Biomedical Parallel Sentence Retrieval using Large Language Models

Sheema Firdous
Fatima Jinnah Women University, Pakistan
sheemafirdous400@gmail.com

Sadaf Abdul Rauf
Fatima Jinnah Women University, Pakistan
Univ. Paris-Saclay, CNRS, LIMSI France
sadaf.abdulrauf@gmail.com

Abstract

We have explored the effect of in domain knowledge during parallel sentence filtering from in domain corpora. Models built with sentences mined from in domain corpora without domain knowledge performed poorly, whereas model performance improved by more than 2.3 BLEU points on average with further domain centric filtering. We have used Large Language Models for selecting similar and domain aligned sentences. Our experiments show the importance of inclusion of domain knowledge in sentence selection methodologies even if the initial comparable corpora are in domain.

1 Introduction

This paper describes FJWU's submission to the biomedical translation task. This year the focus of our research was domain specific parallel corpus mining from Wikipedia using Large Language Models, we explored the potential of the mined sentences using two sentence selection schemes. Neural Machine Translation (NMT) (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Bahdanau et al., 2015; Cho et al., 2014) has witnessed great success over the years (Vaswani et al., 2017; Zhang and Zong, 2020). NMT systems train on parallel corpora to produce translations that capture language intricacies and context with enormous precision as compared to the previous counterpart Statistical Machine Translation (SMT) systems.

Machine translation in the biomedical domain is becoming increasingly important due to the critical nature of medical scientific texts. The majority of these texts are published in English, and the goal of Biomedical Machine Translation is to make them accessible in multiple languages. However, this is a complex undertaking due to the extensive nature of this field and the vast and diverse vocabulary it encompasses. This vocabulary includes specialized terms and non-lexical forms (such as dates and biomedical entities) that pose unique challenges. Consequently, the quality of machine translation output fluctuates depending on the availability of biomedical resources tailored to each target language.

Availability of parallel corpora in reasonable amounts has greatly enhanced the performance of NMT systems, especially for the high-resource languages (Bojar et al., 2018). However, its efficacy remains sub optimal for low-resource languages and domain-specific contexts (Zoph et al., 2016; Koehn and Knowles, 2017; Lample et al., 2018; Chu and Wang, 2020). Performance of NMT system degrades as soon as the application domain deviates from training domain. Domain adaptation (Freitag and Al-Onaizan, 2016), transfer learning (Zoph et al., 2016; Khan et al., 2018; Abdul Rauf et al., 2020), model fusion (Gulcehre et al., 2015), back translation (Sennrich et al., 2015; Ul Haq et al., 2020), fine-tuning (Dakwale and Monz, 2017; Huck et al., 2018), data augmentation (Fadaee et al., 2017), data selective training (Van Der Wees et al., 2017; Knowles and Koehn, 2018), decoding strategies (Park et al., 2020), zero-shot translation (Johnson et al., 2017) are some of the techniques used to address this issue. We will be focusing on domain adaptation using data augmentation and fine tuning.

For this years submission we explore the potential of Large-scale Language Models for extracting parallel sentences from Wikipedia[1]. French-English parallel articles are scraped as detailed in Section 4. For learning sentence embeddings of scraped bilingual data, rather than training encoders from scratch, we leverage the potential of LLM in parallel sentence extraction from our bilingual scraped articles. We used LEALLA-Large, a lightweight system developed by (Mao and Nakagawa, 2023) to compute the language-agnostic low-dimensional sentence embeddings for each

[1]An online multilingual encyclopedia https://en.wikipedia.org/wiki/Main_Page

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 263–270
December 6–7, 2023.

sentence in the English and French parallel articles. Potential parallel sentences are filtered based on the similarity scores. These sentence are then further domain filtered by comparing the closeness with Medline Titles embeddings computed using Transformers MiniLM. Our experiments show the importance of inclusion of domain knowledge in sentence selection methodologies even if the initial comparable corpora are in domain. Our main contributions include:

- Presenting a methodology for domain inclusion in sentence retrieval tasks by using capabilities of Large Language Models
- Highlighting the importance of inculcation of in domain knowledge in sentence retrieval tasks even when the data source is in domain
- Release of the mined parallel corpora to the research community[2]

The paper is structured as follows: Section 2 presents a brief overview of background and related work, Section 3,4 elaborates the data collection pipeline, Section 5 outline the NMT experiments and results, followed by the conclusion of this study.

