School of Informatics

Informatics Research Review
Neural Machine Translation for low-resource languages: Research Review of current techniques

B248125
January 2024

Abstract

Neural Machine translation is a promising approach for translation, but its effectiveness is often defined by the amount of parallel training resources. Acquiring such resources for many of the languages is difficult, especially those at risk of extinction. Many native speakers of those languages could benefit substantially from effective machine translation. This paper is a review of techniques presented by research that are viable options for improving the quality of Neural Machine Translation in cases of low-resource languages. In this text, various techniques have been discussed for improving language translation models. These techniques include transfer learning, using monolingual data, and training on multiple language pairings. Recommendations and conclusions have been drawn based on the analysis of the aspects of each of the techniques including potential issues with bias when using those techniques.

Date: Saturday 20th January, 2024

Supervisor: Xingran Ruan

1 Introduction

1.1 Background

Neural Machine Translation (NMT) is a technique that utilises neural networks to translate text between languages. The typical approach for this is to first procure a large amount of text (also known as corpus) that has been translated (presumably by a human translator) from one of these languages to the other. These corpora are often called parallel corpora. This corpus will then be used to train a model that will learn the relationship between the text in the first and the second languages, allowing it to then translate unseen text into the second language. Koehn & Knowles (2017) hold the view that uncommon words and the amount of training data are key challenges for NMT.

For some language pairings, we do have large parallel corpora, especially for major languages such as English or Mandarin. However, machine translation is also useful in the context of other, less widely used languages (referred to often as low-resource languages). According to Eberhard et al. (2023) there are 7,168 languages in use as of October 29, 2023. The fact that this number is accurate only at the time of writing is highly relevant as 40% of those languages are currently considered endangered (Eberhard et al. 2023). Furthermore, Harrison (2007) highlights the noteworthy concern that many linguists have, that only about half of the languages spoken in 2001 will still be spoken by the year 2101. These two rather pessimistic considerations show that for many language pairs, we are not able to produce large corpora and that for many of them, the time for creating any corpora may be limited.

1.2 Motivation

Procurement of large parallel language corpora is difficult and requires extensive resources for the vast majority of language pairings, whilst it is not as difficult to procure a monolingual language corpus (Edunov et al. 2018). The development of reliable machine translation for less widely used languages will have more than just benefits for people who only speak a minority language. This development could be beneficial to the survival of dying languages. Harrison (2007) mentions the connection between the loss of a native language and the disappearance of the culture and traditions. These considerations point to the notion that there are many benefits to finding procedures that will improve NMT in cases of limited parallel corpora.

1.3 Objectives

The objective is to present several methods that claim to improve the performance of Neural Machine Translation in situations where the parallel data is limited. This paper will not explore any methods for improving Statistical Machine Translation or make comparisons between SMT or NMT. Furthermore, the techniques presented here are limited to only text translation and the consideration will only take into account their performance in that respect. The methods studied will be evaluated and where appropriate compared to produce a set of recommendations for techniques when undertaking this task.

1.4 Research Questions

The set of research questions this paper attempts to answer is as follows:

1. What are the methods for improving NMT in cases where the amount of parallel data is limited?
2. What are the limitations and use cases of these procedures?
3. How feasible are these techniques for organizations with limited resources?
4. Which technique is preferable when the use of one technique is mutually exclusive with another?

1.5 Report Structure

This review begins with an introduction of the problem, which introduces the background and motivation behind this review. Those subsections are followed by an outline of the research questions and objectives set out for this review. The introduction is followed by a literature review which introduces several techniques shown in research. The procedures have been grouped into subsections whenever they had significant similarities in terms of the procedure itself or the circumstance where the technique is useful. This produced a set of three subsections: Transfer Learning approaches, Leveraging Monolingual Data, and Leveraging data between multiple language pairings. The considerations of each section and the included techniques are summarized in the Summary & Conclusion section of this report. Furthermore, this section compares the mutually exclusive techniques and includes several recommendations for different low-resource translation use cases.

