	109k	187k	156k	109k	187k	150k
Unconstrained	CCMatrix	17.51	13.06	24.57	29.30	30.32	37.51	25.38
	Ours	19.22	16.53	25.78	30.21	31.45	37.94	26.86

TABLE V
BLEU SCORES ON DEV2019 TESTSETS OF CCMT2019 BENCHMARK.

Lang		zh→ug	zh→bo	zh→MN	ug→zh	bo→zh	MN→zh	avg.
Constrained	Official	18.10	22.54	20.50	23.32	17.85	41.22	23.92
	Ours	18.87	22.89	18.58	24.88	17.33	35.33	22.98
Unconstrained	Ours	19.66	24.18	19.15	28.40	19.56	38.50	24.91

The training size for unconstrained condition is shown in Table III.

2) *Main Results*: Table IV shows the experimental results from the IWSLT2014 under two conditions: constrained and unconstrained. In the constrained condition, we used a limited number of sentence pairs from our mined datasets, matching the size of the official training data. In this scenario, our MT models performed better than those trained on CCMatrix, with an average improvement of 2.44 BLEU. This highlights the quality of our mined datasets. However, both our models and those trained on CCMatrix fell short of the official benchmarks. The lower performance in the constrained condition might be due to differences in domain and variable data quality, as well as a closer match between the official training set and the test set.

In the unconstrained condition, where we used all our mined data, there were significant improvements. Our models outperformed those built on CCMatrix by an average of 1.48 BLEU, demonstrating again the quality of our mined data. Moreover, by using a larger dataset, our models exceeded the official benchmarks by 4.18 BLEU.

For Mandarin-centric language pairs in the CCMT2019 benchmark, the results were similar, as shown in Table V. Even though there’s a lack of monolingual sentences for Chinese minority languages on the web, we improved translation performance by nearly 1 BLEU with additional mined parallel pairs.

C. Discussions and Analysis

1) *Effectiveness of Iterative Training*: In our data mining framework, we iteratively generate synthetic sentence pairs for the MT system and train a cross-lingual MLM on these sentence pairs to improve model performance. As shown in Table VI, iterative training significantly enhances performance in the bilingual text retrieval task using the Tatoeba corpus. In the first two iterations, the accuracy increased by a cumulative 3.92%. Performance gains slowed down after that. To balance effectiveness and computational resources, we chose to perform two iterations.

TABLE VI
COMPARISON BETWEEN THE BASELINE MODEL AND BOOSTED MODELS IN A SUBSEQUENT FOUR ITERATIONS IN TATOEBEA CORPUS.

Iter. 0	Iter. 1	Iter. 2	Iter. 3	Iter. 4
70.03	72.74	73.95	74.13	74.20

2) *Impact of Similarity Score Threshold*: The threshold of the similarity score affects the quality and quantity of the extracted

parallel sentences. A lower threshold can extract more but noisier data, while a higher threshold generates fewer yet higher-quality data. By evaluating the impact of different thresholds on machine translation performance, we found that a similarity score threshold of around 0.8 yielded the best results. As shown in Figure 3, this finding indicates that higher-quality parallel sentences can improve translation performance.

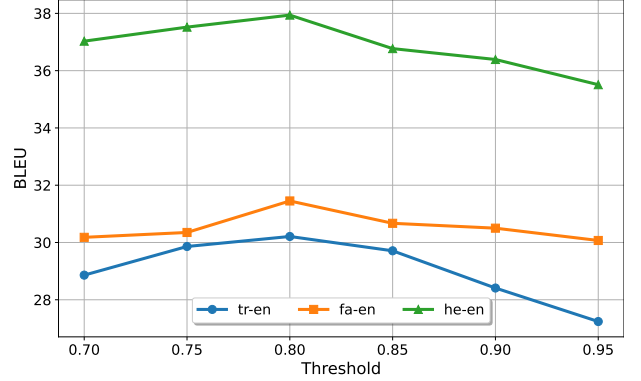


Fig. 3. BLEU scores for different similarity score thresholds on three language directions.

TABLE VII
BLEU SCORES ON FLORES-101 DEVTESTS IN CONSTRAINED CONDITION.

Lang	en→tr	en→fa	en→he	tr→en	fa→en	he→en	avg.
Official	7.22	5.00	12.61	11.60	10.96	18.41	10.97
CCMatrix	9.40	10.86	12.04	15.08	13.93	18.53	13.31
Mined	10.49	11.04	14.16	18.55	14.66	19.71	14.77

3) *Domain Coverage of Mined Corpora*: In our extraction methodology, we employ a strategy to source parallel sentences from a vast array of multi-domain websites utilizing web search techniques. To measure the domain diversity embedded within our harvested corpora, we assess the robustness and domain-wide generalization capabilities of MT systems utilizing the FLORES-101 evaluation datasets [33]. FLORES-101 comprises 3001 sentences curated from 842 articles, categorized into three segments: development (dev), development test (devtest) and test, encapsulating content from three distinct domains further delineated into 10 sub-topical areas. The experimental results, delineated in Table VII, confirm that our extracted corpora exhibit robustness across diverse domains. Notably, systems trained on our corpora demonstrate superior performance, surpassing CCMatrix and the official baselines with 1.46 BLEU and 3.80 BLEU respectively.

