

REFERENCES

- [1] T. Mikolov, Q. V. Le, and I. Sutskever, “Exploiting similarities among languages for machine translation,” *arXiv preprint arXiv:1309.4168*, 2013.
- [2] A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou, “Word translation without parallel data,” *arXiv preprint arXiv:1710.04087*, 2017.
- [3] 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)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 878–891. [Online]. Available: <https://aclanthology.org/2022.acl-long.62>
- [4] A. Conneau and G. Lample, “Cross-lingual language model pretraining,” *NIPS*, vol. 32, 2019.
- [5] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *arXiv preprint arXiv:1301.3781*, 2013.
- [6] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching word vectors with subword information,” *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017. [Online]. Available: <https://aclanthology.org/Q17-1010>
- [7] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. 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*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 8440–8451. [Online]. Available: <https://aclanthology.org/2020.acl-main.747>
- [8] N. de Silva, “Survey on publicly available sinhala natural language processing tools and research,” *arXiv preprint arXiv:1906.02358*, 2019.
- [9] S. Ranathunga and N. de Silva, “Some languages are more equal than others: Probing deeper into the linguistic disparity in the NLP world,” in *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Y. He, H. Ji, S. Li, Y. Liu, and C.-H. Chang, Eds. Online only: Association

for Computational Linguistics, Nov. 2022, pp. 823–848. [Online]. Available: <https://aclanthology.org/2022.aacl-main.62>

- [10] S. L. Smith, D. H. Turban, S. Hamblin, and N. Y. Hammerla, “Offline bilingual word vectors, orthogonal transformations and the inverted softmax,” in *International Conference on Learning Representations*, 2016.
- [11] A. Liyanage, S. Ranathunga, and S. Jayasena, “Bilingual lexical induction for sinhala-english using cross lingual embedding spaces,” in *2021 Moratuwa Engineering Research Conference (MERCon)*, 2021, pp. 579–584.
- [12] 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. Mirzakhlov, 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>
- [13] G. Salton and C. Buckley, “Term-weighting approaches in automatic text retrieval,” *Information processing & management*, vol. 24, no. 5, pp. 513–523, 1988.
- [14] J. Pennington, R. Socher, and C. Manning, “GloVe: Global vectors for word representation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, A. Moschitti, B. Pang, and W. Daelemans, Eds. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. [Online]. Available: <https://aclanthology.org/D14-1162>
- [15] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, “Deep contextualized word representations,” in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, M. Walker, H. Ji, and A. Stent, Eds. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 2227–2237. [Online]. Available: <https://aclanthology.org/N18-1202>

- [16] 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)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: <https://aclanthology.org/N19-1423>
- [17] A. Joulin, P. Bojanowski, T. Mikolov, H. Jégou, and E. Grave, “Loss in translation: Learning bilingual word mapping with a retrieval criterion,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, Eds. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 2979–2984. [Online]. Available: <https://aclanthology.org/D18-1330>
- [18] C. Xing, D. Wang, C. Liu, and Y. Lin, “Normalized word embedding and orthogonal transform for bilingual word translation,” in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, R. Mihalcea, J. Chai, and A. Sarkar, Eds. Denver, Colorado: Association for Computational Linguistics, May–Jun. 2015, pp. 1006–1011. [Online]. Available: <https://aclanthology.org/N15-1104>
- [19] A. Conneau and G. Lample, “Cross-lingual language model pretraining,” *Advances in neural information processing systems*, vol. 32, 2019.
- [20] N. Reimers and I. Gurevych, “Making monolingual sentence embeddings multilingual using knowledge distillation,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 4512–4525. [Online]. Available: <https://aclanthology.org/2020.emnlp-main.365>
- [21] 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>
- [22] K. Heffernan, O. Çelebi, and H. Schwenk, “Bitext mining using distilled sentence representations for low-resource languages,” in *Findings of the Association for Computational Linguistics: EMNLP 2022*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 2101–2112. [Online]. Available: <https://aclanthology.org/2022.findings-emnlp.154>

