Transfer learning in low-resourced MT: An empirical study

Sainik Kumar Mahata[1] and **Dipanjan Saha**[2] and **Dipankar Das**[3] and **Sivaji Bandypodhyay**[4]
Jadavpur University, Kolkata, INDIA
[1]sainik.mahata@gmail.com, [2]sahadipanjan6@gmail.com
[3]dipankar.dipnil2995@gmail.com, [4]sivaji.cse.ju@gmail.com

Abstract

Translation systems rely on a large and good-quality parallel corpus for producing reliable translations. However, obtaining such a corpus for low-resourced languages is a challenge. New research has shown that transfer learning can mitigate this issue by augmenting low-resourced MT systems with high-resourced ones. In this work, we explore two types of transfer learning techniques, namely, cross-lingual transfer learning and multilingual training, both with information augmentation, to examine the degree of performance improvement following the augmentation. Furthermore, we use languages of the same family (Romanic, in our case), to investigate the role of the shared linguistic property, in producing dependable translations.

1 Introduction

For any machine translation (MT) system to produce reliable translations, it needs to be trained using a large and good-quality parallel corpus. Although, these corpora are abundant for European languages with high digital presence (where both the participating languages are high-resourced (HR)), finding one for low-resourced (LR) languages is difficult (one of the participating languages is LR). This leads to the development of a small-sized parallel corpus, which when used in training does not produce robust MT systems and translations.

Over the years, many research works have focused on the effect of transfer learning on MT. This translates from experiments where LR language data have been augmented using HR data to produce better outputs when compared to vanilla MT models trained using LR languages only (Gu et al., 2018). Furthermore, training HR and LR MT models together have also given considerable improvements in the quality of MT output.

This paradigm for multilingual translation can be compared naively to how human individuals learn new languages and acquire new languages. In the Indian context, for instance, a natural English speaker who is conversant in Hindi could pick up Marathi more quickly than someone completely ignorant of any Indic languages. Second language acquisition is associated with cognitive rewiring and anatomical alterations in the human brain, as proposed by Li et al. (2014). Furthermore, Schepens's empirical research (Schepens et al., 2016) shows that bilingual speakers' ease of acquiring a third language is closely correlated with the languages' distance from one another. The authors noted that several factors, including anthropological development, speaker geography, vocabulary exchange, syntactic structural similarities, etc., influence how simple it is to acquire a new language. Pagel et al. (2007) demonstrated in a different work the connection between word usage frequency and the development of Indo-European languages.

The previous works inspired us to look into the cases where LR MT models may gain information when trained incrementally on already trained HR MT models (where the participating languages belong to the same language family), with information augmentation, using various transfer learning approaches. We used two transfer learning approaches to test this. The first one, called the cross-lingual transfer technique with information augmentation, initially trains an HR MT system. After the training is over, the weights of the HR model are saved. Thereafter, an LR MT system is trained, where the saved weights of the HR model are then used to initialize the model training. Furthermore, before sending the information to the decoder, the input sentence is passed through both encoders (both HR and LR) and the resulting context vectors are concatenated. This method allows the transfer learning model to gain from already trained HR model weights as well as to benefit from the already trained HR encoder vector.

In the second approach, called multilingual learning with information augmentation, a multilingual HR MT system is trained with multiple HR language pairs. Similar to the previous approach, the weights are saved again. The same process, including the HR weights and concatenation of the HR encoder and LR encoder vector, is followed before sending information to the decoder. The LR model, in this approach, benefits from the shared linguistic property of the participating HR languages that belong to the same language family.

2 Related Work

The training of NMT systems without the use of parallel or comparable data has been extensively studied. Such studies typically combine designs incorporating several encoders and decoders (Lample et al., 2018; Artetxe et al., 2018), heavily rely on cross-lingual embedding, and perform iterative back translations (Sennrich et al., 2016). Additionally, it has been discovered that training NMT systems in multiple languages enhances translation performance (Dong et al., 2015; Firat et al., 2016; Johnson et al., 2017).

A single encoder that is shared by all languages and a decoder for each language were utilised by Dong et al. (2015) and Sen et al. (2019). A single shared attention mechanism is used by Firat et al. (2016) to propose multi-way cum multilingual NMT with multiple encoders and decoders. Johnson et al. (2017) developed a more straightforward yet efficient method that only required one encoder and one decoder. All of the parallel data were combined into a single corpus after certain unique tokens were added at the start of each sentence. A transfer-learning strategy was put up by Gu et al. (2018) to share lexical and sentence-level representations from several source languages into a single target language.

A similar architecture was described by Zoph et al. (2016) who trained a high-resource language pair (the parent model) first. The weights from the parent model were initially initialised in the child model, which was trained on the language pair with limited resources. They assisted with the English translation of Hausa, Turkish, Uzbek, and Urdu by using a French–English parent model.