2 Literature Review

2.1 Transfer Learning approaches

Transfer Learning is a technique where in order to improve the performance of a model attempting to solve a task by transferring knowledge acquired by a model solving a similar task or a task that is a superset of that initial task (Torrey & Shavlik 2009). In many cases, this transfer of knowledge is done by copying parameters as a pretraining step. Zoph et al. (2016) presents an argument for using transfer learning as a means of improving machine translation for low-resource languages. Their technique is to first train a model on a high-resource language pairing (French - English) and then initialize a new model using the trained weights from the first model and continue training it on a low-resource language pairing (e.g. Uzbek - English). In their experiments, they use a dataset of much lower quality, as it is quite small and the domains from which the data originated were varied, which more closely resembles a challenging real-world translation task. They show a significant improvement in Uzbek to English translation (as the BLEU[1] score improved from 10.7 to 15), even though the pre-training language pair used (French - English) is quite distant from the Uzbek language.

Importantly in this technique after copying some of the weights are fixed such as the English side word embeddings, this is explained as a means of regularization, but it additionally can prevent forgetting, which is very likely given the aforementioned low quality of the English-Uzbek dataset. Their findings show that in their task fixing the weights for only the Target Input and Output Embeddings has yielded the best BLEU results.

The authors also emphasized the fact that if languages chosen for this task are related this should yield even better results. Nguyen & Chiang (2017) have experimented with utilising transfer learning technique presented by Zoph et al. (2016) to also improve Uzbek - English translation. However, their study utilised a different also relatively low-resource language pairing for the pretraining step which is English-Turkish. This pairing is closer related to the target language which is Uzbek, as Turkish and Uzbek are highly related languages. They found that in those cases transfer learning is also beneficial and when combined with the Byte Pair Encoding method for word segmentation (Sennrich et al. 2016) the improvements have been much more substantial.

In conclusion, using transfer learning can be perceived as a viable method for improving machine translation for low-resource languages. It is also important to note that this technique is very viable for organizations with limited resources in general as it only requires the use of parallel data for one other language pairing or the use of a publicly available pre-trained model. Additionally, as can be seen in the first study, the language pairing does not even need to be related to the languages in the original task to achieve great improvements. On the other hand, fine-tuning and deciding on an architecture might become difficult as learning has two distinct phases, each with a different objective.

2.2 Leveraging Monolingual data

Monolingual datasets are only composed of data that are exclusively in one language. Therefore, there are many more available sources for that data making those data sets much bigger for resource-poor languages. Because of the availability of this kind of data it is important to explore the potential ways of utilizing it.

[1] For a detailed explanation of BLEU score please refer to Papineni et al. (2002)

2.2.1 Training with only monolingual data

Conneau et al. (2018) explores a discriminative learning approach to learning translations between languages without using any parallel data whatsoever. The model is tasked with aligning word embeddings from two languages, but without knowing which alignments were correct because asserting that would require parallel data. The objective function of this model is to be able to produce a word embedding for every word from the input data set such that it would be as difficult as possible to classify whether the output originates from the model or the target dataset directly. In practice, this objective function aligns the word embeddings from the input language to the target language.

The rationale behind such an approach is the fact that even though human languages are very different, the words within them use still are mostly concerned with explaining the largely similar reality around us. Therefore if the set of word embeddings for a given language abstracts the syntactical realities of that language to a satisfactory extent and models the context of the uses of that word, there should be a way to align many of the words with any other language using word embeddings alone. In this paper, the anticipated result is that the mathematically optimal alignment will match the linguistical reality.

Several challenges have been encountered in the pursuit of good performance. Firstly, the word embeddings of very uncommon words are unlikely to truly model the contexts and they are unlikely to match between languages. Therefore only the most frequent words have been retained for training along with their nearest neighbours.

When using the aligned model it is important to have a robust way to interpret the produced alignments in the context of word-to-word translation. The use of nearest neighbor selection is not desired as by its nature it is not symmetrical, whilst word translation in most cases is a symmetrical relationship (e.g. cat (EN)→gatto (IT)→cat (EN)). In order to address this issue the proposed solution is to use a measure proposed in the same study called: Cross-Domain Similarity Local Scaling. The summarization of that method is out of the scope of this review, this method can be reviewed in the original paper (Conneau et al. 2018).

