Leveraging Closed-Access Multilingual Embedding for Automatic Sentence Alignment in Low Resource Languages

Idris Abdulmumin
Ahmadu Bello University
Zaria, Nigeria
iabdulmumin@abu.edu.ng

Auwal Abubakar Khalid
Bayero University
Kano, Nigeria
aka2000078.mcs@buk.edu.ng

Shamsuddeen Hassan Muhammad
University of Porto
Porto, Portugal
shamsuddeen2004@gmail.com

Ibrahim Said Ahmad
Northeastern University
Boston, USA
isab7070@gmail.com

Lukman Jibril Aliyu
Arewa Data Science
Nigeria
lukman.j.aliyu@gmail.com

Babangida Sani
Bayero University
Kano, Nigeria
bangis.cs@gmail.com

Bala Mairiga Abduljalil
University Of Maiduguri
Borno, Nigeria
ballaabduljalil@gmail.com

Sani Ahmad Hassan
Ahmadu Bello University
Zaria, Nigeria
myynazee@gmail.com

***Abstract*—The importance of qualitative parallel data in machine translation has long been determined but it has always been very difficult to obtain such in sufficient quantity for the majority of world languages, mainly because of the associated cost and also the lack of accessibility to these languages. Despite the potential for obtaining parallel datasets from online articles using automatic approaches, forensic investigations have found a lot of quality-related issues such as misalignment, and wrong language codes. In this work, we present a simple but qualitative parallel sentence aligner that carefully leveraged the closed access Cohere multilingual embedding,[1] a solution that ranked second in the just concluded #CoHereAIHack 2023 Challenge. [2] The proposed approach achieved 94.96 and 54.83 f1 scores on FLORES and MAFAND-MT, compared to 3.64 and 0.64 of LASER respectively. Our method also achieved an improvement of more than 5 BLEU scores over LASER, when the resulting datasets were used with MAFAND-MT dataset to train translation models. Our code and data are available for research purposes here.[3]**

Keywords—Sentence Alignment, Multilingual Language Models, Machine Translation, Natural Language Processing

I. INTRODUCTION

As globalization continues to blur the lines between cultures and languages, the need for effective language translation tools has surged. Despite the increasing improvement in the quality of automatic translation tools, many languages still tend to be underrepresented in machine translation systems [1]. This is primarily due to the lack of extensive parallel corpora, which are essential for training robust translation models [2]. This has led to a growing interest in development of methods for generating parallel data in low-resource languages through methods such as translation, e.g., in back-translation [3] and self-learning [4]; and extracting potential parallel sentences from large corpora, often sourced from the internet [5]–[7].

A lot of potential parallel sentences exist on the internet, especially on multilingual news and educational sites. Examples of such include the BBC that create contents in many languages such as English, Hausa, Yoruba, etc., and also local news media such as Premium Times and Daily Trust that both have English and Hausa versions of their contents. This, therefore, provides the potential for bridging the lack of parallel data if the parallel sentences can be extracted from these sources, especially through automatic processes. Automatic sentence alignment is the process of identifying which sentences in a source text correspond to which sentences in a target text, enabling the extraction of potential parallel sentences from a large corpora. This is especially possible when the sentences in both languages can be represented in a common vector space—multilingual embeddings—such that the sentences that share semantic similarity are close to each other in the vector space.

Multilingual embeddings are a representation of words or sentences that capture their semantic relationships across multiple languages. These embeddings, often generated using deep learning models like word2vec, fastText, or BERT, encode the meaning and context of linguistic units in a continuous vector space. Traditional alignment methods often relied on linguistic rules and heuristics, which were language-specific and challenging to adapt across different pairs of languages. In contrast, multilingual embeddings offer a universal framework for sentence alignment that transcends language boundaries. By leveraging these embeddings, automatic sentence alignment systems gain a powerful advantage. This method provided the basis for large multilingual parallel data such as the Paracrawl corpus [5] and the CCAligned corpus [6].

