A New Approach to Parameter-Sharing in Multilingual Neural Machine Translation

Benyamin Ahmadnia  
ahmadnia@ucdavis.edu  
Department of Linguistics, University of California, Davis, CA 95616, USA  
School of Professional Advancement, Tulane University, New Orleans, LA 70118, USA

Bonnie J. Dorr  
bdorr@ihmc.us  
Institute for Human and Machine Cognition (IHMC), Ocala, FL 34774, USA

Abstract

In multilingual Machine Translation (MT), the strategy of parameter-sharing not only determines how to make optimal use of the parameter space but also directly influences ultimate translation quality. In this paper, we propose an approach to parameter-sharing in multilingual MT. The core idea of this approach is employing the expert language hierarchies as a fundamental for multilingual architecture: the closer two languages are, the more parameters they share. The proposed approach finds support in observations on the role of language affinity in recent MT literature. We consider the simplest case with two sources and one target language and show the potential of this approach. Our results demonstrate that the hierarchical architecture outperforms bilingual models and multilingual model with full parameter sharing, and show how translation quality of the hierarchical model depends on linguistic distance between source languages.

1 Introduction

Low-resource languages are the languages with fewer technologies and datasets relative to some measure of their international importance (Ahmadnia and Dorr, 2019). In simple words, the languages for which bilingual corpus is sparse, requiring recourse to techniques that are complementary to data-driven MT approaches. The biggest issue with low-resource languages is the difficulty of obtaining sufficient resources. Natural Language Processing (NLP) methods that have been created for analysis of low-resource languages are likely to encounter similar issues to those faced by documentary and descriptive linguists whose primary endeavor is the study of minority languages. Parallel data in other languages can substantially benefit low-resource MT. For example, creating a multilingual model that in which multiple parallel corpora can be combined to train a single model where languages can share some parameters to help each other learn instead of training separate models for each translation direction, is a possible way to overcome the data scarcity bottleneck.

Given that languages have a lot in common, properly organized parameter-sharing strategy could compensate for the lack of training examples in low-resource language pairs. Exploiting language relatedness can substantially enhance translation quality (Tan et al., 2019). Since a lack of systematic approach which would account for the degree of relatedness between languages in a multilingual model throughout the current approaches such as full parameter-sharing (Johnson et al., 2017), or shared encoder side with language-specific decoders (Dong et al., 2015), and with respect to the fact that languages are able to share various characteristics...
on different levels like alphabet, morphology, semantics, we propose an approach depending on the languages’ closeness which makes sense to look for shared parts on different levels. The core idea is to organize both encoder and decoder in a hierarchical fashion, reflecting the degree of relatedness between languages, such that the most related languages share the largest number of parameters.

This paper is organized as follows; Section 2 investigates the previous related work. Section 3 describes the methodology. The experimental results and analysis are covered by Section 4. Conclusions and future work are provided in Section 5.

2 Related Work

The simplest form of parameter sharing was introduced by Johnson et al. (2017) where all parameters are shared and identifier tokens are employed to distinguish between languages. This model requires more resources to capture relationships between languages.

Dong et al. (2015) investigated the translation problem as a multi-task problem with shared encoder and separate decoders for each target language. As a disadvantage of this work, the potential knowledge sharing between target languages is not considered.

Sachan and Neubig (2018) proposed a one-to-many model, where parameters are partially shared between the multiple decoders. Although this work is similar to our approach, instead of having shared parts between individual decoders, we propose creating a hierarchy of decoders.

In the aforementioned works, languages on the encoder or decoder of multilingual models are treated the same regardless of their relatedness degree but in our model, the number of shared parameters in our approach, heavily depends on the degree of kinship between languages.

3 Method Description

We aim at proposing an optimal parameter-sharing strategy between languages in a multilingual Neural MT (NMT). The core idea is to organize languages on the encoder and decoder according to their linguistic similarity. Creating a hierarchical model on each side such that the hierarchies correspond to how languages are connected in phylogenetic trees1.

As an example, assume that sufficient amount of bilingual texts for the following language pairs are available: Russian-Tatar (ru-tt), Russian-Kazakh (ru-kk), Russian-English (ru-en), English-Turkish (en-tr), English-Azerbaijani (en-az), German-Turkish (de-tr), Azerbaijani-Turkish (az-tr). Figure 1 shows a detailed high-level and detailed view of the architecture for a multilingual hierarchical model that are trained using the mentioned corpora.