2 Related Work

Recent work on parallel sentence extraction has focused on lightweight end-to-end word-level and sentence-level embedding methods (Guo et al., 2018; Artetxe and Schwenk, 2018; Yang et al., 2019a). These embedding-based approaches have gained success (Grégoire and Langlais, 2017; Bouamor and Sajjad, 2018; Schwenk, 2018) as these systems outperformed the large-distributed computationally intensive systems (Uszkoreit et al., 2010; Abdul-Rauf and Schwenk, 2009) used to mine parallel documents. Bilingual sentence embeddings, learned from dual-encoder models, have also been used effectively for parallel corpus mining (Guo et al., 2018). Cross-lingual embeddings encode bilingual texts into a single unified vector space allowing nearest-neighbor search can be used to find potential translation candidates. These embedding approaches produce noisy matches that require a re-scoring step in order to obtain a clean parallel sentence retrieval as addressed by (Yang et al., 2019a) who explored using a bi-directional dual encoder with additive margin softmax (Wang et al., 2018) which results in state-of-the-art performance for sentence filtering. Multilingual sentence embedding approaches (Artetxe and Schwenk, 2018; Chidambaram et al., 2018) also show promising results.

Since language-specific models often demand extensive amounts of labeled data for training and can be limited by their language-specific parameters, language-agnostic sentence embedding(Artetxe and Schwenk, 2019; Yang et al., 2019b; Reimers and Gurevych, 2020; Feng et al., 2020; Mao et al., 2022) align multiple languages in a shared embedding space, facilitating parallel sentence alignment that extracts parallel sentences for training translation systems. Among them, LaBSE (Feng et al., 2020) achieved state-of-the-art performance on various bi-text retrieval. The problem of inference speed and computation overhead of large language models was addressed by (Mao and Nakagawa, 2023) who proposed Learning Leight-Weight Language-agnostic Sentence Embeddings (LEALLA) with Knowledge Distillation (Kim and Rush, 2016). They reported significant reduction in computation overhead and inference speed by providing language-agnostic low-dimensional sentence embeddings. We also use LEALLA in the second phase of our pipeline for parallel sentence alignemnent.

3 Wikipedia as a potential resource for biomedical data

Our primary objective was to collect a comprehensive dataset from the biomedical domain, we explored Wikipedia's key biological categories and selected those having a substantial volume of articles. A brief overview of the selected subdomains is given below:

1. **Biodbs** [3] refers to biological databases and contains links of a variety of biological databases.
2. **Genome Reference Consortium** is an international collaboration dedicated to creating and maintaining the most accurate and up-to-date Human Genome [4] reference sequence.

[2] https://github.com/sabdul111/Biomedical-Parallel-Corpus

[3] https://en.wikipedia.org/wiki/List_of_biological_databases

[4] https://en.wikipedia.org/wiki/Human_genome

Domain	Scraped URLs		Scraped Articles		Parallel Articles	Unique Articles
	French	English	French	English		
Biodbs	39.4K	77.3K	39.3K	68.7K	39.3K	1.2K
Human Genome	25.9K	59.1K	25.9K	49K	25.9K	25.9K
Health BioMed	42.8K	122.5K	42.8K	92.5K	42.8K	14.7K
NCBI	64.2K	133.8K	64K	133.6K	64K	51.2K
Pubmed	62.9K	134.5K	62.9K	117.4K	62.9K	22.4K
Total	235.2K	527.2K	234.9K	461.2K	234.9K	115.4K

Table 1: Scraped Data per subdomain

3. **National Institute of Biomedical Imaging and Bio engineering** plays a central role in advancing biomedical engineering research and provides a wealth of data and resources in the domain of Health Biomedical Engineering [5].

