A SURVEY OF NEURAL MACHINE TRANSLATION ON LOW-RESOURCE LANGUAGE PAIRS

A Research Paper submitted to the Department of Computer Science
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By

Anusha Choudhary

May 12, 2023

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

ADVISOR
Yangfeng Ji, Department of Computer Science

In a world seemingly replete with an excess of data and tools to automate the generation of new data, it is easy to forget there nevertheless exists an imbalance in the availability of data across different languages (Joshi et al., 2020), giving rise to Low-Resource Languages (LRL). Due to the inherent nature of data-driven Neural Network-based models, this leads to poor performance quality of such models on low-resource languages. Machine Translation (MT) is one such field afflicted with an imbalance of performance quality due to an imbalance of training data across languages (Koehn & Knowles, 2017).

Although the problem of low-resource MT had been identified as early as 2017, (Koehn & Knowles, 2017), in 2023, low-resource machine translation is a more active field of research than ever before due to the meteoric rise in computational power that supports Neural Machine Translation (NMT) models as well as increasing regional support for languages. With increased volume of research, there is an increasing need to summarize the methodologies being used to improve low-resource machine translation. Thus, to benefit the NMT community, we provide a snapshot of the most recent methodologies being applied to achieve higher translation quality on multilingual NMT models and highlight the emerging trends within those methods. We focus primarily on improvements to multilingual NMT models as multilingual models have been shown to perform better in low-resource conditions (Lakew et al., 2018; Ranathunga et al., 2021).

In Section 1, we recap the methods that have already been summarized by previous surveys to delineate previously established improvements from the new research, and thus provide a starting threshold for the new methods summarized in this paper. Sections 2 and 3 discuss new improvements categorized by the core sub-problem they attempt to solve. For example, Section 2 covers methods focused on the problem of improving cross-lingual transfer. Similarly, Section 3 focuses improving bitext mining and data selection methods while Section 4 summarizes a method that improves tokenization. The paper concludes by discussing the key takeaways gained from surveying these novel methods.

1 Related Work

Past surveys specifically focused on low-resource NMT include Ranathunga et al. (2021) and Haque et al. (2021). Other surveys focused more generally on all low-resource NLP tasks include Hedderich et al. (2021). The improvements to low-resource MT established in previous surveys include the superiority of multilingual NMT over bilingual NMT models, the success of back-translation and other Data Augmentation methods, Transfer Learning and Pivot Learning for linguistically similar languages, Multi-Task Learning, Subword-level encodings, and Denoising (Ranathunga et al., 2021; Haque et al., 2021). With the goal of providing a more recent snapshot of the literature, this paper summarizes only those methods that are sufficiently distinct from previously identified themes and have appeared after the publishing of the two past surveys of low-resource MT.

Haque et al. (2021) identifies morphologically-rich and complex languages (MRCL), mistranslation of rare words, distant language pairs, and universal NMT to be open challenges within low-resource MT. Each of these challenges is addressed by one or more methods discussed in the following sections of this paper.

Additionally, Ranathunga et al. (2021) brings to attention ethical and equity-related concerns within NMT, specifically recommending that “more focus can be

given to those that present new LRL datasets rather than the novelty of the employed technique, when accepting papers to conferences and evaluating value of this type of research." While we recognize that the creation of new LRL datasets is the most straightforward and productive approach towards better-quality low-resource NMT, we nevertheless focus on novel methods in this paper. Novel methods aimed at low-resource languages maximize the number of languages they benefit while minimizing the amount of development time required for the method itself. On the other hand, a new LRL dataset affects only one language or a handful of similar languages and takes large amounts of time and manpower to create. Novel methods can benefit those LRL which are not in the spotlight of public and/or regional attention, or for which adequate manpower is not available. Therefore, there still exists a need to periodically survey the new state-of-the-art structural improvements being made to multilingual NMT models.

