REFERENCES

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.

Abdulmumin, I., Galadanci, B. S., and Isa, A. (2020). Enhanced back-translation for low resource neural machine translation using self-training. In International Conference on Information and Communication Technology and Applications, pages 355–371. Springer.

Açarçiçek, H., Çolakoğlu, T., Hatipoğlu, P. E. A., Huang, C. H., and Peng, W. (2020). Filtering noisy parallel corpus using transformers with proxy task learning. In Proceedings of the Fifth Conference on Machine Translation, pages 940–946.

Akbik, A., Blythe, D., and Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In COLING 2018, 27th International Conference on Computational Linguistics, pages 1638–1649.

Alam, M. M. I., Ahmadi, S., and Anastasopoulos, A. (2024). A morphologically-aware dictionary-based data augmentation technique for machine translation of under-represented languages. arXiv preprint arXiv:2402.01939.

Allen B. Tucker, J. and Nirenburg, S. (1984). Machine translation: A contemporary view. Annual Review of Information Science and Technology, 19:129.

Aoyama, T. and Schneider, N. (2022). Probe-less probing of bert's layer-wise linguistic knowledge with masked word prediction. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop, pages 195–201.

Araabi, A. and Monz, C. (2020). Optimizing transformer for low-resource neural machine translation. In Proceedings of the 28th International Conference on Computational Linguistics, pages 3429–3435.

Artetxe, M., Labaka, G., and Agirre, E. (2018a). Unsupervised statistical machine translation. In ACL.

Artetxe, M., Labaka, G., Lopez-Gazpio, I., and Agirre, E. (2018b). Uncovering divergent linguistic information in word embeddings with lessons for intrinsic and extrinsic evaluation. arXiv preprint arXiv:1809.02094.

Artetxe, M. and Schwenk, H. (2019a). Margin-based parallel corpus mining with multilingual sentence embeddings. In Proceedings of the 57th Annual Meeting of

the Association for Computational Linguistics, pages 3197–3203. Association for Computational Linguistics.

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

Aulamo, M., De Gibert, O., Virpioja, S., and Tiedemann, J. (2023). Unsupervised feature selection for effective parallel corpus filtering. In Proceedings of the 24th Annual Conference of the European Association for Machine Translation, pages 31–38.

Aulamo, M., Virpioja, S., and Tiedemann, J. (2020). Opusfilter: A configurable parallel corpus filtering toolbox. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 150–156.

Azpeitia, A., Etchegoyhen, T., and Garcia, E. M. (2017). Weighted set-theoretic alignment of comparable sentences. In Proceedings of the 10th Workshop on Building and Using Comparable Corpora, pages 41–45.

Azpeitia, A., Etchegoyhen, T., and Garcia, E. M. (2018). Extracting parallel sentences from comparable corpora with stacc variants. In Proceedings of the 11th Workshop on Building and Using Comparable Corpora, pages 48–52.

Bahdanau, D., Cho, K. H., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In 3rd International Conference on Learning Representations, ICLR 2015.

Bala Das, S., Biradar, A., Kumar Mishra, T., and Kr. Patra, B. (2023). Improving multilingual neural machine translation system for indic languages. ACM Transactions on Asian and Low-Resource Language Information Processing, 22(6):1–24.

Bane, F., Uguet, C. S., Stribiżew, W., and Zaretskaya, A. (2022). A comparison of data filtering methods for neural machine translation. In Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas (Volume 2: Users and Providers Track and Government Track), pages 313–325.

Bañón, M., Chen, P., Haddow, B., Heafield, K., Hoang, H., Esplà-Gomis, M., Forcada, M. L., Kamran, A., Kirefu, F., Koehn, P., et al. (2020). Paracrawl: Web-scale acquisition of parallel corpora. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4555–4567.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606.

Bouamor, H. and Sajjad, H. (2018). H2@ bucc18: Parallel sentence extraction from comparable corpora using multilingual sentence embeddings. In Proc. Workshop on Building and Using Comparable Corpora, pages 43–47.

Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., and Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. Computational linguistics, 19(2):263–311.

Brown, P. F., Lai, J. C., and Mercer, R. L. (1991). Aligning sentences in parallel corpora. In 29th Annual Meeting of the Association for Computational Linguistics, pages 169–176, Berkeley, California, USA. Association for Computational Linguistics.

Buck, C. and Koehn, P. (2016a). Findings of the WMT 2016 bilingual document alignment shared task. In Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers, pages 554–563, Berlin, Germany. Association for Computational Linguistics.

Buck, C. and Koehn, P. (2016b). 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, pages 672–678.

Burchell, L., de Gibert, O., Arefyev, N., Aulamo, M., Bañón, M., Fedorova, M., Guillou, L., Haddow, B., Hajič, J., Henriksson, E., et al. (2025). An expanded massive multilingual dataset for high-performance language technologies. arXiv preprint arXiv:2503.10267.

Carlson, L. and Vilkuna, M. (1990). Independent transfer using graph unification. In COLING 1990 Volume 3: Papers presented to the 13th International Conference on Computational Linguistics.

Caswell, I., Chelba, C., and Grangier, D. (2019). Tagged back-translation. In Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers), page 53. Association for Computational Linguistics.

Chaudhary, V., Tang, Y., Guzmán, F., Schwenk, H., and Koehn, P. (2019). Low-resource corpus filtering using multilingual sentence embeddings. WMT 2019, page 261.

Chen, J. and Nie, J.-Y. (2000). Parallel web text mining for cross-language ir. In Content-Based Multimedia Information Access-Volume 1, pages 62–77. RIAO.

Chen, J., Tam, D., Raffel, C., Bansal, M., and Yang, D. (2023). An empirical survey of data augmentation for limited data learning in nlp. Transactions of the Association for Computational Linguistics, 11:191–211.

