Mitigating the Disparity of Machine Translation Quality for Low Resource Languages

Jeffrey T. Miller
mill7707@morris.umn.edu
Division of Science and Mathematics
University of Minnesota, Morris
Morris, Minnesota, USA

Abstract

This paper focuses on the Machine Translation quality of Low Resource Languages (LRL) in contrast to that of High Resource Languages (HRL), both of which are loosely categorized by the size of their respective corpora (data sets of sentence pairs and individual text). This paper also surveys techniques used for improving machine translation for LRL, those of which include Back Translation and Data Augmentation. Neural Machine Translation (NMT) architectures for supervised, semi supervised, and unsupervised models and their respective techniques are examined as well.

Keywords: Machine Translation, Low Resource Languages, High Resource Languages, Natural Language Processing

1 Introduction

There are currently over 7,000 languages spoken on Earth as of 2023 [10]. This variance of spoken languages necessitates the application of automatic translation technology to facilitate international commerce, migration, communication, and diplomacy. For multilingual countries like India and Indonesia, the proliferation of fast, economical, and accurate Machine Translation (MT) has profoundly deep cultural and economic implications as it could further erase the language barriers between separate cultural identities.

Ideally, the quality of this technology would be similar regardless of the language pairs that are being used for the translation model. However, due to social and economic reasons, a majority of world languages are inhibited by suboptimal automatic translation.

This is the quality disparity issue of Low Resource and High Resource Languages (LRL & HRL). There is no concrete methodology that exists to categorize languages into HRLs and LRLs. However, Table 1 provides a good general guide to how languages resource levels are distinguished. Statistical Machine Translation (SMT), not covered extensively in this paper, and Neural Machine Translation (NMT), explained in Section 2.1, require massive data sets of parallel corpora to be trained on to support the complexities of these languages' grammatical structures and the probabilistic models of SMT and NMT.

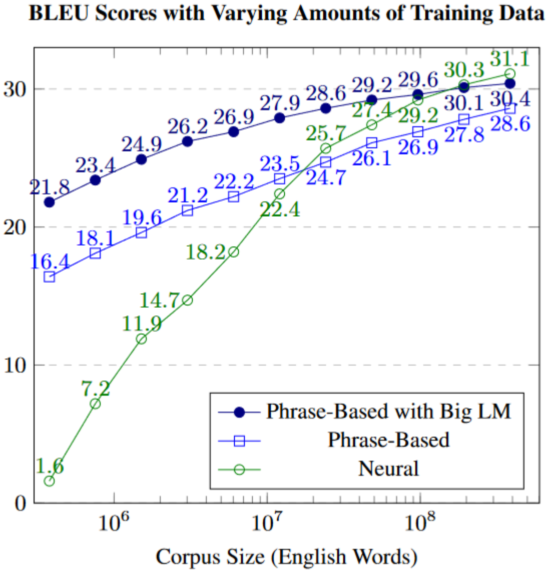


Figure 1. The proportional relationship between BLEU scores and training data [7]

Parallel data is labeled data that consists of novels, social media, television, and radio broadcasts that constitute authentic text for a language's corpus [11] and are aggregated as sentence pairings within a corpus. This is different from monolingual data which, as the name implies, is unlabeled data from a particular language (an untranslated social media post in Tamil or a text document containing War and Peace in its original Russian version). Figure 1 illustrates how the amount of training data available for language pairing schemes can influence the accuracy of the subsequent translation. The amount of training data used for NMT and for encoding words in each language (known as Word Embeddings) show how that data can influence translation quality (measured by a standard of translation quality known as the BLEU score [2])

The translation quality for a given pair of languages is proportional to the size of the parallel corpora for that pair. Western European languages, such as French and English, possess a considerably larger corpora for the use of training than African languages, such as Swahili, regardless of the

Class	Description	Examples	# Langs
0	Have exceptionally limited resources, and have rarely been considered in language technologies.	Solvene, Sinhala	2,191
1	Have some unlabeled data; however, collecting labeled data is challenging	Nepali, Telugu	222
2	A small set of labeled datasets has been collected, and language support communities are there to support the language.	Zulu, Irish	19
3	Has a strong web presence, and a cultural community that backs it. Have been highly benefited by unsupervised pre-training	Afrikaans, Urdu	28
4	Have a large amount of unlabeled data, and lesser, but still a significant amount of labeled data. Have dedicated NLP communities researching these languages.	Russian, Hindi	18
5	Have a dominant online presence. There have been massive investments in the development of resources and technologies	English, Japanese	7

Table 1. Language classifications[6]

number of native speakers. For example, the most prominent publicly available English-Spanish parallel corpus from the online collection OPUS ParaCrawl v9 has 5.0G English tokens (words in the English context) and 5.4G Spanish tokens, whereas the most prominent publicly available English-Swahili parallel corpus WikiMatrix has only 4.4M English tokens to 1.0G Swahili tokens [1].