4. **The National Center for Biotechnology Information (NCBI)** [6] is a U.S. government agency that provides an extensive collection of biomedical and genomic resources.

5. **PubMed** [7] is a widely used online database maintained by the National Library of Medicine (NLM) which provides access to a vast collection of biomedical literature.

4 Parallel Corpus Mining

This section presents an overview of our parallel data creation pipeline. Wikipedia has been extensively used as a data resource for corpus development (Chu et al., 2014; Tufiş et al., 2013; Stefanescu et al., 2012; Karimi et al., 2018; Aghaebrahimian, 2018; Schwenk et al., 2019). We also used Wikipedia's inter language links to mine potential parallel sentences by exploring the potential of Large language models for filtering the closet candidates. Our data preparation pipeline involves three main steps; 1) Domain specific web scraping, 2) Candidate sentence scoring and filtering and 3) Domain adapted filtering.

Parallel article scrapping To extract the bilingual data we used Wikipedia's **Interwiki**[8] (also known as inter language links) property (Adafre and De Rijke, 2006; Otero and López, 2010; Chu et al., 2014; Aghaebrahimian, 2018). English Wikipedia has consistently held the distinction of possessing the highest article count among all language editions of Wikipedia. As of August 2023, there are 6,696,071[9] articles in English containing over 4.3 billion words.

We maximized recall in our article selection procedure by choosing English as the base language since it provided wider coverage of topics. Thus, for each unique English article, the corresponding French article (if found) was scrapped. We named the scrapped articles using the title of the English version, distinguishing them with *.en* for English and *.fr* for French files. At this stage, we had to retrieve the parallel articles since many of the English articles did not have the corresponding French articles (see Table 1). For parallel article retrieval, we compiled a list of all French articles and used this list to retrieve parallel English articles which resulted in our parallel French-English articles. The subdomains (see section § 3) had many overlapping articles which were removed and unique articles from each subdomain were selected.
Table 1 shows the amount of URLs, articles, parallel articles and the corresponding unique articles. At this stage we have unique parallel articles from each subdomain.

Parallel sentence filtering We used a lightweight pre-trained large language model LEALLA-Large (Mao and Nakagawa, 2023) which computes sentence embedding of 256 dimensions by distilling knowledge from LaBSE (Feng et al., 2020). It can be used to mine potential parallel sentences by finding the nearest neighbour of each source sentence in the target side according to cosine similarity, and filtering those below a threshold.

[5] https://en.wikipedia.org/wiki/Biomedical_engineering#Hospital_and_medical_devices

[6] https://en.wikipedia.org/wiki/National_Center_for_Biotechnology_Information

[7] https://en.wikipedia.org/wiki/PubMed

[8] The Interwiki property links the articles across various language editions of Wikipedia.

[9] https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia

Domain	Parallel Sentences		
	Threshold 90	Threshold 85	Threshold 80
Biodbs	1,188	3,240	4,944
Human Genome	25,975	19,849	62,499
Health BioMed	14,677	41,555	66,008
NCBI	65,591	198,692	328,621
Pubmed	16,853	46,273	72,741
Total	124,284	309,609	534,813

Table 2: Parallel Sentences from the unique articles based on similarity threshold computed using LEALLA.

Parallel Sentences	BioFiltered Parallel Sentences		
	Threshold 20	Threshold 10	Threshold 0
Threshold 90	3,602	16,861	47,964
Threshold 85	15,286	64,888	169,215
Threshold 80	23,727	101,845	275,063
Total	42,615	183,594	492,242

Table 3: Bio-Filtered: Parallel sentences from Table 2 selected based on their proximity with Medline titles using MiniLM.

