Nat. Volatiles & Essent. Oils, 2021; 8(6): 6230-6234

NVEO
Natural Volatiles &
Essential Oils

Multilingual Machine Translation: Deep Analysis Of Language-Specific Encoder-Decoders

NANDHINIDEVI.S[1] , Dr. N. Sundararajulu[2]

[1]Assistant Professor, Department of Information Technology, Dhanalakshmi Srinivasan College of Engineering and Technology

[2]Professor, Dhanalakshmi Srinivasan College of Engineering and Technology

Abstract

State-of-the-art multilingual machine translation relies on a shared encoder decoder. In this paper, we propose an alternative approach based on language specific encoder-decoders, which can be easily extended to new languages by learning their corresponding modules. To establish a common interlingua representation, we simultaneously train N initial languages. Our experiments show that the proposed approach improves over the shared encoder-decoder for the initial languages and when adding new languages, without the need to retrain the remaining modules. All in all, our work closes the gap between shared and language-specific encoder-decoders, advancing toward modular multilingual machine translation systems that can be flexibly extended in lifelong learning settings.

Keyword: Multilingual machine translation, Lifelong learning settings.

1. Introduction

Multilingual machine translation is the ability to generate translations automatically across a (large) number of languages. Research in this area has attracted much attention in recent years, from both the scientific and the industrial community. With the recent shift at a neural machine translation paradigm (Bahdanau, Cho, & Bengio, 2015), the opportunities for improvements in this area have dramatically expanded. Thanks to the encoder-decoder architecture, there are viable alternatives to expensive pairwise translations based on classic paradigms.

The main proposal in this direction is the shared encoder-decoder (Johnson et al., 2017) with massive multilingual enhancements (Arivazhagan et al., 2019b). While this approach enables zero-shot translation and is beneficial for low-resource languages, it has multiple drawbacks: **(i)** the entire system has to be retrained when adding new languages or data or alternatively, use an adapter module to add a new language (Bapna, Arivazhagan, & Firat, 2019); **(ii)** the quality of translation drops when adding too many languages or for those with the most resources (Arivazhagan et al., 2019b); **(iii)** the shared vocabulary grows when adding a large number of languages (especially when they do not share alphabets); and **(iv)** the shared encoder is not able to add multiple modalities such as image or speech.

In this paper, we propose a new framework that can be incrementally extended to new languages without the aforementioned limitations. Our proposal is based on language-specific encoders and decoders that rely on a common intermediate representation space. For that purpose, we simultaneously train the initial N languages in all translation directions. New languages are naturally added to the system by training a new module coupled with any of the existing languages, while new data can be easily added by training only the module for the corresponding language.

We evaluate our proposal on three experimental configurations: translation for the jointly trained initial languages, translation when incrementally training a new language, and zero-shot translation. Our results show that the proposed method is competitive in the first two configurations, but still lags behind the shared encoder-decoder in zero-shot translation. In order to further understand our model and as an extension of the previous publication by Escolano, Costa-jussa and Fonollosa (2021), we provide a deeper analysis in the following directions. We study the effect of fine-tuning and our approach shows robustness by avoiding catastrophic forgetting. We analyze why our model does not suffer from the attention mismatch, mentioned in previous works, even though the modules do not share parameters (Firat, Cho, & Bengio, 2016a). To perform this analysis we explore the effect of excluding training data from certain language pairs. We observe that when there are four languages in the initial system, when we train with only parallel data from and to one language, our system is not able to learn all the translation directions. However, when adding one more language (parallel data from and to two languages), our system achieves almost full performance compared to training with all the translation directions from the initial languages in the system. Then, to better understand the nature of the learned representations, we run additional experiments on natural language inference, where the language-specific encoder-decoders internal representation is evaluated in all the encoder layers. Finally, we visualize these representations in a two dimensional space.

Overall, we provide a deep analysis of this new multilingual model based on language-specific encoder-decoders that can incrementally be extended to new languages and that can improve the performance of multilingual machine translation without parameter sharing, closing the existent gap with shared encoder-decoders architecture.

The rest of the paper is organized as follows. Section 2 reviews most related work. Section 3 details the proposed method for multilingual machine translation. Section 4 overviews experiments in machine translation where we compare our proposed method to the shared encoder/decoder which is considered the state-of-the-art in current multilingual approaches. Section 5 provides an in-depth analysis of the intermediate representations created with our proposed method by showing the results in natural language inference and visualizing some intermediate sentence representations. Finally, Section 6 concludes and suggests new research paths for future studies.

2. Proposed System

Our proposed approach trains a separate encoder and decoder for each of the N languages available. We do not share any parameter across these modules, which allows us to add new languages incrementally without retraining the entire system. In contrast to Escolano et al. (2019), we do not force the

intermediate representation to be the same, and therefore, we do not require multi-parallel corpus to train our system.