2 Improving Cross-Lingual Transfer

Bringing languages closer together manifests itself in NMT as bringing the representational embeddings of all languages closer together within multilingual NMT models. Intuitively, just as simultaneously recognizing the similarities and differences between different language speakers is important to bridge the cultural gap, maximizing positive cross-lingual transfer and minimizing negative interference between different language families is the goal within multilingual-NMT to improve its performance on low-resource pairs as well as zero-shot translation. In the next section, we explore the recent novel techniques being implemented to improve cross-lingual transfer and to identify universal language-agnostic parameters across languages in multilingual NMT models.

2.1 Language-Family Adapters

Chronopoulou et al. (2023, p. 1) introduce a novel addition to the mBART-50 (Tang et al. 2020) model to facilitate cross-lingual transfer by considering an adapter layer that is neither language-agnostic nor specific to each language-pair, but instead based on language families according to the World Atlas of Language Structures (WALS). By freezing the pretrained encoder-decoder Transformer model and adding a language-family adapter layer 1) After the input embedding layer and before the output embedding layer and 2) after the Feed-Forward layer in both the encoder and the decoder and fine-tuning only those adapter layers, Chronopoulou et al. (2023) gained better performance in BLEU scores over previous Language-Agnostic adapters as well as Language-Pair adapters across all experimental conditions. The language-family adapters also achieved better performance over the aforementioned adapters on both seen and unseen languages in the mBART-50 model. Since the adapter layers are added to a frozen pre-trained model and only the additional parameters are required to be fine-tuned, this improvement is computationally efficient. Chronoupoulou et al. (2023) does, however, recognize that the model could encode bias as the underlying model has been pre-trained on monolingual data from Common Crawl and no toxicity evaluation of the generated text exists.

2.2 Regularized Sparsely-Gated Mixture of Experts

In the same vein as Chronoupoulou et al. (2023), Costa-jussa et al. (2022) focus on conditionally learning different subsets of

model parameters in their NLLB model to achieve a tradeoff between cross-lingual transfer and interference between different language families. They achieve this by adding a Sparsely-Gated Mixture of Experts (Almahairi et al., 2016; Bengio et al., 2013) to their transformer model. However, Costa-jussa et al. (2023) observe that the addition of Sparsely-Gated MoE results in overfitting in the low-resource scenario and thus implement various regularization methods, of which the Expert Output Masking (EOM) regularizer achieved the best performance.

2.3 Pseudo-Pivot Learning using Discretized Codes

To facilitate knowledge-sharing between languages as well as increasing intermediate model interpretability, Liu & Nieheus (2022) propose a pseudo-pivoting step in multilingual NMT models to map the continuous output of the encoder layer to discretized codes using the sliced Vector Quantized Variational Autoencoder (VQ-VAE) (Van Den Oord & Vinyals, 2017). The underlying pretrained model used by Liu & Nieheus (2022) is the small variant of the M2M-124 (Wenzek et al., 2021) model. Liu & Nieheus (2022) add a soft discretization layer after the encoder layer in the M2M-124 model to "strike a balance between discretization and performance" (p. 191), notably using the same number of slices as attention heads in the model. Upon experimentation, the BLEU score achieved by this discretization method falls behind the benchmark methods of Language-Independent Objective (Pham et al., 2019; Arivazhagan et al., 2019a) and Adversarial Language Classifier (Arivazhagan et al., 2019) in the English-Bridge Language scenario, where the bridge languages include Javanese, Malay, and Tagalog, but is higher than the benchmark methods in the Indonesian-Bridge Language scenario. This renders the proposed discretization layer less useful for learning language-agnostic features and more useful for intermediate model interpretability.

2.4 Layer-Freezing

Transfer Learning is now a widely-accepted method of improving low-resource translation quality, especially for similar language pairs and parent-child pairs that share the same target language (Ranathunga et al., 2021; Haque et al., 2021). Philippy et al. (2023) show correlations between language distance and cross-lingual transfer performance in the different layers of the mBERT (Devlin et al., 2019) model. Since certain layers reflect strong negative Pearson correlation coefficients between language distance and the impact on the representation space and others show positive correlations, Philippy et al. (2023) propose selectively freezing layers that show negative correlation between language distance and the impact on the representation space to improve cross-lingual transfer performance on distant languages in a multilingual language model such as mBERT. The three pilot experiments in layer-freezing provided by Philippy et al. (2023) offer promising improvements in cross-lingual transfer performance (CLTP) using their proposed layer-freezing method. This signifies potential for a similar layer-freezing approach to improve low-resource translation performance in a multilingual NMT model as well; however, extensive further experimentation must be conducted specifically for the translation task, as opposed to simply cross-lingual transfer, to achieve a decisive conclusion on the usefulness of this method. To that end, Chowdhury et al. (2022) offers a promising result in favor of layer-freezing in multilingual NMT, discussed below.