Recently, the LASER toolkit [8] was developed to facilitate the use of multilingual embeddings for sentence alignment. However, empirical investigation of the LASER aligned sentences found that a lot of the sentences are actually not aligned [9], [10]. Some of the sentences were even found to be not in the language they were claimed to be in. This problem is not unconnected to the quality of the multilingual embeddings used. In this work, therefore, we propose using more qualitative

[1] https://docs.cohere.com/docs/multilingual-language-models

[2] https://ai6lagos.devpost.com

[3] https://github.com/abumafrim/Cohere-Align

private multilingual embeddings provided (restrictively) by CoHere. We showed that even when implemented using the simple nearest neighbor method, our method outperforms the LASER toolkit in terms of the quality of the aligned sentences. We evaluated our method on the Hausa-English language pair, and also showed that the parallel data generated by our method trained a better machine translation model than the LASER aligned data.

II. Related Works

[6] applied URL-matching to curate a cross-lingual document dataset from the CommonCrawl corpus. The dataset contains over 392 million document pairs from 8144 language pairs, covering 138 distinct languages. [11] presented an approach based on multilingual sentence embeddings to automatically extract parallel sentences from the content of Wikipedia articles in 96 languages, including several dialects or low-resource languages. [7] showed that margin-based bitext mining in a multilingual sentence space can be successfully scaled to operate on monolingual corpora of billions of sentences. They used 32 Common Crawl snapshots (Wenzek et al., 2019), totaling 71 billion unique sentences. Using one unified approach for 90 languages, they were able to mine 10.8 billion parallel sentences, out of which only 2.9 billion are aligned with English.

[12] moved away from the popular one-for-all multilingual models and focused on training multiple language (family) specific representations, but most prominently enabled all languages to still be encoded in the same representational space. They focused on teacher-student training, allowing all encoders to be mutually compatible for bitext mining and enabling fast learning of new languages. They also combined supervised and self-supervised training, allowing encoders to take advantage of monolingual training data. The approach significantly outperforms the original LASER encoder. They studied very low-resource languages and handled 44 African languages, many of which are not covered by any other model. For these languages, they trained sentence encoders and mined bitexts.

[13] constructed an evaluation set for Cross-Lingual Language Understanding (XLU) by extending the development and test sets of the Multi-Genre Natural Language Inference Corpus (MultiNLI) to 15 languages, including low resource languages such as Swahili and Urdu. In addition, they provided several baselines for multilingual sentence understanding, including two based on machine translation systems and two that use parallel data to train aligned multilingual bag-of-words and LSTM encoders. [14] created a corpus for multilingual document classification. They proposed a new subset of the Reuters corpus with balanced class priors for eight languages. By adding Italian, Russian, Japanese, and Chinese, languages which are very different with respect to syntax and morphology, are covered. They provided baselines for all language transfer directions using multilingual word and sentence embeddings, respectively.

[15] presented an approach to encode a speech signal into a fixed-size representation that minimizes the cosine loss with the existing massively multilingual LASER text embedding space. Sentences are close in this embedding space, independently of their language and modality, either text or audio. Using a similarity metric in that multimodal embedding space, they performed mining of audio in German, French, Spanish, and English from Librivox against billions of sentences from Common Crawl. This yielded more than twenty thousand hours of aligned speech translations. To evaluate the automatically mined speech/text corpora, they trained neural speech translation systems for several language pairs. Adding the mined data achieves significant improvements in the BLEU score on the CoVoST2 and the MUST-C test sets with respect to a very competitive baseline.

III. Methodology

A. CoHere Multilingual Embedding

The CoHere[4] multilingual embedding is a 768- dimensional model that was developed to enable multilingual semantic search, aggregate customer feedback, and cross-lingual zero-shot content moderation across 100 languages, including Hausa.[5] Access to this model is enabled via an API, after authentication using an API key. The key can be generated at https://dashboard.cohere.com/api-keys.

B. CoHere Sentence Aligner

We adapted the evaluation script[6] of Vecmap [16] to create the source-target sentence aligner. The aligner was implemented using the nearest neighbor algorithm, after using the CoHere multilingual embedding model to convert the source and target sentences into a 768-dimensional vector.