Cho, K., van Merriënboer, B., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). 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), pages 1724–1734.

Choi, H., Kim, J., Joe, S., Min, S., and Gwon, Y. (2021). Analyzing zero-shot cross-lingual transfer in supervised nlp tasks. In 2020 25th International Conference on Pattern Recognition (ICPR), pages 9608–9613. IEEE.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, É., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020a). Unsupervised cross-lingual representation learning at scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8440–8451.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, É., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020b). Unsupervised cross-lingual representation learning at scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8440–8451.

Conneau, A. and Lample, G. (2019). Cross-lingual language model pretraining. Advances in neural information processing systems, 32.

Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., et al. (2022). No language left behind: Scaling human-centered machine translation. arXiv preprint arXiv:2207.04672.

Dabre, R., Chu, C., and Kunchukuttan, A. (2020). A survey of multilingual neural machine translation. ACM Computing Surveys (CSUR), 53(5):1–38.

Dabre, R., Fujita, A., and Chu, C. (2019). Exploiting multilingualism through multi-stage fine-tuning for low-resource neural machine translation. 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), pages 1410–1416.

Dara, A. A. and Lin, Y.-C. (2016). Yoda system for wmt16 shared task: Bilingual document alignment. In Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers, pages 679–684.

De Gibert, O., Nail, G., Arefyev, N., Bañón, M., Van Der Linde, J., Ji, S., Zaragoza-Bernabeu, J., Aulamo, M., Ramírez-Sánchez, G., Kutuzov, A., et al. (2024). A new massive multilingual dataset for high-performance language technologies. In Proceedings of the 2024 Joint International Conference on Computational

Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 1116–1128.

de Silva, N. (2019). Survey on publicly available sinhala natural language processing tools and research. arXiv preprint arXiv:1906.02358.

de Silva, N. (2023). Survey on Publicly Available Sinhala Natural Language Processing Tools and Research. arXiv preprint arXiv:1906.02358v20.

de Silva, N. (2025). Survey on Publicly Available Sinhala Natural Language Processing Tools and Research. arXiv preprint arXiv:1906.02358v24.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019a). 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), pages 4171–4186.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019b). 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), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dhananjaya, V., Demotte, P., Ranathunga, S., and Jayasena, S. (2022). Bertifying sinhala-a comprehensive analysis of pre-trained language models for sinhala text classification. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 7377–7385.

Dhar, P., Bisazza, A., and van Noord, G. (2021). Optimal word segmentation for neural machine translation into dravidian languages. In Proceedings of the 8th Workshop on Asian Translation (WAT2021), pages 181–190.

Duan, S., Zhao, H., Zhang, D., and Wang, R. (2020). Syntax-aware data augmentation for neural machine translation. arXiv preprint arXiv:2004.14200.

El-Kishky, A., Chaudhary, V., Guzmán, F., and Koehn, P. (2020). Ccaligned: A massive collection of cross-lingual web-document pairs. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 5960–5969.

El-Kishky, A. and Guzmán, F. (2020). Massively multilingual document alignment with cross-lingual sentence-mover's distance. In Proceedings of the 1st Conference of the

Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, pages 616–625, Suzhou, China. Association for Computational Linguistics.

Epaliyana, K., Ranathunga, S., and Jayasena, S. (2021). Improving back-translation with iterative filtering and data selection for sinhala-english nmt. In 2021 Moratuwa Engineering Research Conference (MERCon), pages 438–443. IEEE.

Espla-Gomis, M., Forcada, M. L., Ortiz-Rojas, S., and Ferrández-Tordera, J. (2016). Bitextor's participation in wmt'16: shared task on document alignment. In Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers, pages 685–691.

Etchegoyhen, T. and Gete, H. (2020). Handle with care: A case study in comparable corpora exploitation for neural machine translation. In Proceedings of The 12th Language Resources and Evaluation Conference, pages 3799–3807.

Fadaee, M., Bisazza, A., and Monz, C. (2017). Data augmentation for low-resource neural machine translation. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 567–573.

Fadaee, M. and Monz, C. (2018). Back-translation sampling by targeting difficult words in neural machine translation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 436–446.

Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., Baines, M., Celebi, O., Wenzek, G., Chaudhary, V., Goyal, N., Birch, T., Liptchinsky, V., Edunov, S., Grave, E., Auli, M., and Joulin, A. (2020). Beyond english-centric multilingual machine translation.

Farhath, F., Ranathunga, S., Jayasena, S., and Dias, G. (2018a). Integration of bilingual lists for domain-specific statistical machine translation for sinhala-tamil. In 2018 Moratuwa Engineering Research Conference (MERCon), pages 538–543. IEEE.

Farhath, F., Theivendiram, P., Ranathunga, S., Jayasena, S., and Dias, G. (2018b). Improving domain-specific smt for low-resourced languages using data from different domains. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).

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

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., and Wang, W. (2022). Language-agnostic bert sentence embedding. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 878–891.

Fernando, A. and Dias, G. (2021). Building a linguistic resource: A word frequency list for sinhala. In Proceedings of the 18th International Conference on Natural Language Processing (ICON), pages 606–610.

Fernando, A. and Ranathunga, S. (2021). Data augmentation to address out of vocabularyproblem in low resource sinhala english neural machine translation. In Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation, pages 61–70.

Fernando, A. and Ranathunga, S. (2025). Linguistic entity masking to improve cross-lingual representation of multilingual language models for low-resource languages. Knowledge and Information Systems.

Fernando, A., Ranathunga, S., and de Silva, N. (2025). Improving the quality of web-mined parallel corpora of low-resource languages using debiasing heuristics. arXiv preprint arXiv:2502.19074.