Chowdhury et al. (2022) used transfer learning to train a very-low-resource language pair, English-Lambani. The child English-Lambani Transformer model with six encoder and decoder layers and eight attention heads was trained using three parent models: English-Kannada, English-Marathi, and English-Gujarati. The experiment measured BLEU scores for all three parent-child transfer pairs while 1) Freezing the first five encoder layers and 2) Freezing the first five decoder layers. The results of the experiment showed increased BLEU scores for all three parent-child pairs when the encoder layers were frozen and decreases in BLEU score when decoder layers were frozen. This can be explained intuitively using the idea from Philippy et al. (2023) of the distance between parent and child languages having opposite impacts on the representation space in the encoder and decoder layers. Chowdhury et al. (2022) provides positive evidence in favor of exploring selective layer-freezing in any MT scenario involving linguistically similar languages, such as transfer learning, pivot learning, and multilingual NMT.

3 Bitext Mining and Data Selection

In a low-resource scenario where parallel data is limited, it becomes all the more crucial to ensure the sentence pairs being used for training are good-quality, i.e., the training pairs are both accurate and diverse. This section reviews an improved sentence-encoder for bitext mining and a new framework for sentence-selection.

3.1 Improving Multilingual Sentence Representations

Moving away from the idea of improving cross-lingual transfer by modifying the architecture of multilingual NMT models, methods like the one introduced by Heffernan et al. (2022) focus on improving sentence representations within the multilingual space for better bitext mining. Although this method is focused on an earlier stage of the model architecture, the underlying intuition of learning representations grouped by language families as seen in Chronopoulou et al. (2023) is shared by Heffernan et al. (2022). The core idea behind the LASER2 and LASER3 sentence encoders introduced by Heffernan et al. (2022) is to jointly train a multilingual teacher-student distillation model alongside a monolingual Masked Language Model (MLM), effectively combining supervised and self-supervised training and enabling the teacher-student model to also learn context from monolingual student data. Through this novel improvement to the LASER (Artetxe & Schwenk, 2019) encoder, Heffernan et al. (2022) achieved better xsim (Artetxe & Schwenk, 2018) error rates for 12 languages as compared to the original LASER encoder and also trained encoders for 50 previously unseen African languages. The LASER3 sentence encoders introduced by Heffernan et al. (2022) were used for the bitext mining of low-resource languages in the NLLB model introduced by Costa-jussa et al. (2022), which indicates that the LASER3 encoder can be considered a novel state-of-the-art sentence-level encoder for the bitext mining of low-resource languages.

3.2 COMET-QE as a metric for Sentence Selection with Active Learning

Pool-based Active Learning in the context of MT (Haffari et al., 2009; Bloodgood & Callison-Burch, 2010) is a method for selecting candidate sentences for translation from a pre-existing pool of monolingual data such that when the translated sentence pair is used as training data for a model, it maximizes uncertainty, thus making the

training data diverse and challenging for the model to learn from. To ensure good quality of training data via Active Learning, it is therefore worthwhile to ask which uncertainty metric for sentence selection yields the best model performance for low-resource languages. Chimoto & Bassett (2022) show that using COMET-QE as the uncertainty metric for sentence selection yields higher BLEU scores on the resulting Transformer-based NMT model for three (artificially-restricted as well as true) very-low-resource languages: Swahili-English, Spanish-English, and Kinyarwanda-English compared to previous state-of-the-art uncertainty metrics such as Round Trip Translation Likelihood (RTTL) and stratified-RTTL, as well as random sentence selection. This initial positive result of COMET-QE with pool-based AL invites potential for further experimentation on using the metric for selecting sentences in other languages, both low and high-resource, and evaluating the performance of the resulting NMT models.