The free CoHere embedding API only allows the conversion of about 6,000 sentences to embeddings per minute. Consequently, we designed the CoHere sentence aligner to sleep for 61 seconds after downloading a batch of source and target sentences. We set the batch size at 2,000 (or the remaining number of sentences) each of the source and target sentences (4,000 altogether) at each iteration until every sentence embedding is obtained. To persist the generated embeddings, in case of any future use, we save them to file, and upload it whenever a previously converted sentence embedding is needed.

C. Datasets and Pre-processing

We crawled 1,000 Hausa and English news articles each. For pre-processing of this data, we used the NLTK [17] sentence tokenizer[7] to split each of the crawled documents into a list of sentences. These sentences were then merged to produce both the target and source files. We removed empty lines, and also sentences that contain fewer than 5 and longer than 80 words, after tokenization with the NLTK's word tokenizer. Table I shows the statistics of the crawled data before and after cleaning. For evaluation, we used the MAFAND-MT [1] train, test, and dev; and FLORES [18] dev and devtest datasets.

TABLE I.
STATISTICS OF MONOLINGUAL HAUSA AND ENGLISH SENTENCES

lang.	***# crawled sents.***	***# cleaned sents.***
hau	13,916	13,560
eng	23,148	22,671

[4] https://docs.cohere.com/docs/multilingual-language-models
[5] https://docs.cohere.com/docs/supported-languages
[6] https://github.com/artetxem/vecmap/blob/master/eval_translation.py
[7] https://www.nltk.org/api/nltk.tokenize.html

D. Evaluation

To evaluate the performance of the developed, CoHere sentence aligner, we used a pre-trained LASER[8] model to create another sentence aligner. We then used datasets where the expected target sentences are known, enabling us to determine the actual quality of the paired sentences, using the f1 metric score.[9] We used the English-Hausa pair of FLORES-200 dev and devtest; and MAFAND-MT train, test and dev sets for this evaluation process.

Using the crawled articles, we used the two aligners to pair the potential source and target sentences. Without a reference text to evaluate this performance, we relied on human evaluation to classify a sample of the generated translations using the following labels: (1) **Not a translation at all**, (2) **Bad**, (3) **Can be considered a translation**, (4) **Good**, and (5) **Perfect**.

Finally, we used the aligned sentences to train various machine translation models, in a semi-supervised set-up—using the labeled MAFAND-MT training data. These models were developed on the MAFAND-MT development set, by fine-tuning a public checkpoint of the M2M-100 [19] seq-to-seq model. This is a transformer-based [20] model that was trained to support direct translation between 100 languages without first relying on the English language. After training, the models were evaluated using both the FLORES devtest and MAFAND-MT test datasets, using the SacreBLEU [21], [22] metric.

IV. Results and Discussion

A. Parallel Data

Table II shows the performances of the two sentence aligners. It can be seen that the performance of the CoHere aligner outrightly outperformed that of the LASER's. It may be argued, though, that since the FLORES data is widely available, the CoHere multilingual embedding may have been trained on the data, since the model's training data is not known. But the same argument cannot be made for the recently released MAFAND-MT datasets, where we observed a relatively lower performance. But comparing this performance with that of LASER's, we can conclude that the CoHere embedding is better, and this resulted in better parallel sentences.

TABLE II.
PERFORMANCES OF COHERE AND LASER AUTO-ENCODER ON FLORES AND MAFAND-MT DATASETS (MEASURED IN F1-SCORE)

data	# sents.	LASER F1	CoHere F1
FLORES dev	997	3.64%	94.08%
FLORES devtest	1,012	3.36%	94.96%
MAFAND-MT dev	1,300	0.64%	49.19%
MAFAND-MT test	1,500	0.43%	54.83%
MAFAND-MT train	3,098	0.33%	39.27%

B. Monolingual Data

For the evaluation on the crawled monolingual data, Figure 1 shows the distribution of qualities of the paired sentences. It can be seen that while the two aligners majorly generated poor pairings, about 4% of the CoHere aligned sentences are perfect translations of each other. Another 22% of the pairings can be considered translations with varying degrees of accuracy. However, this is in total contrast to the LASER aligned sentences where all of the sentences are not even translations of each other.