Upon addressing these challenges the final performance was very promising. The performance in terms of precision of correct word translation is shown to be above 70% for English-Spanish, English-French, and English-German. English-Russian, English-Esperanto, and English-Classical Mandarin were also tested producing much worse results, but their results showed that the model was able to learn to correctly translate over 20% words for these languages. The difference comes from the lexical dissimilarity between the languages and also the differences in linguistical structures leading to cases where a one-to-one mapping from word to word is not possible. This has been manifested by the inclusion of the same metrics for a supervised model that has given parallel data, which also achieved poor, although better, performance in the case of these language pairs.

One potential limitation of this study is the validation set that has been used. The validation that was performed in order to select parameters for the model has been also done in an unsupervised way, which in the context of this study makes sense as their goal was to achieve translations without using any parallel data, but in real-life problems, it is highly plausible that the procurement of small parallel validation set is feasible. In those cases the use of parallel data for validation could improve parameter selection, improving the performance even further.

Another limitation is the learning rate scheduling scheme used in this study, which is to halve the learning rate based on the performance of the validation set. It is plausible that a more modern learning scheduling scheme might be beneficial here. Therefore, experimentation with

Original:	you	think	i	can	not	do	it	but	i	can	-	-
Translated:	du	**denken**	ich	**können**	*nicht*	*tun*	*es*	aber	ich	kann	-	-
LM Improved:	du	**denkst**	ich	**kann**	*es*	*nicht*	*tun*	aber	ich	kann	**es**	**doch**

Table 1: English-German simulated example, where the word-by-word translation, is enhanced with the use of a language model (LM), Importantly adding words that are beneficial in the target language but are not present in the source language. The LM used to improve the sentence was GPT-3 (OpenAI 2023) prompted with 'Fix: <*word-by-word translation*>' to produce the sentence at the bottom of the table.

newer learning scheduling, such as ADAM (Kingma & Ba 2014) methods would potentially be beneficial to the performance as well.

2.2.2 Improving translation using monolingual language models

Gülçehre et al. (2015) shows two techniques for fusing a monolingual generative language model (LM) with a translation model. The rationale is that a monolingual language model will in many cases guess correctly the next word based on the translated sequence up to this word. This could help the model by making the prediction of a function word more likely, which is important in cases where the two languages do not have that word function in common. Therefore the expected outcome is making the translated sentences more grammatical. To illustrate this I have included an English-German handcrafted translation example in Table 1, and as it can be seen from it the use of a language model can confidently and correctly fix even a simple 'word-by-word' translation, which had no regard for grammatical rules and structures. As a part of their study, they have shown two possible ways of fusing the language model for the translation task. One approach is called Shallow Fusion the other is Deep Fusion. The key difference between these techniques is that in Shallow Fusion the LM output is summed to NMT. with weighted factor (β), whilst in the case of Deep Fusion the LM output is passed through an Affine Layer activated by a sigmoid before being added to the model. This makes the Deep Fusion model more flexible, but they add more parameters to the training process.

The findings of this paper are that the inclusion of Deep Fusion in the translation model has improved performance. The improvements shown are up to 2 BLEU in some cases, but the Shallow Fusion has even decreased the performance in some cases. Therefore the inclusion of this technique can be seen as beneficial but it is unlikely to provide ground-breaking improvements, even if the Deep Fusion technique is used. This could be in part explained simply by the nature of training that NMT performs, as it inherently makes the NMT model learn many of the patterns that an LM recognizes, therefore limiting the benefit of mostly duplicated information. An important limitation to note here is that many low-resources languages do not have great-performing language models available, therefore in many cases, the procurement of such a monolingual language model might be an additional and expensive task, as compared to taking an existing language model.

2.2.3 Back Translation

Back translation wide grouping of techniques that exploit an inherent property of translation, which is that a correct translation should be reversible back into the original language without loss of detail or meaning. This assumption is quite idealistic and in some cases, it is not

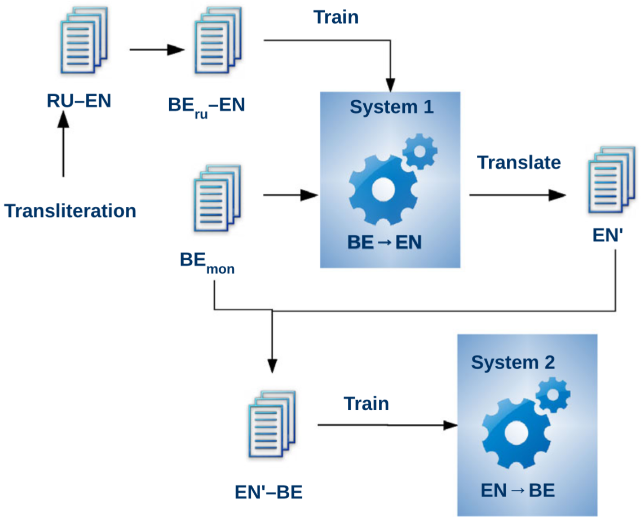

Figure 1: Figure 2 from Karakanta et al. (2018). "Experimental work-flow for the NMT system" (Karakanta et al. 2018)