Fernando, A., Ranathunga, S., and Dias, G. (2020). Data augmentation and terminology integration for domain-specific sinhala-english-tamil statistical machine translation. arXiv preprint arXiv:2011.02821.

Fernando, A., Ranathunga, S., Sachintha, D., Piyarathna, L., and Rajitha, C. (2023). Exploiting bilingual lexicons to improve multilingual embedding-based document and sentence alignment for low-resource languages. Knowledge and Information Systems, 65(2):571–612.

Fernando, S. and Ranathunga, S. (2018). Evaluation of different classifiers for sinhala pos tagging. In 2018 Moratuwa Engineering Research Conference (MERCon), pages 96–101. IEEE.

Fernando, S., Ranathunga, S., Jayasena, S., and Dias, G. (2016). Comprehensive part-of-speech tag set and svm based pos tagger for sinhala. In Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP2016), pages 173–182.

Fonseka, T., Naranpanawa, R., Perera, R., and Thayasivam, U. (2020). English to sinhala neural machine translation. In 2020 International Conference on Asian Language Processing (IALP), pages 305–309. IEEE.

Fung, P. and Cheung, P. (2004). Mining very-non-parallel corpora: Parallel sentence and lexicon extraction via bootstrapping and e. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pages 57–63.

Gala, J., Chitale, P. A., AK, R., Gumma, V., Doddapaneni, S., Kumar, A., Nawale, J., Sujatha, A., Puduppully, R., Raghavan, V., et al. (2023). Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. arXiv preprint arXiv:2305.16307.

Gale, W. A. and Church, K. (1993). A program for aligning sentences in bilingual corpora. Computational Linguistics, 19(1):75–102.

Gao, Y., Hou, F., Jahnke, H., and Wang, R. (2023). Data augmentation with diversified rephrasing for low-resource neural machine translation. In Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track, pages 35–47.

Garcia, X., Niu, Y., and Specia, L. (2023). Low-resource domain-robust unsupervised machine translation via multi-phase adaptation. In Findings of ACL.

Germann, U. (2016). Bilingual document alignment with latent semantic indexing. In Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers, pages 692–696, Berlin, Germany. Association for Computational Linguistics.

Golchin, S., Surdeanu, M., Tavabi, N., and Kiapour, A. (2023). Do not mask randomly: Effective domain-adaptive pre-training by masking in-domain keywords. In Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023), pages 13–21.

Gomes, L. and Lopes, G. (2016). First steps towards coverage-based document alignment. In Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers, pages 697–702.

Gowda, T., Zhang, Z., Mattmann, C., and May, J. (2021). Many-to-English machine translation tools, data, and pretrained models. In Ji, H., Park, J. C., and Xia, R., editors, Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations, pages 306–316, Online. Association for Computational Linguistics.

Goyal, N., Gao, C., Chaudhary, V., Chen, P.-J., Wenzek, G., Ju, D., Krishnan, S., Ranzato, M., Guzmán, F., and Fan, A. (2022). The flores-101 evaluation benchmark for low-resource and multilingual machine translation. Transactions of the Association for Computational Linguistics, 10:522–538.

Grégoire, F. and Langlais, P. (2017). Bucc 2017 shared task: a first attempt toward a deep learning framework for identifying parallel sentences in comparable corpora. In Proceedings of the 10th Workshop on Building and Using Comparable Corpora, pages 46–50.

Guoa, M., Shenb, Q., Yanga, Y., Gea, H., Cera, D., Abregoa, G. H., Stevensa, K., Constanta, N., Sunga, Y.-H., Stropea, B., et al. (2018). Effective parallel corpus mining using bilingual sentence embeddings. WMT 2018, page 165.

Guzmán, F., Chen, P.-J., Ott, M., Pino, J., Lample, G., Koehn, P., Chaudhary, V., and Ranzato, M. (2019). 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), pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Ha, T.-L., Niehues, J., and Waibel, A. (2016). Toward multilingual neural machine translation with universal encoder and decoder. In Cettolo, M., Niehues, J., Stüker, S., Bentivogli, L., Cattoni, R., and Federico, M., editors, Proceedings of the 13th International Conference on Spoken Language Translation, Seattle, Washington D.C. International Workshop on Spoken Language Translation.

Haddow, B., Bawden, R., Miceli-Barone, A. V., Helcl, J., and Birch, A. (2022). Survey of low-resource machine translation. Computational Linguistics, 48(3):673–732.

Hangya, V. and Fraser, A. (2018). An unsupervised system for parallel corpus filtering. In Proceedings of the Third Conference on Machine Translation: Shared Task Papers, pages 882–887.

Hangya, V. and Fraser, A. (2019). Unsupervised parallel sentence extraction with parallel segment detection helps machine translation. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1224–1234.

Heffernan, K., Çelebi, O., and Schwenk, H. (2022). Bitext mining using distilled sentence representations for low-resource languages. In Findings of the Association for Computational Linguistics: EMNLP 2022, pages 2101–2112.

Herold, C., Rosendahl, J., Vanvinckenroye, J., and Ney, H. (2022). Detecting various types of noise for neural machine translation. In Findings of the Association for Computational Linguistics: ACL 2022, pages 2542–2551.

Hu, J., Johnson, M., Firat, O., Siddhant, A., and Neubig, G. (2021a). Explicit alignment objectives for multilingual bidirectional encoders. In Proceedings of the 2021

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3633–3643.

Hu, J., Johnson, M., Firat, O., Siddhant, A., and Neubig, G. (2021b). Explicit alignment objectives for multilingual bidirectional encoders. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3633–3643.

Ion, R., Ceauşu, A., and Irimia, E. (2011). An expectation maximization algorithm for textual unit alignment. In Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web, pages 128–135.