To the best of our knowledge, there hasn't been any published research on how, when all languages belong to the same language family, a system trained in an HR language can improve the translation output of a system trained in an LR language. Furthermore, we didn't merely use the HR model's weights to initialise the LR model. In addition, the encoder vectors of the LR model were concatenated with the HR model's vectors for each source sentence, fed with the same sentence, and then sent to the decoder for prediction.

3 Data

As mentioned earlier, we wanted to test whether MT models trained using HR languages, benefit an LR MT model when used in accordance to transfer learning approach. Furthermore, we wanted to keep the participating languages belonging to the same language family, so that we can investigate the concept of shared linguistic property. For this, we stuck to experimenting with the Romanic language family, which consists of languages like French, Italian, Spanish, Portuguese and Romanian.

For experimentation purposes, we considered (French, Italian, Spanish, and Portuguese) – English language pairs as HR and Romanian – English language pairs as LR for the second transfer learning approach. On the contrary, For the first approach of transfer learning, we considered French – English as the HR language and Romanian – English as the LR language. For this, the amount of data for HR language pairs was adjusted to 100k and 50k for the LR language pair. We named these experiments as $\mathrm{MT}_{\mathrm{SameFamily}}$ as a whole.

Moreover, to find whether languages from the same language family only aid in better translations, we experimented with a separate language family (Slavik, in our case), which consisted of Bulgarian, Czech, Polish and Slovak languages. These languages again, were considered as HR languages and 100k parallel sentences (HR – English) of these were considered for the second transfer learning approach. For the first transfer learning approach, we considered Bulgarian – English language pair as the HR language. We named these experiments as $\mathrm{MT}_{\mathrm{DiffFamily}}$ as a whole.

The parallel corpus for the above-mentioned languages was extracted from the Europarl project[1].

4 Methodology

The proposed approach starts with training in the cross-lingual transfer technique with information

[1] https://www.statmt.org/europarl/

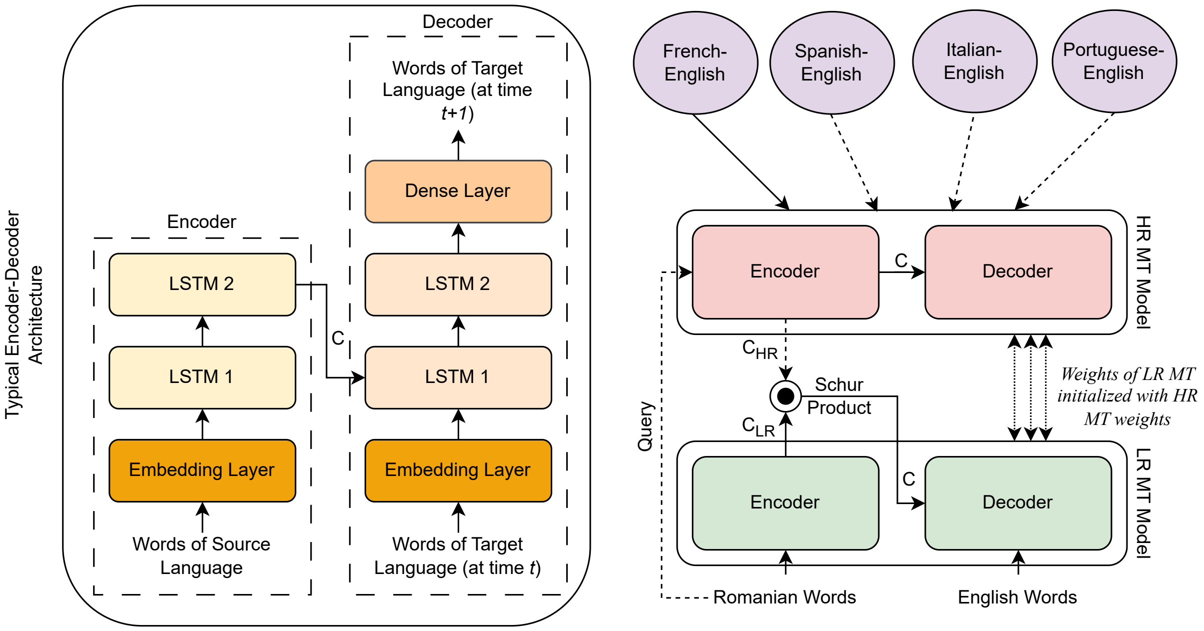

Figure 1: Transfer learning based MT architecture.

augmentation. For this approach, the HR MT model (essentially a seq-to-seq model based on LSTM cells) was first trained using the French – English language pair, for the $\text{MT}_{\text{SameFamily}}$ experiments and Bulgarian – English language pair, for the $\text{MT}_{\text{DiffFamily}}$ experiments. The hyper-parameters for training these models were as follows. The batch size was set to 64, the number of epochs was set to 100, the activation function was softmax, the optimizer chosen was rmsprop and the loss function used was sparse categorical cross-entropy. The learning rate was set to 0.001.