In Figure 1, there are four source and five target languages connected through the chain of hierarchically organized encoders and decoders. In the encoder side, all source languages have their own respective encoders $e_{ru}$, $e_{en}$, $e_{de}$, and $e_{az}$. The parameters of these encoders are not shared with any other languages since they are intended to learn language-specific features. In detail, the outputs of some encoders are stacked together and passed to the shared encoders. These encoders are: $e_{en-de}$ and $e_{ru-en-de}$ that are shared among two or more languages so that they can capture knowledge common to these languages. Finally, there is the last encoder ($e_{ru-en-de-es}$) that is shared by all source languages which combines the outputs of all remaining encoders together.

This architecture has been designed to enable knowledge sharing between different languages on various levels. It can be seen that English and German are connected first2. Then, Russian is added in3. Next, all these languages are joined Azerbaijanian4. Logic on the decoder

---

1The closer two languages are, the more parameters they share.
2They come from the Germanic branch of the Indo-European language family.
3It belongs to a different branch of the same language family.
4It comes from an entirely different Turkic language family.
Figure 1: A high-level view of a multilingual hierarchical model. $e$ is a shared encoder, and $d$ is a decoder.

side is analogous—most dissimilar languages split first. We propose having the same number of parameters along any path from source to target. For example, if we compare ru-tt and az-en translation paths in Figure 1, although the latter has considerably less shared parameters, their total number should be the same for both paths. That is why some encoder blocks are longer than others, pointing that, for example, the number of parameters in $e_{az}$ should be the same as there are cumulatively in $e_{ru}$ and $e_{ru-en-de}$.

In NMT, model parameters are learned by maximizing the conditional probability $P(t|s)$ of reconstructing target sentence ($t$) given source sentence ($s$). In a multilingual setting, there are multiple source and target sentences $(s_1, s_2, s_3, ...), (t_1, t_2, t_3, ...)$. Then, for each available bilingual dataset $(i, j)$, the aim is at maximizing $P(t_i|s_j)$. Assuming some parts of the model are shared, when maximizing the conditional probabilities, only those parameters are updated that depend on the path from one language to another. Considering a simple example in Figure 2. There are two possible translation paths, and individual encoder parameters $\theta_{e_1}, \theta_{e_2}$ will only be updated when the parameters for the corresponding translation path are updated.

Figure 2: Simple multi-source hierarchical architecture. $e_1, e_2$ are individual encoders, $e$ is a shared encoder, and $d$ is a decoder.

As for the training procedure in a general multi-source multi-target case, there are several options. One possible way is to alternate between all translation directions in a system, but in this case, the possible concern is that the model parameters may start to oscillate between these directions. Therefore, we propose to simultaneously feed data from all source languages, stack representations at points where specific encoders merge into shared ones, and pass them through the chain of decoders down to individual decoders. There are many other decisions to be made, such as what should be the proportion between samples from different language pairs,
should samples be over-sampled, should they be weighted, etc.

4 Experiments and Results

For the implementation, we selected three languages; 1) Russian (as a high-resource language from the East-Slavic subgroup within the “Indo-European” language family), 2) Kazakh, and 3) Tatar. The last two languages belong to the “Kipchak” sub-branch of the Turkic language family and hence are quite related. Kazakh and Tatar languages are far from Russian both by the number of speakers and the amount of linguistic resources, but Kazakh is a higher-resource language than Tatar in general. All three languages use the Cyrillic script.

Taking two bilingual corpora, Tatar-Russian and Kazakh-Russian. We train two separate bilingual Tatar-to-Russian and Kazakh-to-Russian translation models. Then, we employ the same corpora, and we create the multilingual model as shown in Figure 2, where Tatar and Kazakh languages are on the source side and Russian is on the target side. Both bilingual and multilingual models meet the same data, the same number of epochs, and the same number of parameters for each translation path.

To understand whether introducing a hierarchy into an encoder is beneficial, following Johnson et al. (2017), we train the model with a fully shared encoder for both source languages. We compare this model to the hierarchical model from the first set of experiments. Model architecture and size correspond to that of bilingual models. Data exposure is the same as in the hierarchical multilingual model.

Finally, to check if language relatedness indeed matters, we replace the Kazakh language with Ukrainian (the language from the same subgroup as Russian), also using the Cyrillic script. The Ukrainian language is as far from Tatar as Russian is. So, we want to test if using non-related language reduces the performance of the hierarchical multilingual model.