- [23] Y. J. Kim, A. A. Awan, A. Muzio, A. F. C. Salinas, L. Lu, A. Hendy, S. Rajbhandari, Y. He, and H. H. Awadalla, “Scalable and efficient moe training for multitask multilingual models,” *arXiv preprint arXiv:2109.10465*, 2021.
- [24] Y. Tang, C. Tran, X. Li, P.-J. Chen, N. Goyal, V. Chaudhary, J. Gu, and A. Fan, “Multilingual translation from denoising pre-training,” in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Online: Association for Computational Linguistics, Aug. 2021, pp. 3450–3466. [Online]. Available: <https://aclanthology.org/2021.findings-acl.304>
- [25] J. Kreutzer, I. Caswell, L. Wang, A. Wahab, D. van Esch, N. Ulzii-Orshikh, A. Tapo, N. Subramani, A. Sokolov, C. Sikasote *et al.*, “Quality at a glance: An audit of web-crawled multilingual datasets,” *TACL*, vol. 10, pp. 50–72, 2022.
- [26] I. Wijesiri, M. Gallage, B. Gunathilaka, M. Lakjeewa, D. Wimalasuriya, G. Dias, R. Paranavithana, and N. De Silva, “Building a wordnet for sinhala,” in *Proceedings of the seventh global wordnet conference*, 2014, pp. 100–108.
- [27] S. Ranathunga and N. de Silva, “Some languages are more equal than others: Probing deeper into the linguistic disparity in the nlp world,” *arXiv preprint arXiv:2210.08523*, 2022.
- [28] F. Guzmán, P.-J. Chen, M. Ott, J. Pino, G. Lample, P. Koehn, V. Chaudhary, and M. Ranzato, “The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English,” in *EMNLP-IJCNLP*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 6098–6111. [Online]. Available: <https://aclanthology.org/D19-1632>
- [29] M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard *et al.*, “No language left behind: Scaling human-centered machine translation,” *arXiv preprint arXiv:2207.04672*, 2022.
- [30] R. A. Hameed, N. Pathirennelhelage, A. Ihalapathirana, M. Z. Mohamed, S. Ranathunga, S. Jayasena, G. Dias, and S. Fernando, “Automatic creation of a sentence aligned sinhala-tamil parallel corpus,” in *Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WS-SANLP2016)*, 2016, pp. 124–132.
- [31] M. Bañón, P. Chen, B. Haddow, K. Heafield, H. Hoang, M. Esplà-Gomis, M. L. Forcada, A. Kamran, F. Kirefu, P. Koehn *et al.*, “Paracrawl: Web-scale acquisition of parallel corpora,” in *ACL*, 2020, pp. 4555–4567.

- [32] C. Vasantharajan and U. Thayasivam, “Tamizhi-net ocr: Creating a quality large scale tamil-sinhala-english parallel corpus using deep learning based printed character recognition (pcr),” *arXiv preprint arXiv:2109.05952*, 2021.
- [33] A. Wasala and R. Weerasinghe, “Ensitiip: a tool to unlock the english web,” in *11th international conference on humans and computers, Nagaoka University of Technology, Japan*, 2008, pp. 20–23.
- [34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *NIPS*, vol. 30, 2017.
- [35] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, “Albert: A lite bert for self-supervised learning of language representations,” *arXiv preprint arXiv:1909.11942*, 2019.
- [36] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized bert pretraining approach,” *arXiv preprint arXiv:1907.11692*, 2019.
- [37] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, “Electra: Pre-training text encoders as discriminators rather than generators,” *arXiv preprint arXiv:2003.10555*, 2020.
- [38] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence embeddings using Siamese BERT-networks,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3982–3992. [Online]. Available: <https://aclanthology.org/D19-1410>
- [39] K. Ahuja, S. Sitaram, S. Dandapat, and M. Choudhury, “On the calibration of massively multilingual language models,” in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 4310–4323. [Online]. Available: <https://aclanthology.org/2022.emnlp-main.290>
- [40] A. Cooper Stickland, S. Sengupta, J. Krone, S. Mansour, and H. He, “Robustification of multilingual language models to real-world noise in crosslingual zero-shot settings with robust contrastive pretraining,” in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, A. Vlachos and I. Augenstein, Eds. Dubrovnik, Croatia: Association

for Computational Linguistics, May 2023, pp. 1375–1391. [Online]. Available: <https://aclanthology.org/2023.eacl-main.100>