LEALLA Embedding vector is computed for each sentence in the French and English article. Thus for each French(source) sentence we have N potential matching sentences, where N is the number of sentences in English(target) article. The dot-product is then used to compute the similarity between each source and N target candidate sentences. The top 10 candidate sentences are retrieved for each sentence. At this stage we have a sorted list of potential parallel sentences from each subdomain.

It is important to note that these are potential bio med domain sentences since these are mined from in-domain articles. We focus on both precision and recall at this stage. Our sentence retrieval is recall oriented, given that English articles were roughly double the French articles, thus using French sentence as prompt to retrieve the matching English sentences promised a wider search space. For final parallel corpus creations we selected the sentences on similarity threshold. We report three thresholds (thresholds 80, 85, and 90) to retrieve parallel sentences from the retrieved top-10 sentence pairs. We are working on lower threshold sentences. A higher threshold indicates a greater degree of parallelism between the sentences. Table 2 shows the number of parallel sentences retrieved using different thresholds for each subdomain. We call these *LLMfilter* sentences for reference.

In domain filtering We did a second level selection from the *LLMfilter* parallel sentences extracted in the previous step. Even though these sentences come from bio-medical articles and are in-domain but our hypothesis is that there will be many sentences that may categorize as general domain. Our second filter is to ensure collection of purely biomedical sentences. For this we select Medline titles (Jimeno Yepes et al., 2017) as biomedical representative dataset since titles contain the main domain terminologies. An embedding was generated for Medline Titles using sentence transformers `paraphrase-multilingual-MiniLM-L12-v2`[10] which was then used to remove the out-domain sentences, striving to retain an optimal amount of in-domain sentences (pertaining to the biomedical domain). Dot product of each sentence with the Medline titles embedding was used to compute the similarity score(ranging from -1 to 1). We selected thresholds 20, 10, and 0 which correspond to 0.2, 0.1, and 0.0 respectively in the similarity score. Table 3 shows the number of sentences per threshold, we call these *Biofilter* sentences for reference.

Post-processing involved the removal of exceptionally short sentences, special characters, and sentences in languages other than the intended source and target languages. Duplicated and identical sentences were also removed from both English and French sides.

5 Translation performance on retrieved sentences

We used Transformer base (Vaswani et al., 2017) architecture provided by Fairseq (Ott et al., 2019)

[10] https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

Model Name	Fine-tuning	WMT20 testset *LLMfilter*	Model Name	*Biofilter*
B1	-	19.52		
S1	B1 =>t90	18.12	SB1	20.29
S2	B1 =>t85	18.41	SB2	20.29
S3	B1 =>t80	18.54	SB3	20.58
S4	B1 =>t90-t85-t80	**18.78**	SB4	**21.11**
B2	-	38.71		
L5	B2 =>t90	19.69	LB1	21.81
L6	B2 =>t85	20.57	LB2	21.88
L7	B2 =>t80	**20.62**	LB3	22.07
L8	B2 =>t90-t85-t80	20.36	LB4	**22.43**

Table 4: BLEU scores on fine tuned datasets. B1 and B2 denote the baselines. B1 is trained on the biomedical texts provided by the WMT'23 organizers, while B2 is a big model trained on general domain and biomed data.

as `transformer_iwslt_en_de`. The ReLU activation function was used in all encoder and decoder layers. We optimize with Adam (Kingma and Ba, 2015), set up with a maximum learning rate of 0.0005 and an inverse square root decay schedule, as well as 4000 warmup updates.

All corpora were segmented into subword units using Sentence Piece (Kudo and Richardson, 2018) with a vocabulary of 32K units. We share the decoder input and output embedding matrices. Models are trained with mixed precision and a batch size of 4096 tokens on a single GPU. Systems were trained until convergence based on the BLEU score on the development sets. Evaluation was performed using SacreBleu (Post, 2018). Scores are chosen based on the best score on the development set (Medline 18, 19), and the corresponding scores for that checkpoint are reported on Medline 20 test set.