For Tatar-to-Russian translation, we utilized the first 300K sentence pairs from the private Tatar-Russian bilingual corpora provided by the Institute of Applied Semiotics, Academy of Sciences of the Republic of Tatarstan. For Kazakh-to-Russian translation, we acquired the bilingual corpus from WMT’19, cleaned it and used the first 700K sentence pairs. Finally, for Ukrainian-to-Russian, we took the data from JW300 dataset (Agić and Vujić, 2019) and used the 700K sentence pairs. After filtering sentences by maximum length (50 BPE (Sennrich et al., 2016) tokens) only about 250K Tatar-Russian pairs and about 450K Kazakh-Russian and Ukrainian-Russian pairs remained. Since Tatar-Russian pair is less in quantity, we reiterated them within every epoch such that on each step the system sees both languages in equal proportions. For comparability, bilingual Tatar-Russian model was exposed to the data in the same fashion, with reiterating.

Our multilingual NMT system implementation is based on Transformer architecture (Vaswani et al., 2017). We set the number of layers in individual encoders to 2, number of layers in shared encoder to 2, and the number of layers in decoder to 4. We also set the dimensional model to 256, the number of hidden layer to 1024, number of heads to 8, dropout to 0.1, and the number of epochs to 20. Bilingual models also use this architecture; the difference is that there is a single 4-layer encoder.

We trained the models with two batch sizes (128 and 256) because translation quality seems to highly depend on it. We compare bilingual models with a particular batch size to multilingual models with the same batch size. For multilingual models, half of the samples in a batch comes from the first source language and the other half—from the second. To train the hierarchical model, we simultaneously feed input data from both source languages to respective encoders and stack their outputs before passing to the shared encoder. On the decoder side, the final model outputs are compared to expected outputs for both languages (also stacked).

The results of the first set of experiments are summarized in Table 1.
As seen, the hierarchical model surpassed bilingual models in both translation directions and for both batch sizes. However, the most improvements, up to 1.26 BLEU, were obtained for low-resource tt-ru translation. This suggests that both directions are able to benefit from information captured by a shared encoder, but it is more helpful to the weakest pair.

In the second set of experiments, the goal was to check whether dividing the encoder to individual and shared parts is beneficial for learning. From Table 2, we observe that in all but one case the hierarchical model performs better than the one with single encoder, and the difference is considerable especially for kk-ru direction (0.89 BLEU). This result is important because it empirically confirms our assumption about the usefulness of explicitly defining parameter sharing strategy for a multilingual model.

The third set of experiments, Table 3, shows another interesting results. Tatar language is not able to learn much from the Ukrainian language. For batch size 128, the BLEU score only decreases, and for batch size 256, it increases only by 0.07 BLEU. For the higher-resource uk-ru translation, although the scores improve, less than they did for kk-ru translation.

Based on these positive results for the simple multi-source case, we conclude that the idea of introducing a hierarchy reflecting language similarity and controlling between-language parameter sharing, into multilingual architecture, has strong potential and should be explored further.
5 Conclusions and Future Work

In this paper, we proposed a hierarchical approach in multilingual NMT. The proposed approach was tested on the simple problem with two related source languages and only one target language. Our experimental results showed that the system is functioning, and in the case when source languages are related, both of them can exploit of dividing the encoder into language-specific and shared parts. We also found positive evidence supporting the assumption about the usefulness of explicitly defining parameter-sharing strategy of a multilingual model showing that parameter-sharing between related languages reaps more benefits.

As the future work, we would like to test the proposed approach on a larger set of languages on both the encoder and decoder sides of the translation model and explore different approaches to parameter-sharing.

Acknowledgements

The authors would like to express their sincere gratitude to Dr. Raul Aranovich (University of California, Davis) for all his support. The authors also would like to acknowledge the financial support received from the Linguistics Department at UC Davis.

References

Agić, Ž. and Vulić, I. (2019). JW300: A wide-coverage parallel corpus for low-resource languages. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3204–3210.

Ahmadnia, B. and Dorr, B. J. (2019). Augmenting neural machine translation through round-trip training approach. Open Computer Science, 9(1):268–278.

Dong, D., Wu, H., He, W., Yu, D., and Wang, H. (2015). Multi-task learning for multiple language translation. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, pages 1723–1732.

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.

Sachan, D. S. and Neubig, G. (2018). Parameter sharing methods for multilingual self-attentional translation models. In Proceedings of the Third Conference on Machine Translation.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 1715–1725.

Tan, X., Chen, J., He, D., Xia, Y., Qin, T., and Liu, T.-Y. (2019). Multilingual neural machine translation with language clustering. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 963–973.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In Proceedings of the 31st Conference on Advances in Neural Information Processing Systems, pages 5998–6008.