4 Improving Tokenization

The inherent long-tailed nature of morphologically-rich low-resource languages is a challenge as it makes NMT prone to mistranslating rare words (Haque et al., 2021). Naturally, the solution to the long-tail problem is to represent tokens in a way that maximizes the frequency of each token. To this end, Signoroni et al. (2021) provide a novel tokenization strategy called High-Frequency Tokenizer (HFT) which achieves improvement in NMT performance for four low-resource languages over state-of-the-art subword segmentation methods such as Byte-Pair Encoding (BPE) and Unigram, as well as improved scores on other metrics for evaluating tokenizers such as Frequency at 95% Class Rank and Mean Sequence Length. HFT uses the advantage of pretokenization and its learning algorithm works by first considering all characters as possible subwords and then iteratively considering only the top K (where K is 5% of the vocabulary) candidate subwords with the highest frequency and removing all multicharacter subwords with frequency lower than the last added candidate.

Signoroni et al. (2021) also introduce a new metric for evaluating tokenizers, Frequency Rank Weighted Average, which is intended to give higher weights to the low frequency tokens in a vocabulary while also considering overall vocabulary length (p. 3).

Notably, Signoroni et al. (2021) used default Transformers when testing the performance of the NMT task using HFT. Therefore, there is potential to further test the performance of HFT when used in conjunction with optimizations known to increase Transformer performance, such as data filtering, back-translation and multilingual NMT.

Conclusion

This paper attempted to summarize seven new methods aimed at improving the quality of low-resource Machine Translation primarily in multilingual NMT models. In considering the results of these seven methods, we observe two qualitative takeaways:

1) The introduction of Language-Family Adapters (Chronopoulou et al. 2023) and Regularized Sparsely-Gated MoE (Costa-jussa et al., 202) indicates that architectural changes in Transformer models are still yielding performance improvements for the NMT task, so the core structure of the state-of-the-art Transformer model for NMT cannot be considered to

have converged to a uniform architecture yet.

2) There is a general qualitative trend in the literature of moving away from learning universal language-agnostic representations and instead focusing on language families, as seen in the inconsistent performance of sliced VQ-VAE (Liu & Nieheus, 2022) and strong performance of Language-Family Adapters (Chronopoulou et al. 2023) and Regularized Sparsely-Gated MoE (Costa-jussa et al., 2022).

Overall, we observe through the literature review performed in this survey that while no overarching high-level themes emerge in the improvements being made to low-resource NMT, the NMT community is currently focused on making finer low-level improvements to the high-level optimization techniques surveyed two years ago by Ranathunga et al. (2021) and Haque et al. (2021).

REFERENCES

Almahairi, A. Ballas, N., Cooijmans, T., Zheng, Y., Larochelle, H. & Courville, A. (2016) *Dynamic capacity networks. Proceedings of the 33rd International Conference on International Conference on Machine Learning, ICML'16* (pp. 2091 – 2100).

Arivazhagan, N., Bapna, A., Firat, O., Aharoni, R., Johnson, M., & Macherey, W. (2019). The missing ingredient in zero-shot neural machine translation. (Unpublished). *arXiv:1903.07091*.

Artetxe, M., & Schwenk, H. (2018). Margin-based parallel corpus mining with multilingual sentence embeddings. *Association for Computation and Linguistics.* https://doi.org/10.48550/arXiv.1811.01136

Artetxe, M., & 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.

Bengio, Y., Léonard, N., & Courville, A. (2013). Estimating or propagating gradients through stochastic neurons for conditional computation. (Unpublished). *arXiv:1308.3432*. http://arxiv.org/abs/1308.3432

Bloodgood, M., & Callison-Burch, C. (2014). Bucking the trend: Large-scale cost-focused active learning for statistical machine translation. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 854–864). https://doi.org/10.48550/arXiv.1410.5877

Chimoto, E. A., & Bassett, B. A. (2022). COMET-QE and active learning for low-resource machine translation. (Unpublished). *arXiv:2210.15696*. https://doi.org/10.48550/arXiv.2210.15696