achievable. For example, the German concept of 'schadenfreude' (the pleasure acquired from the misfortunes of others) or 'weltschmerz' (feeling of melancholy and world-weariness) (*Cambridge Dictionary* 2023) has no equivalent in many languages, therefore in the case of these words, the underlying assumption of the back translation does not hold. Fortunately, statistically speaking, those words are rare and most human interactions are back translatable (especially in the case of closely related languages).

One interesting way to harness this property is to treat NMT models as autoencoders for human language. Autoencoders are neural networks tasked with effectively copying their input into the output (Goodfellow et al. 2016), and the real goal of the model is to produce a representation (encoding) of the input in the process of copying. An example autoencoder task would be to copy an image made out of 1 000 000 pixels, to the output, but inside the model, a layer is constrained to only representing 10 000 pixels. This will make the model compress and extract only the most relevant 1% of the original information from the original image that is required to reconstruct it.

Going back to language, one desirable outcome from an autoencoder working through sentences would be that in one of the layers, a language-agnostic, synthesized, and compact representation of the sentence meaning is produced. This representation of the input sentence could be very useful for a model. Currey et al. (2017) present a very simple approach for achieving this. The solution presented is to augment the training set with some monolingual data, where the target and source are set to the same sentence. This augmentation has shown improved translation in the vast majority of experiments, but the performance changes range from a 0.3 BLEU decrease to a 1.2 BLEU increase. Additionally, the authors argue that a major reason for that is that the resulting model was more likely to preserve words that should not be translated, such as person names. Overall it can be seen as an easy-to-implement technique that can be advantageous in low-resource settings.

Karakanta et al. (2018) proposes another way of exploiting back translation. Their approach entails creating a model transliterating (translation character by character) from a high-resource related language into a low-resource language, in this case, Russian into Belorussian. This transliterating model is then used to create an English-Belorussian dataset based on an English-

Russian dataset. A new model is trained on this dataset, which is then used to translate large amounts of Belorussian monolingual data, resulting in a new English-Belorussian dataset. A visual representation of this approach has been included in 1. The final model resulting from this pipeline has achieved 10.83 BLEU on Belorussian to English translation, which given the low amount of resources and the distance between the English and Belorussian language is still a very respectable result. It is worth noting the transliteration task was trained using Wikipedia article names, which is a resource available for a multitude of languages. Therefore, the pipeline shown in this paper can be used for many translation tasks, provided a high-resource-related language can be found. An interesting future direction for this approach using data from knowledge bases such as Wikidata which encompass millions of concepts and many of them have labels in a large variety of languages, for the transliteration step.

2.3 Leveraging data between multiple language pairings

The multilingual approach to neural machine translation is an approach where the model simultaneously learns to translate between all possible permutations of language pairs for which the data was procured (Dong et al. 2015). This technique stands in stark contrast to others as the most prevalent approach to Machine Translation problems is to solve them for a pair of languages, rather than considering a set of languages and enabling translation between all pairs. A possible way of approaching multilingual tasks is to add a special token to the beginning of the sequence which instructs the model what the target language is. Johnson et al. (2017) show several reasons why this may be desirable, amongst them the possibility of using this technique for language with no parallel data and improving translation for languages where the data is limited. Another interesting implication of models trained in this manner is that they will likely include a language-agnostic representation of the input sentence somewhere in the layers of the models, which is the goal of using autoencoder strategies discussed in the previous section.