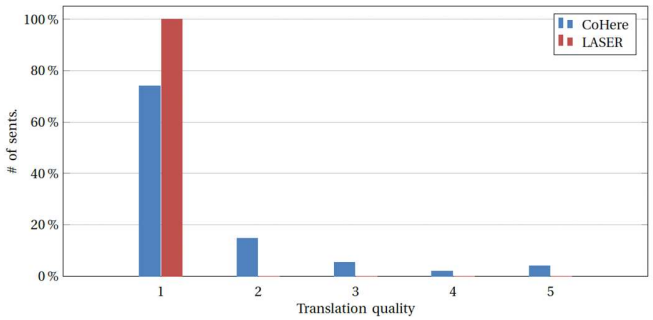

Fig. 1. Distribution of quality after human evaluation of the aligned monolingual crawled sentences.

On further scrutiny, we realized that the CoHere aligner was able to generate sentences that mimic natural translation, where the length ratio of the source and target sentences are similar, an average source to target ratio of 1: 1.2, see Figure 2. Contrastingly, however, the LASER aligner generated about 3 times the lengths of the source sentences (an average ratio of 1: 2.6). Furthermore, only about 3% of the translations are unique, meaning about 97% are the same even though the sources are different. On the other hand, about 30% of the CoHere aligned sentences are unique.

C. Machine Translation

Finally, Table III shows the performance of the various models that were trained for eng → hau and hau → eng translations. Across the test sets, and translation direction, the CoHere sentences were shown to be more beneficial to the model. On MAFAND-MT, we obtained +0.23 and +3.11 improvements on the translation directions respectively. On FLORES, the performance was similar on eng→hau, but even better in the other direction—an improvement of more than +5 BLEU scores.

[8] https://github.com/facebookresearch/LASER

[9] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html

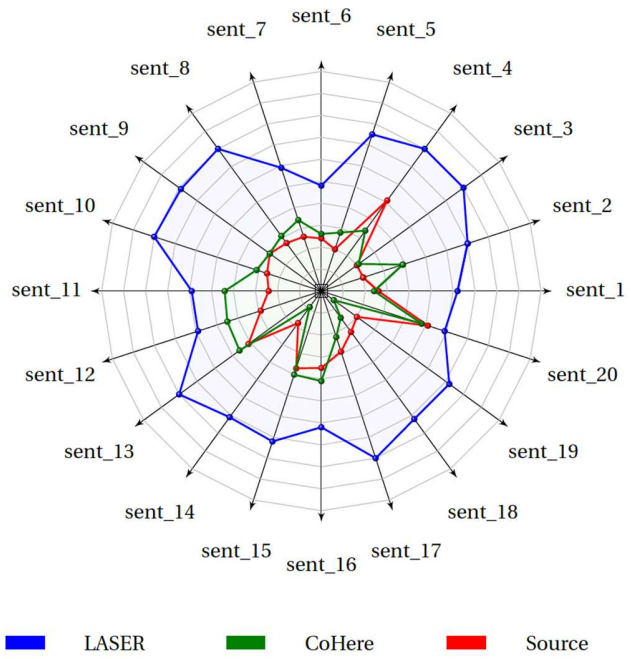

Fig. 2. Demonstrating how the CoHere aligner paired sentences with similar lengths compared to the bigger sentences in the LASER aligner.

V. Conclusion and Future Work

In this work, we showed the efficacy of extracting parallel sentences using a multilingual embedding model for English and Hausa machine translation tasks. We compared the performance of our model with the LASER model and showed that our approach yielded better performance. By this, we showed that the quality of the embeddings used in automatic sentence alignment determines the accuracy of the paired sentences. In the future, we aim to further investigate our approach on more similarity matching algorithms such as inverted nearest neighbor, inverted softmax, and cross-domain similarity local scaling [23] algorithms. We also aim to deploy this parallel data generation technique to improve the performances of other low resource machine translation tasks, such as using other Nigerian languages.

Ethics Statement

The aim of this work was to create parallel sentences for low resource languages and the dataset to be used strictly for research and non-commercial purposes. The CoHere multilingual free tier API and monolingual datasets used in this work allows such usage. This work, therefore, does not raise any ethical concern.