For fine-tuned systems, the process starts with models trained to convergence, based on BLEU score on dev sets. Training then resumes using a selected portion of the training corpus using the same parameters and criterion as for the base systems.

Baseline We trained a smaller model B1 on the biomedical texts provided by the WMT'23 organizers: Edp, Medline abstracts and titles (Jimeno Yepes et al., 2017), Scielo (Neves et al., 2016) and the Ufal Medical corpus[11] consisting of Cesta, Ecdc, Emea (OpenSubtitles), PatTR Medical and Subtitles. We used a bigger model B2 by (Xu et al., 2021) trained on WMT14 general domain corpus and WMT and supplementary biomed data including B1 data.

[11] https://ufal.mff.cuni.cz/ufal_medical_corpus

5.1 Results and Discussion

Table 4 presents the results using the two data selection methods. *LLMfilter* column shows the BLEU scores on Medline 20 testset for sentences filtered based on the sentence similarity score, whereas *Biofilter* are the sentences which were selected from the *LLMfilter* based on their closeness with the Biomedical Medline titles. Both filters used LLMs for computing similarity as detailed in section 4.

B1 represents a smaller baseline model trained on all biomed data provided by WMT organizers having a BLEU score of 19.52. This was further fine-tuned using each threshold dataset i.e. threshold 90, 85, and 80 (represented by t90, t85, and t80 respectively in 4), and finally with a concatenation of the 3 thresholds. Concatenation refers to the union of t90, t85, and t80. We did this to upsample the higher quality corpora (i.e. t90) to analyze the impact on MT. Evidently, none of the *LLMfilter* sentences improved the initial bio med baseline. The *Biofilter* sentences on the other hand helped improve the scores even when a small amount is added e.g. for t90 and the scores improved consistently with the increase in the number of sentences with SB4 yielding an increase of 1.59 BLEU points from the baseline. For the larger baseline B2, though none of the filtering schemes help improve the initial high score but still the supremacy of *Biofilter* sentences over *LLMfilter* is evident.

Arguably, both *LLMfilter* and *Biofilter* contain in-domain sentences as these have been selected from biomedical articles. The models built using the same thresholds for the two schemes have a difference of more than 2 BLEU points on average

with *Biofilter* systems being superior. Our results demonstrate the importance of inculcation of in-domain knowledge in sentence retrieval tasks even if the data source is in-domain as there are many sentences that do not pertain specifically to the domain and affect the results of domain-centered translation.

6 Conclusion

In this study, we explored the potential of large language models for parallel sentence extraction from domain-adapted bilingual corpus extracted from Wikipedia. On our dataset, we experimented with two data selection schemes and assessed the NMT performance for the biomedical domain. Our findings demonstrate that merely web-mining from in-domain corpus is not sufficient to improve domain-specific NMT performance but there is also a need for further filtering out out-domain sentences to improve the domain-specific NMT systems. Leveraging large language models to extract in-domain parallel sentences resulted in improved NMT performance by outperforming the baseline with 2 BLEU points.

Acknowledgments

This study is funded by the National Research Program for Universities (NRPU) by Higher Education Commission of Pakistan (5469/Punjab/NRPU/R&D/HEC/2016).

References

Abdul Rauf, S., Rosales Núñez, J. C., Pham, M. Q., and Yvon, F. (2020). Limsi @ wmt 2020. In *Proceedings of the Fifth Conference on Machine Translation*, pages 803–812, Online. Association for Computational Linguistics.

Abdul-Rauf, S. and Schwenk, H. (2009). On the use of comparable corpora to improve SMT performance. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 16–23, Athens, Greece. Association for Computational Linguistics.