- [41] M. Artetxe, G. Labaka, and E. Agirre, “Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations,” in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018, pp. 5012–5019.
- [42] Y. Li, F. Liu, N. Collier, A. Korhonen, and I. Vulić, “Improving word translation via two-stage contrastive learning,” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 4353–4374. [Online]. Available: <https://aclanthology.org/2022.acl-long.299>
- [43] A. v. d. Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” *arXiv preprint arXiv:1807.03748*, 2018.
- [44] M. Artetxe, G. Labaka, and E. Agirre, “A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, I. Gurevych and Y. Miyao, Eds. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 789–798. [Online]. Available: <https://aclanthology.org/P18-1073>
- [45] E. Grave, A. Joulin, and Q. Berthet, “Unsupervised alignment of embeddings with wasserstein procrustes,” in *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2019, pp. 1880–1890.
- [46] P. O. Aboagye, Y. Zheng, M. Yeh, J. Wang, Z. Zhuang, H. Chen, L. Wang, W. Zhang, and J. Phillips, “Quantized Wasserstein Procrustes alignment of word embedding spaces,” in *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, K. Duh and F. Guzmán, Eds. Orlando, USA: Association for Machine Translation in the Americas, Sep. 2022, pp. 200–214. [Online]. Available: <https://aclanthology.org/2022.amta-research.15>
- [47] F. Guzmán, P.-J. Chen, M. Ott, J. Pino, G. Lample, P. Koehn, V. Chaudhary, and M. Ranzato, “The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Hong Kong,

China: Association for Computational Linguistics, Nov. 2019, pp. 6098–6111. [Online]. Available: <https://aclanthology.org/D19-1632>

- [48] 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. Sarriás, 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*, D. Jurafsky, J. Chai, N. Schlueter, and J. Tetreault, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 4555–4567. [Online]. Available: <https://aclanthology.org/2020.acl-main.417>
- [49] C. Vasantharajan, L. Tharmalingam, and U. Thayasivam, “Adapting the tesseract open-source ocr engine for tamil and sinhala legacy fonts and creating a parallel corpus for tamil-sinhala-english,” in *2022 International Conference on Asian Language Processing (IALP)*. IEEE, 2022, pp. 143–149.
- [50] K. Wickramasinghe and N. de Silva, “Sinhala-English parallel word dictionary dataset,” in *IEEE 17th International Conference on Industrial and Information Systems (ICIIS)*, Peradeniya, Sri Lanka, Aug. 2023, p. 6.
- [51] H. S. Saadi, V. Hangya, T. Eder, and A. Fraser, “Comparative analysis of cross-lingual contextualized word embeddings,” in *Proceedings of the The 2nd Workshop on Multi-lingual Representation Learning (MRL)*, D. Ataman, H. Gonen, S. Ruder, O. Firat, G. Gül Sahin, and J. Mirzakhlov, Eds. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, Dec. 2022, pp. 64–75. [Online]. Available: <https://aclanthology.org/2022.mrl-1.6>
- [52] A. Irvine and C. Callison-Burch, “A comprehensive analysis of bilingual lexicon induction,” *Computational Linguistics*, vol. 43, no. 2, pp. 273–310, Jun. 2017. [Online]. Available: <https://aclanthology.org/J17-2001>
- [53] A. Conneau, R. Rinott, G. Lample, A. Williams, S. Bowman, H. Schwenk, and V. Stoyanov, “XNLI: Evaluating cross-lingual sentence representations,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, Eds. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 2475–2485. [Online]. Available: <https://aclanthology.org/D18-1269>
- [54] J. Zhang, B. Ji, N. Xiao, X. Duan, M. Zhang, Y. Shi, and W. Luo, “Combining static word embeddings and contextual representations for bilingual lexicon induction,” in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Online:

Association for Computational Linguistics, Aug. 2021, pp. 2943–2955. [Online]. Available: <https://aclanthology.org/2021.findings-acl.260>