We denote the encoder and the decoder for the ith language in the system as ei and di, respectively. For language-specific scenarios, both the encoder and decoder are considered independent modules that can be freely interchanged to work in all translation directions.

3. Experiments in Multilingual Machine Translation

In this section we review the machine translation experiments in different settings. Since the main difference between the shared and the language-specific encoders-decoders lies in whether they retrain the entire system when adding new languages, we accordingly design our experiments to compare this aspect of the systems.

We denote the encoder and the decoder for the ith language in the system as ei and di, respectively. For language specific scenarios, both the encoder and decoder are considered independent modules that can be freely interchanged to work in all translation directions. In what follows, we describe the proposed method in two steps: joint training and incremental training.

a. Joint training The straightforward approach is to train independent encoders and decoders foreach language. The main difference from the standard pairwise training is that, in this case, there isonly one encoder and one decoder for each language, which will be used for all translation directionsinvolving that language.

b. Incremental training Once we have our jointly trained model forNlanguages, the next stepis to add new languages. Since parameters are not shared between the independent encoders anddecoders, the basic joint training enables the addition of new languages without the need to retrainthe existing modules.

4. Experiments in Multilingual Machine Translation

In this section we review the machine translation experiments in different settings. Since the main difference between the shared and the language-specific encoders-decoders lies in whether they retrain the entire system when adding new languages, we accordingly design our experiments to compare this aspect of the systems.

5. Analysis of the Intermediate Representations

In this section, we want to better understand the capabilities of our model and we analyze the quality of the intermediate representations by means of a probing classification task. This method has been proposed before as a measure of the cross-lingual capabilities of NMT systems (Eriguchi et al., 2018; Siddhant et al., 2020; McCann et al., 2017; Conneau et al., 2018), using natural language inference and visualization techniques.

6. Conclusions

In this paper, we present a novel method to train language-specific encoders-decoders that allows incremental additions of new languages in the system without having to retrain the entire system or, add any adapter. We believe that this approach can be particularly useful for situations in which a rapid extension of an existing machine translation system is critical.

For the initial languages in the system, the language-specific encoder decoders outperform the shared architecture by 0.4 BLEU points on average. When adding a new language, the language-specific encoder-decoders outperforms the shared ones by 3.4 BLEU points on average and, most importantly, the training of this new language was done in only one day, as opposed to the week taken by the shared system. Additionally, by design, there is no variation in the quality of languages in the initial system when adding a new language.

A further analysis of our model in fine-tuning shows more robustness by avoiding catastrophic forgetting. Moreover, we do not need parallel data among all language pairs in the initial system to learn translations from and to all languages; however, we at least need parallel data with more than one language. In this sense, language-specific encoders-decoders could take further benefit from incremental training with more than one language in the initial system.

We also examine the quality of the intermediate cross-lingual representation created with our proposed model in the application of natural language inference. We see that the higher the encoder layers are, the better the quality. Additionally, an intuitive visualization example shows that the sentences in different languages appear close in the space, but not exactly at the same point. When compared to the shared system we observe that parameter sharing provides better cross-lingual representations for the probing task, which correlates with the difference in performance of the systems on zero-shot translation.

Our work substantially closes the existing gap between the language-specific and the shared encoders-decoders, while maintaining the flexibility that results from not sharing parameters. Similar to Arivazhagan et al. (2019b), our results suggest that the shared architecture is beneficial for languages that share the same script because of their joint vocabulary and is detrimental for languages that do not share the same script, due to negative transfer between languages in the system. This behavior is not observed on the language-specific system as each language has its own vocabulary and embeddings.

In the future, we would like to further compare the shared and language specific encoders-decoders in cases where the languages do not share scripts (e.g. Chinese, Arabic, Russian and Greek) to see if our model has even more advantages over the shared system under these conditions.

7. References

Aharoni, R., Johnson, M., & Firat, O. (2019). Massively multilingual neural machine translation. 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), pp. 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Arivazhagan, N., Bapna, A., Firat, O., Aharoni, R., Johnson, M., & Macherey, W. (2019a). The missing ingredient in zero-shot neural machine translation. ArXiv, abs/1903.07091.

Arivazhagan, N., Bapna, A., Firat, O., Lepikhin, D., Johnson, M., Krikun, M., Chen, M. X., Cao, Y., Foster, G., Cherry, C., Macherey, W., Chen, Z., & Wu, Y. (2019b) Massively multilingual neural machine translation in the wild: Findings and challenges. CoRR, abs/1907.05019. Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In Bengio, Y., & LeCun, Y. (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.

Bapna, A., Arivazhagan, N., & Firat, O. (2019). Simple, scalable adaptation for neural machine translation..

Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S., Schwenk, H., & Stoyanov, V. (2018). XNLI: Evaluating cross-lingual sentence representations. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Dong, D., Wu, H., He, W., Yu, D., & 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 (Volume 1: Long Papers), pp. 1723–1732, Beijing, China. Association for Computational Linguistics.