IV. CONCLUSION

In this paper, we introduce a robust framework designed for the extraction of high-quality parallel texts from monolingual corpora, with a specific focus on zero-and low-resource languages. By employing bilingual lexicons and a minimal parallel dataset as auxiliary resources, our cross-lingual language model demonstrates a marked improvement in generating language-agnostic sentence embeddings. Model performance is further enhanced through iterative training on both mined and augmented data. Multiple experimental results substantiate the effectiveness of our framework. In the future, we aim to explore the potential of document-level pretraining by leveraging naturally occurring examples of translation in expansive web-based documents.

REFERENCES

- [1] D. Bahdanau, K. H. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” in *3rd International Conference on Learning Representations, ICLR 2015*.
- [2] L. Barrault, O. Bojar, M. R. Costa-jussà, and e. Federmann, Christian, “Findings of the 2019 conference on machine translation (WMT19),” in *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*.
- [3] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, “Learning phrase representations using rnn encoder–decoder for statistical machine translation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- [4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*.
- [5] L. Campbell, “Ethnologue: Languages of the world.”
- [6] R. Dabre, C. Chu, and A. Kunchukuttan, “A survey of multilingual neural machine translation,” *ACM Computing Surveys (CSUR)*.
- [7] R. Wang, X. Tan, R. Luo, T. Qin, and T.-Y. Liu, “A survey on low-resource neural machine translation,” *arXiv preprint arXiv:2107.04239*.
- [8] P. Koehn, “Europarl: A parallel corpus for statistical machine translation,” in *Proceedings of machine translation summit x: papers*.
- [9] M. Ziemski, M. Junczys-Dowmunt, and B. Pouliquen, “The united nations parallel corpus v1. 0,” in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*.
- [10] J. Uszkoreit, J. Ponte, A. Popat, and M. Dubiner, “Large scale parallel document mining for machine translation,” in *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*.
- [11] D. S. Munteanu and D. Marcu, “Improving machine translation performance by exploiting non-parallel corpora,” *Computational Linguistics*.
- [12] A. A. Dara and Y.-C. Lin, “Yoda system for wmt16 shared task: Bilingual document alignment,” in *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*.
- [13] L. Gomes and G. Lopes, “First steps towards coverage-based document alignment,” in *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*.
- [14] M. Artetxe and H. Schwenk, “Margin-based parallel corpus mining with multilingual sentence embeddings,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.
- [15] F. Grégoire and P. Langlais, “A deep neural network approach to parallel sentence extraction,” *arXiv preprint arXiv:1709.09783*.
- [16] H. Schwenk, “Filtering and mining parallel data in a joint multilingual space,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.
- [17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*.
- [18] G. Lample and A. Conneau, “Cross-lingual language model pretraining,” *arXiv preprint arXiv:1901.07291*.
- [19] M. Artetxe and H. Schwenk, “Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond,” *Transactions of the Association for Computational Linguistics*.
- [20] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, “Language-agnostic BERT sentence embedding,” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.
- [21] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, É. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, “Unsupervised cross-lingual representation learning at scale,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.
- [22] I. Kvapilíková, M. Artetxe, G. Labaka, E. Agirre, and O. Bojar, “Unsupervised multilingual sentence embeddings for parallel corpus mining,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*.
- [23] C.-c. Tien and S. Steinert-Threlkeld, “Bilingual alignment transfers to multilingual alignment for unsupervised parallel text mining,” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.
- [24] P. Zweigenbaum, S. Sharoff, and R. Rapp, “Overview of the third bucc shared task: Spotting parallel sentences in comparable corpora,” in *Proceedings of 11th Workshop on Building and Using Comparable Corpora*.
- [25] X. Pan, M. Wang, L. Wu, and L. Li, “Contrastive learning for many-to-many multilingual neural machine translation,” 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)*.
- [26] R. Rei, C. Stewart, A. C. Farinha, and A. Lavie, “COMET: A neural framework for MT evaluation,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- [27] F. Kepler and e. Trénous, Jonay, “Unbabel’s participation in the WMT19 translation quality estimation shared task,” in *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*.
- [28] G. Lample, A. Conneau, L. Denoyer, and M. Ranzato, “Unsupervised machine translation using monolingual corpora only,” in *International Conference on Learning Representations*, 2018.
- [29] H. Schwenk, G. Wenzek, and e. Edunov, Sergey, “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)*.
- [30] M. Ott, S. Edunov, and e. Baevski, Alexei, “fairseq: A fast, extensible toolkit for sequence modeling,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*.
- [31] R. Sennrich, B. Haddow, and A. Birch, “Neural machine translation of rare words with subword units,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.
- [32] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: a method for automatic evaluation of machine translation,” in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.
- [33] N. Goyal, C. Gao, and e. Chaudhary, Vishrav, “The Flores-101 evaluation benchmark for low-resource and multilingual machine translation,” *Transactions of the Association for Computational Linguistics*.