Adafre, S. F. and De Rijke, M. (2006). Finding similar sentences across multiple languages in wikipedia. In *Proceedings of the Workshop on NEW TEXT Wikis and blogs and other dynamic text sources*.

Aghaebrahimian, A. (2018). Deep neural networks at the service of multilingual parallel sentence extraction. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1372–1383, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Artetxe, M. and Schwenk, H. (2018). Margin-based parallel corpus mining with multilingual sentence embeddings. *arXiv preprint arXiv:1811.01136*.

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

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bojar, O., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., and Monz, C. (2018). Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Bouamor, H. and Sajjad, H. (2018). Parallel sentence extraction from comparable corpora using multilingual sentence embeddings. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC), Miyazaki, Japan*, pages 7–12.

Chidambaram, M., Yang, Y., Cer, D., Yuan, S., Sung, Y.-H., Strope, B., and Kurzweil, R. (2018). Learning cross-lingual sentence representations via a multi-task dual-encoder model. *arXiv preprint arXiv:1810.12836*.

Cho, K., Van Merriënboer, B., Bahdanau, D., and Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Chu, C., Nakazawa, T., and Kurohashi, S. (2014). Constructing a chinese—japanese parallel corpus from wikipedia. In *LREC*, pages 642–647.

Chu, C. and Wang, R. (2020). A survey of domain adaptation for machine translation. *Journal of information processing*, 28:413–426.

Dakwale, P. and Monz, C. (2017). Fine-tuning for neural machine translation with limited degradation across in-and out-of-domain data. In *Proceedings of Machine Translation Summit XVI: Research Track*, pages 156–169.

Fadaee, M., Bisazza, A., and Monz, C. (2017). Data augmentation for low-resource neural machine translation. *arXiv preprint arXiv:1705.00440*.

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., and Wang, W. (2020). Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Freitag, M. and Al-Onaizan, Y. (2016). Fast domain adaptation for neural machine translation. *arXiv preprint arXiv:1612.06897*.

Grégoire, F. and Langlais, P. (2017). A deep neural network approach to parallel sentence extraction. *arXiv preprint arXiv:1709.09783*.

Gulcehre, C., Firat, O., Xu, K., Cho, K., Barrault, L., Lin, H.-C., Bougares, F., Schwenk, H., and Bengio, Y. (2015). On using monolingual corpora in neural machine translation. *arXiv preprint arXiv:1503.03535*.

Guo, M., Shen, Q., Yang, Y., Ge, H., Cer, D., Abrego, G. H., Stevens, K., Constant, N., Sung, Y.-H., Strope, B., et al. (2018). Effective parallel corpus mining using bilingual sentence embeddings. *arXiv preprint arXiv:1807.11906*.

Huck, M., Stojanovski, D., Hangya, V., and Fraser, A. (2018). Lmu munich's neural machine translation systems at wmt 2018. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 648–654.

Jimeno Yepes, A., Névéol, A., Neves, M., Verspoor, K., Bojar, O., Boyer, A., Grozea, C., Haddow, B., Kittner, M., Lichtblau, Y., Pecina, P., Roller, R., Rosa, R., Siu, A., Thomas, P., and Trescher, S. (2017). Findings of the WMT 2017 biomedical translation shared task. In *Proceedings of the Second Conference on Machine Translation*, pages 234–247, Copenhagen, Denmark. Association for Computational Linguistics.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., et al. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Kalchbrenner, N. and Blunsom, P. (2013). Recurrent continuous translation models. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1700–1709.

Karimi, A., Ansari, E., and Sadeghi Bigham, B. (2018). Extracting an English-Persian parallel corpus from comparable corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Khan, A., Panda, S., Xu, J., and Flokas, L. (2018). Hunter nmt system for wmt18 biomedical translation task: Transfer learning in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 655–661.