- [55] K. Hämmerl, J. Libovický, and A. Fraser, “Combining static and contextualised multilingual embeddings,” in *Findings of the Association for Computational Linguistics: ACL 2022*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 2316–2329. [Online]. Available: <https://aclanthology.org/2022.findings-acl.182>
- [56] G. Glavaš, R. Litschko, S. Ruder, and I. Vulić, “How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, A. Korhonen, D. Traum, and L. Màrquez, Eds. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 710–721. [Online]. Available: <https://aclanthology.org/P19-1070>
- [57] I. Vulić, G. Glavaš, R. Reichart, and A. Korhonen, “Do we really need fully unsupervised cross-lingual embeddings?” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 4407–4418. [Online]. Available: <https://aclanthology.org/D19-1449>
- [58] K. Wickramasinghe and N. de Silva, “Sinhala-English word embedding alignment: Introducing datasets and benchmark for a low resource language,” in *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, C.-R. Huang, Y. Harada, J.-B. Kim, S. Chen, Y.-Y. Hsu, E. Chersoni, P. A. W. H. Zeng, B. Peng, Y. Li, and J. Li, Eds. Hong Kong, China: Association for Computational Linguistics, Dec. 2023, pp. 424–435. [Online]. Available: <https://aclanthology.org/2023.paclic-1.42>
- [59] A. Magueresse, V. Carles, and E. Heetderks, “Low-resource languages: A review of past work and future challenges,” *arXiv preprint arXiv:2006.07264*, 2020.
- [60] M. A. Hedderich, L. Lange, H. Adel, J. Strötgen, and D. Klakow, “A survey on recent approaches for natural language processing in low-resource scenarios,” *arXiv preprint arXiv:2010.12309*, 2020.
- [61] C. Xing, D. Wang, C. Liu, and Y. Lin, “Normalized word embedding and orthogonal transform for bilingual word translation,” in *NAACL:HLT*, 2015, pp. 1006–1011.

- [62] A. Joulin, P. Bojanowski, T. Mikolov, H. Jégou, and E. Grave, “Loss in translation: Learning bilingual word mapping with a retrieval criterion,” in *EMNLP*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 2979–2984. [Online]. Available: <https://aclanthology.org/D18-1330>
- [63] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, “Unsupervised cross-lingual representation learning at scale,” in *ACL*. Online: Association for Computational Linguistics, jul 2020, pp. 8440–8451.
- [64] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, “Language-agnostic BERT Sentence Embedding,” in *ACL*, 2022, pp. 878–891.
- [65] A. Fernando, S. Ranathunga, D. Sachintha, L. Piyaarathna, and C. Rajitha, “Exploiting bilingual lexicons to improve multilingual embedding-based document and sentence alignment for low-resource languages,” *Knowledge and Information Systems*, pp. 1–42, 2022.
- [66] X. Saralegi, I. Manterola, and I. San Vicente, “Analyzing methods for improving precision of pivot based bilingual dictionaries,” in *EMNLP*, 2011, pp. 846–856.
- [67] A. H. Nasution, Y. Murakami, and T. Ishida, “Constraint-based bilingual lexicon induction for closely related languages,” in *LREC*, 2016, pp. 3291–3298.
- [68] M. Wushouer, D. Lin, T. Ishida, and K. Hirayama, “A constraint approach to pivot-based bilingual dictionary induction,” *TALLIP*, vol. 15, no. 1, pp. 1–26, 2015.
- [69] P. F. Brown, J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, J. Lafferty, R. L. Mercer, and P. S. Roossin, “A statistical approach to machine translation,” *Computational linguistics*, vol. 16, no. 2, pp. 79–85, 1990.
- [70] 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*, 2016, pp. 679–684.
- [71] 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*, 2016, pp. 697–702.
- [72] C. Buck and P. Koehn, “Quick and reliable document alignment via tf/idf-weighted cosine distance,” in *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, 2016, pp. 672–678.