After the HR MT models were trained, the weights were saved. Thereafter, the LR MT model was trained using the Romanian – English parallel corpus. The weights of the LR MT model were initialized with the saved weights of the HR MT model. This time around, the source Romanian sentence was fed to the encoder of the LR MT model. At the same time, this sentence was also sent to the trained encoder of the HR MT model. Both resultant context vectors, C_{HR} and C_{LR} were concatenated using Schur product operation. This concatenated context vector was then sent to the decoder part to complete the training process of the LR MT model. The hyper-parameters for this model were kept alike with the HR MT model.

For training the multilingual learning technique with information augmentation, the HR MT model was trained using a mixture of language pairs, with sentences belonging to the same language family for $\text{MT}_{\text{SameFamily}}$ and different language family for $\text{MT}_{\text{DiffFamily}}$. For this, the sentences from language families were coupled together. This gave rise to two source corpora, that consisted of a mixture of French, Spanish, Italian and Portuguese languages and Bulgarian, Czech, Polish and Slovak languages. The target corpus consisted of all English sentences.

The same procedure of training HR MT models first, saving the weights, passing Romanian sentences through the encoder of the LR MT model and HR MT model to get concatenated context vector, etc., was followed. The same hyper-parameters were used to train the second transfer learning approach as well. The whole architecture followed for training both our approaches has been depicted in Figure 1.

Also, a vanilla MT model, using seq-to-seq architecture was trained using 50k sentences of Romanian – English and this acted as our baseline system through which we could investigate the performance gain via transfer learning. The same hyper-parameters, as mentioned above, were used to train this baseline model.

5 Results and Discussion

The results of the above experiments were quantified using automated MT evaluation metrics. For the automated metrics, we used BLEU (Papineni et al., 2002), ChrF (Popović, 2015) and TER (Snover et al., 2006) to evaluate the translation quality. The results have been shown in Table 1.

Ln Family	MT Models	BLEU	TER	ChrF
	Baseline MT Model	11.26	89.43	23.15
Romanic	**Cross-lingual Transfer Learning**	12.25	86.98	25.87
	Multilingual Learning	13.88	80.47	27.91
Slavik	**Cross-lingual Transfer Learning**	10.87	91.27	21.52
	Multilingual Learning	11.13	90.15	22.71

Table 1: Automated evaluation of the MT models for both language families.

From Table 1, we can see that both transfer learning approaches of $\text{MT}_{\text{SameFamily}}$ outshine the baseline model. Though the difference is not substantial, it is because we have used subsets of the actual parallel corpora to perform our experiments. As neural machine translation (NMT) is a data-hungry process, using full datasets for training purposes can generate much more substantial gains.

Also, we found out that the multilingual transfer learning approach produced better translations as compared to the cross-lingual approach. This is because this type of transfer learning allows knowledge to be transferred so that all languages can gain from one another. In our case, the multilingual transfer learning MT model benefits from the knowledge of all Romanic languages as compared to only one in the case of the cross-lingual transfer learning MT model.

For experiments of $\text{MT}_{\text{DiffFamily}}$, we see that both the transfer learning approaches do not produce improvements in the quality of translation output.

6 Conclusion and Future Work

In this work, we see the benefit of using transfer learning approaches for low-resourced machine translation. In our case, we experimented with the Romanic language family, which consisted of languages like French, Italian, Spanish, Portuguese and Romanian. Another set of experiments with the Salvik language family, which consisted of languages like Bulgarian, Czech, Polish and Slovak were also performed.

While it is a known fact that transferring knowledge of a high-resourced MT model into a low-resourced MT model aids in producing better translation quality, we wanted to test the same concerning a language family. For our experiments, we considered languages of the same family (French – English, Spanish – English, Italian – English and Portuguese – English) and different language family (Bulgarian – English, Czech – English, Polish – English and Slovak – English) as high-resourced and Romanian – English as low-resourced.

Experimentation, using two types of transfer learning approaches shows that performance (BLEU, TER and ChrF) does increase when languages of the same language family are used in the transfer learning setting, as this approach takes advantage of the common linguistic property shared by the languages of a same family.

References

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *ICLR (Poster)*. OpenReview.net.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 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)*, pages 1723–1732.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor OK Li. 2018. Universal neural machine translation for extremely low resource languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 344–354.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *ICLR (Poster)*. OpenReview.net.

Ping Li, Jennifer Legault, and Kaitlyn A. Litcofsky. 2014. Neuroplasticity as a function of second language learning: Anatomical changes in the human brain. *Cortex*, 58:301 – 324.

Mark Pagel, Quentin D Atkinson, and Andrew Meade. 2007. Frequency of word-use predicts rates of lexical evolution throughout indo-european history. *Nature*, 449(7163):717.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Job J Schepens, Frans van der Slik, and Roeland Van Hout. 2016. L1 and l2 distance effects in learning l3 dutch. *Language Learning*, 66(1):224–256.

Sukanta Sen, Kamal Kumar Gupta, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Multilingual unsupervised nmt using shared encoder and language-specific decoders. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3083–3089.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. 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.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.