Kim, Y. and Rush, A. M. (2016). Sequence-level knowledge distillation. *arXiv preprint arXiv:1606.07947*.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Knowles, R. and Koehn, P. (2018). Context and copying in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3034–3041, Brussels, Belgium. Association for Computational Linguistics.

Koehn, P. and Knowles, R. (2017). Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*.

Kudo, T. and Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Lample, G., Ott, M., Conneau, A., Denoyer, L., and Ranzato, M. (2018). Phrase-based & neural unsupervised machine translation. *arXiv preprint arXiv:1804.07755*.

Mao, Z., Chu, C., and Kurohashi, S. (2022). Ems: efficient and effective massively multilingual sentence representation learning. *arXiv preprint arXiv:2205.15744*.

Mao, Z. and Nakagawa, T. (2023). Lealla: Learning lightweight language-agnostic sentence embeddings with knowledge distillation. *arXiv preprint arXiv:2302.08387*.

Neves, M., Yepes, A. J., and Névéol, A. (2016). The Scielo Corpus: a parallel corpus of scientific publications for biomedicine. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2942–2948, Portorož, Slovenia. European Language Resources Association (ELRA).

Otero, P. G. and López, I. G. (2010). Wikipedia as multilingual source of comparable corpora. In *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora, LREC*, pages 21–25.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019). fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038*.

Park, C., Yang, Y., Park, K., and Lim, H. (2020). Decoding strategies for improving low-resource machine translation. *Electronics*, 9(10):1562.

Post, M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Reimers, N. and Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. *arXiv preprint arXiv:2004.09813*.

Schwenk, H. (2018). Filtering and mining parallel data in a joint multilingual space. *arXiv preprint arXiv:1805.09822*.

Schwenk, H., Chaudhary, V., Sun, S., Gong, H., and Guzmán, F. (2019). Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *arXiv preprint arXiv:1907.05791*.

Sennrich, R., Haddow, B., and Birch, A. (2015). Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Stefanescu, D., Ion, R., and Hunsicker, S. (2012). Hybrid parallel sentence mining from comparable corpora. In *Proceedings of the 16th annual conference of the European association for machine translation*, pages 137–144.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Tufiş, D., Ion, R., Dumitrescu, Ş. D., and Stefanescu, D. (2013). Wikipedia as an smt training corpus. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 702–709.

Ul Haq, S., Abdul Rauf, S., Shaukat, A., and Saeed, A. (2020). Document level NMT of low-resource languages with backtranslation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 442–446, Online. Association for Computational Linguistics.

Uszkoreit, J., Ponte, J., Popat, A., and Dubiner, M. (2010). Large scale parallel document mining for machine translation.

Van Der Wees, M., Bisazza, A., and Monz, C. (2017). Dynamic data selection for neural machine translation. *arXiv preprint arXiv:1708.00712*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, F., Cheng, J., Liu, W., and Liu, H. (2018). Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25(7):926–930.

Xu, J., Pham, M. Q., Abdul Rauf, S., and Yvon, F. (2021). LISN @ WMT 2021. In *Proceedings of the Sixth Conference on Machine Translation*, pages 232–242, Online. Association for Computational Linguistics.

Yang, Y., Abrego, G. H., Yuan, S., Guo, M., Shen, Q., Cer, D., Sung, Y.-H., Strope, B., and Kurzweil, R. (2019a). Improving multilingual sentence embedding using bi-directional dual encoder with additive margin softmax. *arXiv preprint arXiv:1902.08564*.

Yang, Y., Cer, D., Ahmad, A., Guo, M., Law, J., Constant, N., Abrego, G. H., Yuan, S., Tar, C., Sung, Y.-H., et al. (2019b). Multilingual universal sentence encoder for semantic retrieval. *arXiv preprint arXiv:1907.04307*.

Zhang, J. and Zong, C. (2020). Neural machine translation: Challenges, progress and future. *Science China Technological Sciences*, 63(10):2028–2050.

Zoph, B., Yuret, D., May, J., and Knight, K. (2016). Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*.