Isabelle, P. and Macklovitch, E. (1986). Transfer and mt modularity. In Coling 1986 Volume 1: The 11th International Conference on Computational Linguistics.

Isuranga, U., Sandaruwan, J., Athukorala, U., and Dias, G. (2020). Improved cross-lingual document similarity measurement.

Iyyer, M., Manjunatha, V., Boyd-Graber, J., and Daumé III, H. (2015). Deep unordered composition rivals syntactic methods for text classification. In Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers), pages 1681–1691.

Jain, M., Punia, R., and Hooda, I. (2020). Neural machine translation for tamil to english. Journal of Statistics and Management Systems, 23(7):1251–1264.

Jakubina, L. and Langlais, P. (2016). Bad luc@ wmt 2016: a bilingual document alignment platform based on lucene. In Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers, pages 703–709.

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

Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., and Levy, O. (2020). Spanbert: Improving pre-training by representing and predicting spans. Transactions of the Association for Computational Linguistics, 8:64–77.

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.

Khayrallah, H. and Koehn, P. (2018). On the impact of various types of noise on neural machine translation. In Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, pages 74–83.

Klein, G., Kim, Y., Deng, Y., Nguyen, V., Senellart, J., and Rush, A. M. (2018). Opennmt: Neural machine translation toolkit. In Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track), pages 177–184.

Kocmi, T., Zouhar, V., Federmann, C., and Post, M. (2024). Navigating the metrics maze: Reconciling score magnitudes and accuracies. In Ku, L.-W., Martins, A., and Srikumar, V., editors, Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1999–2014, Bangkok, Thailand. Association for Computational Linguistics.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In Proceedings of machine translation summit x: papers, pages 79–86.

Koehn, P., Chaudhary, V., El-Kishky, A., Goyal, N., Chen, P.-J., and Guzmán, F. (2020). Findings of the wmt 2020 shared task on parallel corpus filtering and alignment. In Proceedings of the Fifth Conference on Machine Translation, pages 726–742.

Koehn, P., Guzmán, F., Chaudhary, V., and Pino, J. (2019). Findings of the wmt 2019 shared task on parallel corpus filtering for low-resource conditions. In Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2), pages 54–72.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open source toolkit for statistical machine translation. In Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions, pages 177–180.

Koehn, P., Khayrallah, H., Heafield, K., and Forcada, M. L. (2018). Findings of the wmt 2018 shared task on parallel corpus filtering. In Proceedings of the Third Conference on Machine Translation: Shared Task Papers, pages 726–739.

Koehn, P. and Knowles, R. (2017). Six challenges for neural machine translation. In Proceedings of the First Workshop on Neural Machine Translation, pages 28–39.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, pages 127–133.

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

Kreutzer, J., Caswell, I., Wang, L., Wahab, A., van Esch, D., Ulzii-Orshikh, N., Tapo, A. A., Subramani, N., Sokolov, A., Sikasote, C., et al. (2022b). Quality at a glance: An audit of web-crawled multilingual datasets. Transactions of the Association for Computational Linguistics, 10:50–72.

Krupakar, H. and Milton, R. S. (2016). Improving the performance of neural machine translation involving morphologically rich languages. ArXiv, abs/1612.02482.

Kudugunta, S., Caswell, I., Zhang, B., Garcia, X., Xin, D., Kusupati, A., Stella, R., Bapna, A., and Firat, O. (2024). Madlad-400: A multilingual and document-level large audited dataset. Advances in Neural Information Processing Systems, 36.

Kumarasinghe, K., Dias, G., and Herath, I. (2021). Sinmorphy: A morphological analyzer for the sinhala language. In 2021 Moratuwa Engineering Research Conference (MERCon), pages 681–686. IEEE.

Kvapilíková, I., Artetxe, M., Labaka, G., Agirre, E., and Bojar, O. (2020). Unsupervised multilingual sentence embeddings for parallel corpus mining. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 255–262.

Lai, G., Xie, Q., Liu, H., Yang, Y., and Hovy, E. (2017). Race: Large-scale reading comprehension dataset from examinations. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 785–794.

Lakmal, D., Ranathunga, S., Peramuna, S., and Herath, I. (2020). Word embedding evaluation for sinhala. In Proceedings of The 12th Language Resources and Evaluation Conference, pages 1874–1881.

Lample, G. and Conneau, A. (2018). Phrase-based & neural unsupervised machine translation. In EMNLP.

Latief, A. D., Jarin, A., Yaniasih, Y., Afra, D. I. N., Nurfadhilah, E., Pebiana, S., Hidayati, N. N., and Fajri, R. (2024). Latest research in data augmentation for low resource language text translation: A review. In 2024 International Conference on Computer, Control, Informatics and its Applications (IC3INA), pages 185–190. IEEE.

Lee, E.-S. A., Thillainathan, S., Nayak, S., Ranathunga, S., Adelani, D. I., Su, R., and McCarthy, A. D. (2022). Pre-trained multilingual sequence-to-sequence models: A hope for low-resource language translation? arXiv preprint arXiv:2203.08850.

Leong, C., Wong, D. F., and Chao, L. S. (2018). Um-paligner: Neural network-based parallel sentence identification model. In 11th Workshop on Building and Using Comparable Corpora, page 53.

Leveling, J., Ganguly, D., Dandapat, S., and Jones, G. (2012). Approximate sentence retrieval for scalable and efficient example-based machine translation. In Kay, M. and Boitet, C., editors, Proceedings of COLING 2012, pages 1571–1586, Mumbai, India. The COLING 2012 Organizing Committee.

Levine, Y., Lenz, B., Lieber, O., Abend, O., Leyton-Brown, K., Tennenholtz, M., and Shoham, Y. (2020). Pmi-masking: Principled masking of correlated spans. In International Conference on Learning Representations.