- [73] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” *arXiv preprint arXiv:1409.0473*, 2014.
- [74] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey *et al.*, “Google’s neural machine translation system: Bridging the gap between human and machine translation,” *arXiv preprint arXiv:1609.08144*, 2016.
- [75] S. Thillainathan, S. Ranathunga, and S. Jayasena, “Fine-Tuning Self-Supervised Multilingual Sequence-To-Sequence Models for Extremely Low-Resource NMT,” in *2021 Moratuwa Engineering Research Conference (MERCon)*. IEEE, 2021, pp. 432–437.
- [76] A. Fernando and S. Ranathunga, “Data augmentation to address out-of-vocabulary problem in low-resource sinhala-english neural machine translation,” *arXiv preprint arXiv:2205.08722*, 2022.
- [77] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching word vectors with subword information,” *TACL*, vol. 5, pp. 135–146, 2017.
- [78] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, “Bag of tricks for efficient text classification,” *arXiv preprint arXiv:1607.01759*, 2016.
- [79] R. Perera, T. Fonseka, R. Naranpanawa, and U. Thayasivam, “Improving english to sinhala neural machine translation using part-of-speech tag,” *arXiv preprint arXiv:2202.08882*, 2022.
- [80] N. de Silva, “Sinhala text classification: observations from the perspective of a resource poor language,” *ResearchGate*, 2015.
- [81] D. Upeksha, C. Wijayarathna, M. Siriwardena, L. Lasandun, C. Wimalasuriya, N. De Silva, and G. Dias, “Implementing a corpus for sinhala language,” in *Symposium on Language Technology for South Asia 2015*, 2015.
- [82] D. Upeksha, C. Wijayarathna, M. Siriwardena, L. Lasandun, C. Wimalasuriya, N. de Silva, and G. Dias, “Comparison between performance of various database systems for implementing a language corpus,” in *International Conference: Beyond Databases, Architectures and Structures*. Springer, 2015, pp. 82–91.
- [83] J. Tiedemann, “Parallel data, tools and interfaces in OPUS,” in *LREC*, May 2012, pp. 2214–2218.
- [84] N. Reimers and I. Gurevych, “Making monolingual sentence embeddings multilingual using knowledge distillation,” in *EMNLP*, 2020, pp. 4512–4525.

- [85] C.-Y. Lin, “Rouge: A package for automatic evaluation of summaries,” in *Text summarization branches out*, 2004, pp. 74–81.
- [86] D. Lakmal, S. Ranathunga, S. Peramuna, and I. Herath, “Word embedding evaluation for sinhala,” in *LREC*, 2020, pp. 1874–1881.
- [87] A. Søgaard, S. Ruder, and I. Vulić, “On the limitations of unsupervised bilingual dictionary induction,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, I. Gurevych and Y. Miyao, Eds. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 778–788. [Online]. Available: <https://aclanthology.org/P18-1072>
- [88] 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)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 5960–5969. [Online]. Available: <https://aclanthology.org/2020.emnlp-main.480>
- [89] J. Tiedemann, “Finding alternative translations in a large corpus of movie subtitle,” in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, Eds. Portorož, Slovenia: European Language Resources Association (ELRA), May 2016, pp. 3518–3522. [Online]. Available: <https://aclanthology.org/L16-1559>
- [90] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, “Bag of tricks for efficient text classification,” in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, M. Lapata, P. Blunsom, and A. Koller, Eds. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 427–431. [Online]. Available: <https://aclanthology.org/E17-2068>
- [91] D. Lakmal, S. Ranathunga, S. Peramuna, and I. Herath, “Word embedding evaluation for Sinhala,” in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, Eds. Marseille, France: European Language Resources Association, May 2020, pp. 1874–1881. [Online]. Available: <https://aclanthology.org/2020.lrec-1.231>

- [92] M. S. Maučec, Z. Kačič, and B. Horvat, “Modelling highly inflected languages,” *Information Sciences*, vol. 166, no. 1-4, pp. 249–269, 2004.
- [93] E. Kuriyozov, Y. Doval, and C. Gómez-Rodríguez, “Cross-lingual word embeddings for Turkic languages,” in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, Eds. Marseille, France: European Language Resources Association, May 2020, pp. 4054–4062. [Online]. Available: <https://aclanthology.org/2020.lrec-1.499>
- [94] M. Izbicki, “Aligning word vectors on low-resource languages with Wiktionary,” in *Proceedings of the Fifth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2022)*, A. K. Ojha, C.-H. Liu, E. Vylomova, J. Abbott, J. Washington, N. Oco, T. A. Pirinen, V. Malykh, V. Logacheva, and X. Zhao, Eds. Gyeongju, Republic of Korea: Association for Computational Linguistics, Oct. 2022, pp. 107–117. [Online]. Available: <https://aclanthology.org/2022.loresmt-1.14>