Chowdhury, A., Deepak, K. T., & Prasanna, S. M. (2022, October). Machine translation for a very low-resource language-layer freezing approach on transfer learning. *Proceedings of the Fifth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2022)* (pp. 48-55). https://aclanthology.org/2022.loresmt-1.7.pdf

Chronopoulou, A., Stojanovski, D., & Fraser, A. (2022). Language-family adapters for multilingual neural machine translation. (Unpublished). *arXiv:2209.15236*. https://doi.org/10.48550/arXiv.2209.15236

Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., ... & Wang, J. (2022). No language left behind: Scaling human-centered machine translation (Unpublished). *arXiv:2207.04672*. https://doi.org/10.48550/arXiv.2207.04672

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep

bidirectional transformers for language understanding. (Unpublished). *arXiv:1810.04805*. https://doi.org/10.48550/arXiv.1810.04805

Haffari, G., Roy, M., & Sarkar, A. (2009, June). Active learning for statistical phrase-based machine translation. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 415-423).

Haque, R., Liu, C.H., Way, A. (2021, December). Recent advances of low-resource machine translation. *Machine Translation*, 35*(2*). https://doi.org/10.1007/s10590-021-09281-1

Hedderich, M. A., Lange, L., Adel, H., Strötgen, J., & Klakow, D. (2020). A survey on recent approaches for natural language processing in low-resource scenarios. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 2545-2568). https://aclanthology.org/2021.naacl-main.201.pdf

Heffernan, K., Çelebi, O., & Schwenk, H. (2022). Bitext mining using distilled sentence representations for low-resource languages. (Unpublished). *arXiv:2205.12654*. https://doi.org/10.48550/arXiv.2205.12654

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. *Proceedings of the 58th annual meeting of the association of computational linguistics*. https://doi.org/10.18653/v1/2020.acl-main.560

Koehn, P., & Knowles, R. (2017). Six challenges for neural machine translation. *Proceedings for the first workshop on neural machine translation*. https://doi.org/10.48550/arXiv.1706.03872

Lakew, S. M., Cettolo, M., & Federico, M. (2018). A comparison of transformer and recurrent neural networks on multilingual neural machine translation. *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 641–652).

Liu, D. & Niehues, J. (2022). Learning an artificial language for knowledge-sharing in multilingual translation. *Proceedings of the Seventh Conference on Machine Translation (WMT)* (pp. 188–202). https://www.statmt.org/wmt22/pdf/2022.wmt-1.12.pdf

Pham, N. Q., Niehues, J., Ha, T. L., & Waibel, A. (2019). Improving zero-shot translation with language-independent constraints. (Unpublished). *arXiv:1906.08584*.

Philippy, F., Guo, S., & Haddadan, S. (2023). Identifying the correlation between language distance and cross-lingual transfer in a multilingual representation space. (Unpublished). *arXiv:2305.02151*. https://doi.org/10.48550/arXiv.2305.02151

Ranathunga, S., Lee, E. S. A., Prifti Skenduli, M., Shekhar, R., Alam, M., & Kaur, R. (2021).

Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, *55*(11), pp. 1-37. https://doi.org/10.48550/arXiv.2106.15115

Signoroni, E., & Rychlý, P. (2022). HFT: High frequency tokens for low-resource NMT. *Proceedings of the 29th International Conference on Computational Linguistics* (pp. 56–63). https://aclanthology.org/2022.loresmt-1.8.pdf

Tang, Y., Tran, C., Li, X., Chen, P. J., Goyal, N., Chaudhary, V., ... & Fan, A. (2020). Multilingual translation with extensible multilingual pretraining and finetuning. (Unpublished). *arXiv:2008.00401*. https://doi.org/10.48550/arXiv.2008.00401

Van Den Oord, A., & Vinyals, O. (2017). Neural discrete representation learning. *Advances in neural information processing systems*, *30*.

Wenzek, G., Chaudhary, V., Fan, A., Gomez, S., Goyal, N., Jain, S., ... & Guzmán, F. (2021, November). Findings of the WMT 2021 shared task on large-scale multilingual machine translation. In *Proceedings of the Sixth Conference on Machine Translation* (pp. 89-99).