Acknowledgements

This work was made possible by the mentorship program at the Arewa Data Science. The work is a continuation of the participation of a group of mentees in the #CoHEREAIHack, where they finished as the first runners up.

References

[1] D. Adelani, J. Alabi, A. Fan, J. Kreutzer, X. Shen, M. Reid, D. Ruiter, D. Klakow, P. Nabende, E. Chang, T. Gwadabe, F. Sackey, B. F. P. Dossou, C. Emezue, C. Leong, M. Beukman, S. Muhammad, G. Jarso, O. Yousuf, A. Niyongabo Rubungo, G. Hacheme, E. P. Wairagala, M. U. Nasir, B. Ajibade, T. Ajayi, Y. Gitau, J. Abbott, M. Ahmed, M. Ochieng, A. Aremu, P. Ogayo, J. Mukiibi, F. Ouoba Kabore, G. Kalipe, D. Mbaye, A. A. Tapo, V. Memdjokam Koagne, E. Munkoh- Buabeng, V. Wagner, I. Abdulmumin, A. Awokoya, H. Buzaaba, B. Sibanda, A. Bukula, and S. Manthalu, "A few thousand translations go a long way! leveraging pre-trained models for African news translation," in Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 3053–3070. [Online]. Available: https://aclanthology.org/2022.naacl-main.223

[2] D. Adelani, M. M. I. Alam, A. Anastasopoulos, A. Bhagia, M. R. Costajussà, J. Dodge, F. Faisal, C. Federmann, N. Fedorova, F. Guzmán, S. Koshelev, J. Maillard, V. Marivate, J. Mbuya, A. Mourachko, S. Saleem, H. Schwenk, and G. Wenzek, "Findings of the WMT '22 shared task on large-scale machine translation evaluation for African languages," in Proceedings of the Seventh Conference on Machine Translation (WMT). Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, Dec. 2022, pp. 773–800. [Online]. Available: https://aclanthology.org/2022.wmt-1.72

[3] R. Sennrich, B. Haddow, and A. Birch, "Improving neural machine translation models with monolingual data," in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 86–96. [Online]. Available: https://aclanthology.org/P16-1009

[4] I. Abdulmumin, B. S. Galadanci, A. Isa, H. A. Kakudi, and I. I. Sinan, "A Hybrid Approach for Improved Low Resource Neural Machine Translation using Monolingual Data," Engineering Letters, vol. 29, no. 4, pp. 339–350, 2021. [Online]. Available: http://www.engineeringletters.com/issues_v29/issue_4/EL_29_4_21.pdf

[5] M. Bañón, P. Chen, B. Haddow, K. Heafield, H. Hoang, M. Esplà-Gomis, M. L. Forcada, A. Kamran, F. Kirefu, P. Koehn, S. Ortiz Rojas, L. Pla Sempere, G. Ramírez-Sánchez, E. Sarrías, M. Strelec, B. Thompson, W. Waites, D. Wiggins, and J. Zaragoza, "ParaCrawl: Web-scale acquisition of parallel corpora," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, Jul. 2020, pp. 4555–4567. [Online]. Available: https://aclanthology.org/2020.acl-main.417

[6] A. El-Kishky, V. Chaudhary, F. Guzmán, and P. Koehn, "CCAligned: A massive collection of cross-lingual web-document pairs," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020). Online: Association for Computational Linguistics, November 2020, pp. 5960–5969. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-main.480

[7] H. Schwenk, G. Wenzek, S. Edunov, E. Grave, A. Joulin, and A. Fan, "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). Online: Association for Computational Linguistics, Aug. 2021, pp. 6490–6500. [Online]. Available: https://aclanthology.org/2021.acl-long.507

[8] M. Artetxe and H. Schwenk, "Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond," Transactions of the Association for Computational Linguistics, vol. 7, pp. 597–610, 2019. [Online]. Available: https://aclanthology.org/Q19-1038

[9] I. Abdulmumin, M. Beukman, J. Alabi, C. C. Emezue, E. Chimoto, T. Adewumi, S. Muhammad, M. Adeyemi, O. Yousuf, S. Singh, and T. Gwadabe, "Separating grains from the chaff: Using data filtering to improve multilingual translation for low-resourced African languages," in Proceedings of the Seventh Conference on Machine Translation (WMT). Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, Dec. 2022, pp. 1001–1014. [Online]. Available: https://aclanthology.org/2022.wmt-1.98