2.3.1 Zero-Shot

Johnson et al. (2017) illustrate the possibility of translating between languages where no parallel data has been provided in training. This is referred to as Zero-Shot Translation or implicit bridging. To show that multilingual models have this ability they have devised an experiment, where a model that did not train on any parallel data between Portuguese and Spanish, but rather only on data for the English-Portuguese pair and English-Spanish pair, still was able to produce relatively good translation as compared to baseline model, trained explicitly on Portuguese to Spanish data (explicitly bridged), as the BLEU score has only dropped by 21.4%. Furthermore, when the addition of a corpus of parallel Portuguese to Spanish data was considered, the multilingual model was able to surpass the performance of the baseline model when using only a fraction of the parallel dataset used by the baseline model.

It is important to note that the research group responsible for this study had access to extensive parallel language data that is exclusive to their organization, which was used for the zero-shot experiments. Many organizations, especially those attempting to preserve languages with low economic utility to large organizations, may not be able to procure many diverse parallel datasets for multiple languages, making this approach less viable in those cases.

2.3.2 Universal Lexical Representation

Gu et al. (2018) have focused their attention on the use of multilingual language models for low-resource languages, where finding parallel data with any language is difficult. Universal Lexical Representation (URL) is a technique they have proposed for improving performance in those cases. URL involves considering the mapping from a word to an embedding as a probabilistic distribution over a shared universal space of embeddings given the word embedding learned from monolingual data. This ensures that similar words from across languages can be aligned together, whilst preserving the original word embedding learned from monolingual data.

This method has been tested with three languages where for each the amount of parallel data was severely limited (thousands of sentences only). The languages considered were: Romanian, Latvian, and Korean. The language model was largely trained on European languages. An important thing to note about the languages used is that Lithuanian was not included. Therefore the languages considered show three very distinct use cases:

1. closely related languages are present in the training data (as **Romanian**, Italian, and Spanish are all Romance languages)
2. with similar properties and roots are present in the training data (as **Latvian**, Czech, and Russian are from the Balto-Slavic family of languages)
3. only very distant languages were present in the training data (**Korean** as opposed to all other languages in the data set that are Indo-European)

The findings were quite good when considering how little parallel data was used for this task. In the case of Romanian where the model has the closest relation to the other languages, the final achieved BLEU score when also utilizing back translation was only 21.1% behind the baseline trained directly on parallel English to Romanian data. Unfortunately in the case of Latvian, the final performance was quite poor, but not insignificant, but the use of ULR has helped significantly, as BLEU has risen 57.3%. Finally, in the case of the Korean language, the performance was very poor, but again ULR has proven to be a viable technique to help in this case as BLEU has increased by 124%. Therefore, even though the use of ULR can improve performance significantly unless there are large amounts of data available for closely related languages, translation to a low-resource language cannot be simply solved by using a multilingual model.

3 Summary & Conclusion

3.1 Summary of the techniques

This paper has shown several techniques for improving machine translation for low-resource languages. Transfer learning is a technique of incorporating weights from models that have been trained for a similar or related task. This has been shown to improve translation substantially when the parallel data is limited. Lexical proximity has a large effect on the extent to which this technique is influential. Another benefit of this approach is that any organization undertaking this task could likely utilize publicly available translation models for high-resource languages. Additionally, it is important to note that transfer learning is a technique that is inherently difficult to combine with multilingual approaches, and due to the nature of those approaches, the benefits of combining these two techniques would be questionable.

Procedures leveraging monolingual data can be highly relevant to many use cases, as monolingual data is relatively straightforward to acquire for many languages. Additionally, for some languages, pre-trained monolingual language models are available, which can lessen the resource burden of the translation task whilst improving the quality of the produced translations. Techniques such as copying the monolingual data to augment the dataset can be used relatively easily and this procedure will work without a need for large changes in the architecture to encompass those techniques. Finally, it is important to mention the technique for aligning word embeddings purely using monolingual data, as the results produced by this technique are quite promising and could be very valuable in cases of rare words that never appear in a parallel dataset. Additionally, in less theoretical use cases, the alignment could be improved by the use of some parallel data.

The last set of discussed techniques was concerned with including additional language pairs in the training dataset. Although these procedures can produce impressive performance improvements, they require multiple parallel datasets, which may be difficult to acquire, and the effort of procuring them could potentially be better spent on producing a parallel dataset for the language pair in the task set out in the beginning, depending on the circumstances. Furthermore, the lexical proximity of the languages in the training set has a large influence on the quality of the outcomes when using this technique. Therefore those techniques can be mostly exploited in the cases of large organizations and for languages where large datasets of parallel data are widely available, such as in the case of the official languages of the European Union and the Europarl dataset (Koehn 2005).