Li, B. and Gaussier, E. (2013). Exploiting comparable corpora for lexicon extraction: Measuring and improving corpus quality. In Building and using comparable corpora, pages 131–149. Springer.

Liu, D., Ma, N., Yang, F., and Yang, X. (2019). A survey of low resource neural machine translation. In 2019 4th International Conference on Mechanical, Control and Computer Engineering (ICMCCE), pages 39–393. IEEE.

Liu, H., Hou, R., and Lepage, Y. (2024). High-quality data augmentation for low-resource nmt: Combining a translation memory, a gan generator, and filtering. arXiv preprint arXiv:2408.12079.

Liu, X., He, J., Liu, M., Yin, Z., Yin, L., and Zheng, W. (2023). A scenario-generic neural machine translation data augmentation method. electronics 2023, 12, 2320. doi.org/10.3390/electronics12102320, 4.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. Transactions of the Association for Computational Linguistics, 8:726–742.

Lopes, A., Farajian, M. A., Bawden, R., Zhang, M., and Martins, A. F. (2020). Document-level neural mt: A systematic comparison. In Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, pages 225–234.

Lu, H., Huang, H., Zhang, D., Wei, F., and Lam, W. (2024). Revamping multilingual agreement bidirectionally via switched back-translation for multilingual neural machine translation. In Findings of the Association for Computational Linguistics: EACL 2024, pages 264–275.

Luong, M.-T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1412–1421.

Ma, X. (2006). Champollion: A robust parallel text sentence aligner. In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), pages 489–492, Genoa, Italy. European Language Resources Association (ELRA).

Ma, X. and Liberman, M. (1999). Bits: A method for bilingual text search over the web. In Machine Translation Summit VII, pages 538–542.

Mager, M., Bhatnagar, R., Neubig, G., Vu, N. T., and Kann, K. (2023). Neural machine translation for the indigenous languages of the americas: An introduction. In Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 109–133.

Mahata, S., Das, D., and Bandyopadhyay, S. (2017). Bucc2017: A hybrid approach for identifying parallel sentences in comparable corpora. In Proceedings of the 10th Workshop on Building and Using Comparable Corpora, pages 56–59.

Maimaiti, M., Liu, Y., Luan, H., and Sun, M. (2022). Data augmentation for low-resource languages nmt guided by constrained sampling. International Journal of Intelligent Systems, 37(1):30–51.

Medveď, M., Jakubíček, M., and Kovář, V. (2016). English-french document alignment based on keywords and statistical translation. In Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers, pages 728–732.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. arXiv preprint arXiv:1310.4546.

Minh-Cong, N.-H., Van-Vinh, N., and Le-Minh, N. (2023a). A fast method to filter noisy parallel data wmt2023 shared task on parallel data curation. In Proceedings of the Eighth Conference on Machine Translation, pages 359–365.

Minh-Cong, N.-H., Vinh, N. V., and Le-Minh, N. (2023b). A fast method to filter noisy parallel data WMT2023 shared task on parallel data curation. In Koehn, P., Haddow, B., Kocmi, T., and Monz, C., editors, Proceedings of the Eighth Conference on Machine Translation, pages 359–365, Singapore. Association for Computational Linguistics.

Moon, H., Park, C., Koo, S., Lee, J., Lee, S., Seo, J., Eo, S., Jang, Y., Kim, H., Lee, H.-g., et al. (2023). Doubts on the reliability of parallel corpus filtering. Expert Systems with Applications, 233:120962.

Morin, E., Hazem, A., Boudin, F., and Loginova-Clouet, E. (2015). LINA: Identifying comparable documents from Wikipedia. In Proceedings of the Eighth Workshop on Building and Using Comparable Corpora, pages 88–91, Beijing, China. Association for Computational Linguistics.

Munteanu, D. S. and Marcu, D. (2002). Processing comparable corpora with bilingual suffix trees. In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002), pages 289–295.

Munteanu, D. S. and Marcu, D. (2005). Improving machine translation performance by exploiting non-parallel corpora. Computational Linguistics, 31(4):477–504.

Nag, S., Kale, M., Lakshminarasimhan, V., and Singhavi, S. (2020). Incorporating bilingual dictionaries for low resource semi-supervised neural machine translation. arXiv preprint arXiv:2004.02071.

Nagao, H. and Tsujii, J. (1986). The transfer phase of the mu machine translation system. In Coling 1986 Volume 1: The 11th International Conference on Computational Linguistics.

Nagao, M., Tsujii, J., Mitamura, K., Hirakawa, H., and Kume, M. (1980). A machine translation system from japanese into english-another perspective of mt systems. In COLING 1980 Volume 1: The 8th International Conference on Computational Linguistics.

Nagy, A., Lakatos, D. P., Barta, B., Nanys, P., and Ács, J. (2023). Data augmentation for machine translation via dependency subtree swapping. arXiv preprint arXiv:2307.07025.

Naranpanawa, R., Perera, R., Fonseka, T., and Thayasivam, U. (2020). Analyzing subword techniques to improve english to sinhala neural machine translation. International Journal of Asian Language Processing, 30(04):2050017.

Nastase, V. and Merlo, P. (2023). Grammatical information in bert sentence embeddings as two-dimensional arrays. In Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023), pages 22–39.

Nastase, V. and Merlo, P. (2024). Tracking linguistic information in transformer-based sentence embeddings through targeted sparsification. In Proceedings of the 9th Workshop on Representation Learning for NLP (RepL4NLP-2024), pages 203–214.

Ni, J., Abrego, G. H., Constant, N., Ma, J., Hall, K., Cer, D., and Yang, Y. (2022). Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. In Findings of the Association for Computational Linguistics: ACL 2022, pages 1864–1874.