[10] J. Kreutzer, I. Caswell, L. Wang, A. Wahab, D. van Esch, N. Ulzii-Orshikh, A. Tapo, N. Subramani, A. Sokolov, C. Sikasote, M. Setyawan, S. Sarin, S. Samb, B. Sagot, C. Rivera, A. Rios, I. Papadimitriou, S. Osei, P. O. Suarez, I. Orife, K. Ogueji, A. N. Rubungo, T. Q. Nguyen, M. Müller, A. Müller, S. H. Muhammad, N. Muhammad, A. Mnyakeni, J. Mirzakhalov, T. Matangira, C. Leong, N. Lawson, S. Kudugunta, Y. Jernite, M. Jenny, O. Firat, B. F. P. Dossou, S. Dlamini, N. de Silva, S. Çabuk Ballı, S. Biderman, A. Battisti, A. Baruwa, A. Bapna, P. Baljekar, I. A. Azime, A. Awokoya, D. Ataman, O. Ahia, O. Ahia, S. Agrawal, and M. Adeyemi, "Quality at a glance: An audit of web-crawled multilingual datasets," Transactions of the Association for Computational Linguistics, vol. 10, pp. 50–72, 2022. [Online]. Available: https://aclanthology.org/2022.tacl-1.4

[11] H. Schwenk, V. Chaudhary, S. Sun, H. Gong, and F. Guzmán, "Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. arxiv cs," 1907.
[12] K. Heffernan, O. Çelebi, and H. Schwenk, "Bitext mining using distilled sentence representations for low-resource languages," 2022.
[13] A. Conneau, G. Lample, R. Rinott, A. Williams, S. R. Bowman, H. Schwenk, and V. Stoyanov, "Xnli: Evaluating cross-lingual sentence representations," 2018.
[14] H. Schwenk and X. Li, "A corpus for multilingual document classification in eight languages," 2018.
[15] P.-A. Duquenne, H. Gong, and H. Schwenk, "Multimodal and multilingual embeddings for large-scale speech mining," Advances in Neural Information Processing Systems, vol. 34, pp. 15 748–15 761, 2021.
[16] M. Artetxe, G. Labaka, and E. Agirre, "Learning principled bilingual mappings of word embeddings while preserving monolingual invariance," in Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 2289–2294. [Online]. Available: https://aclanthology.org/D16-1250
[17] S. Bird, E. Klein, and E. Loper, Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc.", 2009.
[18] N. Goyal, C. Gao, V. Chaudhary, P.-J. Chen, G. Wenzek, D. Ju, S. Krishnan, M. Ranzato, F. Guzmán, and A. Fan, "The Flores-101 evaluation benchmark for low-resource and multilingual machine translation," Transactions of the Association for Computational Linguistics, vol. 10, pp. 522–538, 2022. [Online]. Available: https://aclanthology.org/2022.tacl-1.30
[19] A. Fan, S. Bhosale, H. Schwenk, Z. Ma, A. El-Kishky, S. Goyal, M. Baines, O. Celebi, G. Wenzek, V. Chaudhary, N. Goyal, T. Birch, V. Liptchinsky, S. Edunov, E. Grave, M. Auli, and A. Joulin, "Beyond english-centric multilingual machine translation," J. Mach. Learn. Res., vol. 22, no. 1, jan 2021.
[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in Proceedings of the 31st International Conference on Neural Information Processing Systems, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6000–6010.
[21] 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. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, Jul. 2002, pp. 311–318. [Online]. Available: https://aclanthology.org/P02-1040
[22] M. Post, "A call for clarity in reporting BLEU scores," in Proceedings of the Third Conference on Machine Translation: Research Papers. Belgium, Brussels: Association for Computational Linguistics, Oct. 2018, pp. 186–191. [Online]. Available: https://www.aclweb.org/anthology/W18-6319
[23] G. Lample, A. Conneau, M. Ranzato, L. Denoyer, and H. Jégou, "Word translation without parallel data," in International Conference on Learning Representations, 2018.