3.2 Risk associated with some techniques

Many of the techniques shown in this review utilized languages outside of the original task, for example in the case of the study presented by Karakanta et al. (2018), the task of translating English into Belorussian was aided by the use of Russian. The choice to select Russian is quite obvious due to their lexical proximity to Belorussian but inherently will cause any of the produced models to be biased towards sometimes using words and/or grammatical rules from the high-resource language used. This is a very important consideration especially where preservation of a language is part of the motivation behind a translation model. The choice of the high-resource language used has to be considered in the context of the causes of the endangerment of that language. In the case of Belorussian, a large reason for its decline is the use of Russian in Belarus. Smolicz & Radzik (2004) show the extent of the issue by comparing the number of schools providing education in Russian or Belorussian in Belorus, and only

61.5% of schools in Belarus teach exclusively in Belorussian as of 2004, and in some parts of the country such as the capital city only 1.7% of schools teach exclusively in Belorussian, whilst a staggering 58.7% of schools teach exclusively in Russian. More instances of this kind exist, and those techniques that utilize another language should be used with caution, especially if their goal is helping in the efforts to preserve a language.

3.3 Future research

This paper has presented several potential future research opportunities for each of the techniques. Firstly in the case of word embedding alignment methodology presented by Conneau et al. (2018), there is great potential in exploring the possibility of using a different learning rate scheduler and experimenting with replacing the unsupervised validation methodology with a supervised validation set, potentially leading to better model architecture and hyperparameter choices. Gülçehre et al. (2015) research could be revisited with new language models that have been introduced since the year of study as the field of monolingual language models has seen great improvements. Additionally, a low-resource language pair could be used in potential experiments to see the impact in those cases. The research conducted by Karakanta et al. (2018) included a stage where the training data for transliteration was obtained from Wikipedia articles. This process could be trialed with the use of multilingual data from knowledge bases such as Wikidata.

3.4 Conclusion

This paper has reviewed several important procedures that can improve Neural Machine Translation. The methods have been described and where appropriate compared. The majority of the techniques shown in this paper yield substantial improvements to translation when used in isolation, and many of them could be combined producing potentially even better outcomes. Multilingual approaches and transfer learning are amongst the ones demonstrating the largest performance improvements, but they come with substantial drawbacks in cases where the data or models required for them might be difficult to acquire, and additionally, it is complicated to combine these techniques. On the other hand, monolingual approaches do not bear the same performance rewards, but they can be seen as more feasible and less resource-heavy procedures that also can yield notable improvement, whilst still allowing for the possibility of them being combined with other techniques presented here. Some techniques quantifiably may not appear as promising, but they are interesting use cases when taken into the context of other techniques, such as word embedding alignment or back translation by copying monolingual datasets. Crucially it is important to consider that many of the techniques may introduce undesirable properties to the resulting model, which could impact the perceived usefulness.

References

Cambridge Dictionary (2023).
URL: *https://dictionary.cambridge.org/*

Conneau, A., Lample, G., Ranzato, M., Denoyer, L. & Jégou, H. (2018), Word translation without parallel data, *in* 'Proceedings of the Sixth International Conference on Learning Representations'.
URL: *https://arxiv.org/abs/1710.04087*

Currey, A., Miceli Barone, A. V. & Heafield, K. (2017), Copied monolingual data improves low-resource neural machine translation, *in* O. Bojar, C. Buck, R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn & J. Kreutzer, eds, 'Proceedings of the Second Conference on Machine Translation', Association for Computational Linguistics, Copenhagen, Denmark, pp. 148–156.
URL: *https://aclanthology.org/W17-4715*

Dong, D., Wu, H., He, W., Yu, D. & Wang, H. (2015), Multi-task learning for multiple language translation, *in* C. Zong & M. Strube, eds, '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)', Association for Computational Linguistics, Beijing, China, pp. 1723–1732.
URL: *https://aclanthology.org/P15-1166*

Eberhard, D. M., Simons, G. F. & Fennig, C. D. (2023), 'Ethnologue: Languages of the world'.
URL: *http://www.ethnologue.com*

Edunov, S., Ott, M., Auli, M. & Grangier, D. (2018), Understanding back-translation at scale, *in* E. Riloff, D. Chiang, J. Hockenmaier & J. Tsujii, eds, 'Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing', Association for Computational Linguistics, Brussels, Belgium, pp. 489–500.
URL: *https://aclanthology.org/D18-1045*

Goodfellow, I., Bengio, Y. & Courville, A. (2016), *Deep Learning*, MIT Press. `http://www.deeplearningbook.org`.