Nissanka, L., Pushpananda, B., and Weerasinghe, A. (2020). Exploring neural machine translation for sinhala-tamil languages pair. In 2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer), pages 202–207. IEEE.

Novák, A., Tihanyi, L., and Prószéky, G. (2008). The metamorpho translation system. In Proceedings of the Third Workshop on Statistical Machine Translation, pages 111–114.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. Computational linguistics, 29(1):19–51.

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. In Proceedings of NAACL-HLT 2019: Demonstrations, pages 48–53.

Papavassiliou, V., Prokopidis, P., and Piperidis, S. (2016). The ilsp/arc submission to the wmt 2016 bilingual document alignment shared task. In Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers, pages 733–739.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pages 311–318.

Peng, W., Huang, C., Li, T., Chen, Y., and Liu, Q. (2020). Dictionary-based data augmentation for cross-domain neural machine translation. arXiv preprint arXiv:2004.02577.

Popović, M. (2015). chrf: character n-gram f-score for automatic mt evaluation. In Proceedings of the tenth workshop on statistical machine translation, pages 392–395.

Popović, M. (2017). chrf++: words helping character n-grams. In Proceedings of the second conference on machine translation, pages 612–618.

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

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

Pramodya, A. (2023). Exploring low-resource neural machine translation for sinhala-tamil language pair. In Proceedings of the 8th Student Research Workshop associated with the International Conference Recent Advances in Natural Language Processing, pages 87–97.

Pramodya, A., Pushpananda, R., and Weerasinghe, R. (2020). A comparison of transformer, recurrent neural networks and smt in tamil to sinhala mt. In 2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer), pages 155–160. IEEE.

Priyadarshani, H., Rajapaksha, M., Ranasinghe, M., Sarveswaran, K., and Dias, G. (2019). Statistical machine learning for transliteration: Transliterating names between sinhala, tamil and english. In 2019 International Conference on Asian Language Processing (IALP), pages 244–249. IEEE.

Prószéky, G. (2005). An approach to machine translation via the rule-to-rule hypothesis. In Proceedings of the 10th EAMT Conference: Practical applications of machine translation.

Pushpananda, R. (2019). Improving sinhala-tamil translation through deep learning techniques.

Rajitha, M., Piyarathna, L., Nayanajith, M., and Surangika, S. (2020). Sinhala and english document alignment using statistical machine translation. In 2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer), pages 29–34. IEEE.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2383–2392.

Ramesh, A., Parthasarathy, V. B., Haque, R., and Way, A. (2021a). Comparing statistical and neural machine translation performance on hindi-to-tamil and english-to-tamil. Digital, 1(2):86–102.

Ramesh, A., Uhana, H. U., Parthasarathy, V. B., Haque, R., and Way, A. (2021b). Augmenting training data for low-resource neural machine translation via bilingual word embeddings and bert language modelling. In 2021 International Joint Conference on Neural Networks (IJCNN), pages 1–8. IEEE.

Ranathunga, S. and de Silva, N. (2022). 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, pages 823–848.

Ranathunga, S., De Silva, N., Menan, V., Fernando, A., and Rathnayake, C. (2024a). Quality does matter: A detailed look at the quality and utility of web-mined parallel corpora. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pages 860–880.

Ranathunga, S., Lee, E.-S. A., Prifti Skenduli, M., Shekhar, R., Alam, M., and Kaur, R. (2023). Neural machine translation for low-resource languages: A survey. ACM Computing Surveys, 55(11):1–37.

Ranathunga, S., Lee, E.-S. A., Skenduli, M. P., Shekhar, R., Alam, M., and Kaur, R. (2021). Neural machine translation for low-resource languages: A survey. arXiv preprint arXiv:2106.15115.

Ranathunga, S., Ranasinghea, A., Shamala, J., Dandeniyaa, A., Galappaththia, R., and Samaraweeraa, M. (2024b). A multi-way parallel named entity annotated corpus for english, tamil and sinhala. arXiv preprint arXiv:2412.02056.

Rathnayake, H., Sumanapala, J., Rukshani, R., and Ranathunga, S. (2022). Adapter-based fine-tuning of pre-trained multilingual language models for code-mixed and code-switched text classification. Knowledge and Information Systems, 64(7):1937–1966.

Reimers, N., Gurevych, I., Reimers, N., Gurevych, I., Thakur, N., Reimers, N., Daxenberger, J., and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics.

Resnik, P. (1998). Parallel strands: A preliminary investigation into mining the web for bilingual text. In Conference of the Association for Machine Translation in the Americas, pages 72–82. Springer.

Resnik, P. (1999). Mining the web for bilingual text. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, pages 527–534.

Resnik, P. and Smith, N. A. (2003). The web as a parallel corpus. Computational Linguistics, 29(3):349–380.

Rossenbach, N., Rosendahl, J., Kim, Y., Graça, M., Gokrani, A., and Ney, H. (2018). The rwth aachen university filtering system for the wmt 2018 parallel corpus filtering task. In Proceedings of the Third Conference on Machine Translation: Shared Task Papers, pages 946–954.

Roy, A., Ray, P., Maheshwari, A., Sarkar, S., and Goyal, P. (2024). Enhancing low-resource nmt with a multilingual encoder and knowledge distillation: A case study. In Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024), pages 64–73.

Sachintha, D., Piyarathna, L., Rajitha, C., and Ranathunga, S. (2021). Exploiting parallel corpora to improve multilingual embedding based document and sentence alignment. arXiv preprint arXiv:2106.06766.

San, M. E., Usanavasin, S., Thu, Y. K., and Okumura, M. (2024). A study for enhancing low-resource thai-myanmar-english neural machine translation. ACM Transactions on Asian and Low-Resource Language Information Processing, 23(4):1–24.

Sánchez-Martínez, F., Perez-Ortiz, J. A., Galiano Jimenez, A., and Oliver, A. (2024). Findings of the WMT 2024 shared task translation into low-resource languages of Spain: Blending rule-based and neural systems. In Haddow, B., Kocmi, T., Koehn, P., and Monz, C., editors, Proceedings of the Ninth Conference on Machine Translation, pages 684–698, Miami, Florida, USA. Association for Computational Linguistics.

Sarikaya, R., Maskey, S., Zhang, R., Jan, E.-E., Wang, D., Ramabhadran, B., and Roukos, S. (2009). Iterative sentence-pair extraction from quasi-parallel corpora for machine translation. In Tenth Annual Conference of the International Speech Communication Association, pages 432–435.

Sarveswaran, K. and Dias, G. (2020). Thamizhiudp: A dependency parser for tamil. In Proceedings of the 17th International Conference on Natural Language Processing (ICON), pages 200–207.

Scao, T. L., Fan, A., Akiki, C., Pavlick, E., et al. (2022). Bloom: A 176b-parameter open-access multilingual language model. arXiv preprint arXiv:2211.05100.

Schwenk, H., Chaudhary, V., Sun, S., Gong, H., and Guzmán, F. (2021a). Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. In

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 1351–1361.

Schwenk, H., Wenzek, G., Edunov, S., Grave, É., Joulin, A., and Fan, A. (2021b). 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), pages 6490–6500.

Sen, S., Hasanuzzaman, M., Ekbal, A., Bhattacharyya, P., and Way, A. (2021). Neural machine translation of low-resource languages using smt phrase pair injection. Natural Language Engineering, 27:271–292.

Sennrich, R., Haddow, B., and Birch, A. (2016a). 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), pages 86–96.

Sennrich, R., Haddow, B., and Birch, A. (2016b). 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), pages 1715–1725.

Shi, L., Niu, C., Zhou, M., and Gao, J. (2006). A DOM tree alignment model for mining parallel data from the web. In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, pages 489–496.

Shi, S., Wu, X., Su, R., and Huang, H. (2022). Low-resource neural machine translation: Methods and trends. ACM Transactions on Asian and Low-Resource Language Information Processing, 21(5):1–22.

Shliazhko, A., Oguejiofor, A., Agafonova, A., et al. (2022). mgpt: Few-shot learners go multilingual. In Findings of the Association for Computational Linguistics: EMNLP 2022, pages 1492–1509.

Sloto, S., Thompson, B., Khayrallah, H., Domhan, T., Gowda, T., and Koehn, P. (2023). Findings of the wmt 2023 shared task on parallel data curation. In Proceedings of the Eighth Conference on Machine Translation, pages 95–102.

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.

Steingrímsson, S. (2023). A sentence alignment approach to document alignment and multi-faceted filtering for curating parallel sentence pairs from web-crawled data. In Proceedings of the Eighth Conference on Machine Translation, pages 366–374.

Steingrímsson, S., Loftsson, H., and Way, A. (2023). Filtering matters: Experiments in filtering training sets for machine translation. In Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa), pages 588–600.

Stocke, A. (2011). Srilm at sixteen: Update and outlook. In Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Waikoloa, Hawaii, Dec. 2011.

Stojanovski, D. (2021). Modeling contextual information in neural machine translation. PhD thesis, lmu.

Su, T., Peng, X., Thillainathan, S., Guzmán, D., Ranathunga, S., and Lee, E.-S. (2024). Unlocking parameter-efficient fine-tuning for low-resource language translation. In Findings of the Association for Computational Linguistics: NAACL 2024, pages 4217–4225.

Sun, S., Zhuang, S., Wang, S., and Zuccon, G. (2025). An investigation of prompt variations for zero-shot llm-based rankers. In European Conference on Information Retrieval, pages 185–201. Springer.

Sun, Y., He, J., Xia, M., and Neubig, G. (2021). Contrastive learning for unsupervised neural machine translation. In ACL.

Sun, Y., Wang, S., Li, Y., Feng, S., Chen, X., Zhang, H., Tian, X., Zhu, D., Tian, H., and Wu, H. (2019). Ernie: Enhanced representation through knowledge integration. arXiv preprint arXiv:1904.09223.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. Advances in Neural Information Processing Systems, 27:3104–3112.

Takase, S. and Kiyono, S. (2023). Lessons on parameter sharing across layers in transformers. In Proceedings of The Fourth Workshop on Simple and Efficient Natural Language Processing (SustaiNLP), pages 78–90.

Tan, X., Ren, Y., He, D., Qin, T., Zhao, Z., and Liu, T.-Y. (2019). Multilingual neural machine translation with knowledge distillation. arXiv e-prints, pages arXiv–1902.

Tang, Y., Tran, C., Li, X., Chen, P.-J., Goyal, N., Chaudhary, V., Gu, J., and Fan, A. (2021). Multilingual translation from denoising pre-training. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 3450–3466.

Tars, M., Tattar, A., and Fishel, M. (2022). Cross-lingual transfer from large multilingual translation models to unseen under-resourced languages. Baltic Journal of Modern Computing, 10(3):435–446.

Tennage, P., Herath, A., Thilakarathne, M., Sandaruwan, P., and Ranathunga, S. (2018a). Transliteration and byte pair encoding to improve tamil to sinhala neural machine translation. In 2018 Moratuwa Engineering Research Conference (MERCon), pages 390–395. IEEE.

Tennage, P., Sandaruwan, P., Thilakarathne, M., Herath, A., and Ranathunga, S. (2018b). Handling rare word problem using synthetic training data for sinhala and tamil neural machine translation. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).