Gu, J., Hassan, H., Devlin, J. & Li, V. O. (2018), Universal neural machine translation for extremely low resource languages, *in* M. Walker, H. Ji & A. Stent, eds, 'Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)', Association for Computational Linguistics, New Orleans, Louisiana, pp. 344–354.
URL: *https://aclanthology.org/N18-1032*

Gülçehre, Ç., Firat, O., Xu, K., Cho, K., Barrault, L., Lin, H., Bougares, F., Schwenk, H. & Bengio, Y. (2015), 'On using monolingual corpora in neural machine translation', *CoRR* **abs/1503.03535**.
URL: *http://arxiv.org/abs/1503.03535*

Harrison, K. D. (2007), *When Languages Die: The Extinction of the World's Languages and the Erosion of Human Knowledge*, Oxford University Press.
URL: *https://doi.org/10.1093/acprof:oso/9780195181920.001.0001*

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M. & Dean, J. (2017), 'Google's multilingual neural machine translation system: Enabling zero-shot translation', *Transactions of the Association for Computational Linguistics* **5**, 339–351.
URL: *https://aclanthology.org/Q17-1024*

Karakanta, A., Dehdari, J. & Genabith, J. (2018), 'Neural machine translation for low-resource languages without parallel corpora', *Machine Translation* **32**.

Kingma, D. & Ba, J. (2014), 'Adam: A method for stochastic optimization', *International Conference on Learning Representations* .

Koehn, P. (2005), Europarl: A parallel corpus for statistical machine translation, *in* 'Proceedings of Machine Translation Summit X: Papers', Phuket, Thailand, pp. 79–86.
URL: *https://aclanthology.org/2005.mtsummit-papers.11*

Koehn, P. & Knowles, R. (2017), Six challenges for neural machine translation, *in* T. Luong, A. Birch, G. Neubig & A. Finch, eds, 'Proceedings of the First Workshop on Neural Machine Translation', Association for Computational Linguistics, Vancouver, pp. 28–39.
URL: *https://aclanthology.org/W17-3204*

Nguyen, T. Q. & Chiang, D. (2017), Transfer learning across low-resource, related languages for neural machine translation, *in* G. Kondrak & T. Watanabe, eds, 'Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)', Asian Federation of Natural Language Processing, Taipei, Taiwan, pp. 296–301.
URL: *https://aclanthology.org/I17-2050*

OpenAI (2023), 'Chatgpt'. Accessed on December 2, 2023.
URL: *https://www.openai.com/chatgpt*

Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. (2002), Bleu: a method for automatic evaluation of machine translation, *in* P. Isabelle, E. Charniak & D. Lin, eds, 'Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics', Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pp. 311–318.
URL: *https://aclanthology.org/P02-1040*

Sennrich, R., Haddow, B. & Birch, A. (2016), Neural machine translation of rare words with subword units, *in* K. Erk & N. A. Smith, eds, 'Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)', Association for Computational Linguistics, Berlin, Germany, pp. 1715–1725.
URL: *https://aclanthology.org/P16-1162*

Smolicz, J. J. & Radzik, R. (2004), 'Belarusian as an endangered language: can the mother tongue of an independent state be made to die?', *International Journal of Educational Development* **24**(5), 511–528. Education in Transitional States.
URL: *https://www.sciencedirect.com/science/article/pii/S0738059303000725*

Torrey, L. & Shavlik, J. (2009), *Transfer Learning.*
URL: *https://ftp.cs.wisc.edu/machine-learning/shavlik-group/torrey.handbook09.pdf*

Zoph, B., Yuret, D., May, J. & Knight, K. (2016), Transfer learning for low-resource neural machine translation, *in* J. Su, K. Duh & X. Carreras, eds, 'Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing', Association for Computational Linguistics, Austin, Texas, pp. 1568–1575.
URL: *https://aclanthology.org/D16-1163*