Tennage, P., Sandaruwan, P., Thilakarathne, M., Herath, A., Ranathunga, S., Jayasena, S., and Dias, G. (2017). Neural machine translation for sinhala and tamil languages. In 2017 International Conference on Asian Language Processing (IALP), pages 189–192. IEEE.

Thillainathan, S., Ranathunga, S., and Jayasena, S. (2021). Fine-tuning self-supervised multilingual sequence-to-sequence models for extremely low-resource NMT. In 2021 Moratuwa Engineering Research Conference (MERCon), pages 432–437. IEEE.

Thompson, B. and Koehn, P. (2019). Vecalign: Improved sentence alignment in linear time and space. 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), pages 1342–1348.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), pages 2214–2218.

Udawatta, P., Udayangana, I., Gamage, C., Shekhar, R., and Ranathunga, S. (2024). Use of prompt-based learning for code-mixed and code-switched text classification. World Wide Web, 27(5):63.

Uszkoreit, J., Ponte, J., Popat, A., and Dubiner, M. (2010). Large scale parallel document mining for machine translation. In Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 1101–1109.

Varga, D., Halácsy, P., Kornai, A., Nagy, V., Németh, L., and Trón, V. (2007). Parallel corpora for medium density languages. Amsterdam Studies In The Theory And History Of Linguistic Science Series 4, 292:247.

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, 30:5998–6008.

Velayuthan, M., Jayakody, D., De Silva, N., Fernando, A., and Ranathunga, S. (2024). Back to the stats: Rescuing low resource neural machine translation with statistical methods. In Proceedings of the Ninth Conference on Machine Translation, pages 901–907.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2018a). Glue: A multi-task benchmark and analysis platform for natural language understanding. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 353–355.

Wang, J., Lu, Y., Weber, M., Ryabinin, M., Adelani, D., Chen, Y., Tang, R., and Stenetorp, P. (2025). Multilingual language model pretraining using machine-translated data. arXiv preprint arXiv:2502.13252.

Wang, X., Pham, H., Dai, Z., and Neubig, G. (2018b). Switchout: an efficient data augmentation algorithm for neural machine translation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 856–861.

Wang, Z., Wang, P., Liu, K., Wang, P., Fu, Y., Lu, C.-T., Aggarwal, C. C., Pei, J., and Zhou, Y. (2024). A comprehensive survey on data augmentation. arXiv preprint arXiv:2405.09591.

Weaver, W. (1955). Translation. In Machine Trans Languages, volume 14, pages 15–23.

Weller-Di Marco, M. and Fraser, A. (2022). Findings of the WMT 2022 shared tasks in unsupervised MT and very low resource supervised MT. In Koehn, P., Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussà, M. R., Federmann, C., Fishel, M., Fraser, A., Freitag, M., Graham, Y., Grundkiewicz, R., Guzman, P., Haddow, B., Huck, M., Jimeno Yepes, A., Kocmi, T., Martins, A., Morishita, M., Monz, C., Nagata, M., Nakazawa, T., Negri, M., Névéol, A., Neves, M., Popel, M., Turchi, M., and Zampieri, M., editors, Proceedings of the Seventh Conference on Machine Translation (WMT), pages 801–805, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Werbos, P. J. (1990). Backpropagation through time: what it does and how to do it. Proceedings of the IEEE, 78(10):1550–1560.

Wettig, A., Gao, T., Zhong, Z., and Chen, D. (2023). Should you mask 15% in masked language modeling? In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 2977–2992.

Winiwarter, W. (2007). Jetcat–japanese-english translation using corpus-based acquisition of transfer rules. JOURNAL OF COMPUTERS, 2(9):27.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2021a). mT5: A massively multilingual pre-trained text-to-text transformer. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 483–498, Online. Association for Computational Linguistics.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2021b). mt5: A massively multilingual pre-trained text-to-text transformer. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), page 483–498.

Yang, Y., Cer, D., Ahmad, A., Guo, M., Law, J., Constant, N., Abrego, G. H., Yuan, S., Tar, C., Sung, Y.-H., et al. (2020). Multilingual universal sentence encoder for semantic retrieval. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 87–94.

Yang, Z., Li, Y., Liu, L., Li, R., and Li, M. (2023). Grammar-aware representation learning for unsupervised machine translation. In EMNLP.

Yazar, B. K., Şahın, D. Ö., and Kiliç, E. (2023). Low-resource neural machine translation: A systematic literature review. IEEE Access, 11:131775–131813.

Zafarian, A., Sadeghi, A. P. A., Azadi, F., Ghiasifard, S., Panahloo, Z. A., Bakhshaei, S., and Ziabary, S. M. M. (2015). Aut document alignment framework for bucc workshop shared task. In Proceedings of the Eighth Workshop on Building and Using Comparable Corpora, pages 79–87.

Zhang, B., Nagesh, A., and Knight, K. (2020). Parallel corpus filtering via pre-trained language models. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8545–8554.

Zhang, J., Luan, H., Sun, M., Zhai, F., Xu, J., Zhang, M., and Liu, Y. (2018). Improving the transformer translation model with document-level context. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 533–542.

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

Zhou, Y., Guo, C., Wang, X., Chang, Y., and Wu, Y. (2024). A survey on data augmentation in large model era. arXiv e-prints, pages arXiv–2401.

Ziemski, M., Junczys-Dowmunt, M., and Pouliquen, B. (2016). The united nations parallel corpus v1. 0. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 3530–3534.

Zoph, B., Yuret, D., May, J., and Knight, K. (2016). Transfer learning for low-resource neural machine translation. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1568–1575.

Zweigenbaum, P., Sharoff, S., and Rapp, R. (2018). Overview of the third bucc shared task: Spotting parallel sentences in comparable corpora. In Proceedings of 11th Workshop on Building and Using Comparable Corpora, pages 39–42.