The purpose of this paper is to provide an overview of NMT systems as well as survey the practiced techniques for improving the translation quality of LRL. In Section 2, the subject of NMT systems is elaborated upon in supervised, semi-supervised, and unsupervised contexts, the section also provides a cursory explanation for how Artificial Neural Networks (ANNs) and the encoder-decoder system works. In Section 3, the LRL techniques of Data Augmentation and Transfer Learning are explained. Section 4 will cover how these techniques affect the quality of translation. Lastly, Section 5 will cover the conclusion of this survey.

2 Background

As of 2013, NMT has become the standard translation model for Machine Translation (MT), replacing SMT [10]. This section covers the structure of the NMT translation model for supervised, semi supervised, and unsupervised architectures as well as provides a cursory understanding as to how Neural Networks operate. It also discusses instances when the NMT model is multilingual, instead of bilingual.

NMT models can be placed into three categories: supervised, semi-supervised, and unsupervised. The distinctions of these categories is dependent on the size of the data set for the language pair. The supervised section will cover NMT for HRLs; and the semi supervised and unsupervised sections will cover NMT for LRLs.

2.1 ANNs, RNNs, and the Encoder-Decoder System

To provide an explanation for how NMT works, Artificial Neural Networks (ANNs) themselves must be described in some detail. ANNs, as the name implies, are systems of interconnected nodes inspired by the advanced biological neural networks that exist in our brains. The basic structure can be represented in Figure 3. Neurons in the human brain are represented here with nodes that will activate and trigger more nodes depending on whether a certain threshold is met with the connections of all the nodes of the first layer to nodes of the second layer. This process repeats through an arbitrary number of "hidden layers" until it reaches the output layer, where it should be representative of our desired result. For example, if a picture of a human face were fed into an ANN, then the ANN would be able to to consistently produce an integer that denotes that person's age. If the result is not correct, then the weights (connections) within the ANN are adjusted/tweaked to provide a correct output for what was inputted. This is the reason why parallel corpora are crucial to the development of an efficient NMT, the bilingual/multilingual data can serve to moderate the internal thresholds of the ANN until it is finely tuned to the two languages of the corpus. To reference the previous example, training data containing one million images of human faces, that are labeled with the person's respective age, can be fed through an ANN, concurrently. When this training stage has been completed, we should have an ANN that possess parameters (weights) that are sophisticated enough to produce an accurate age for an image introduced outside of the training dataset.

To understand how this fits into the context of Natural Language Processing (NLP), it helps to break the sentence structure into its basic components, tokens (words) and grammar (context). Words can be processed into small vector representations known as Word Embeddings through an ANN called an auto-encoder. In order for NLP to work within the

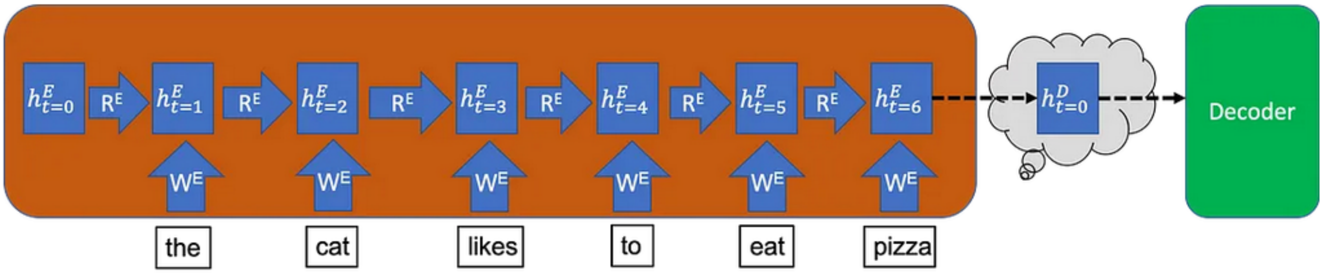


Figure 2. Diagram of an encoder [8]

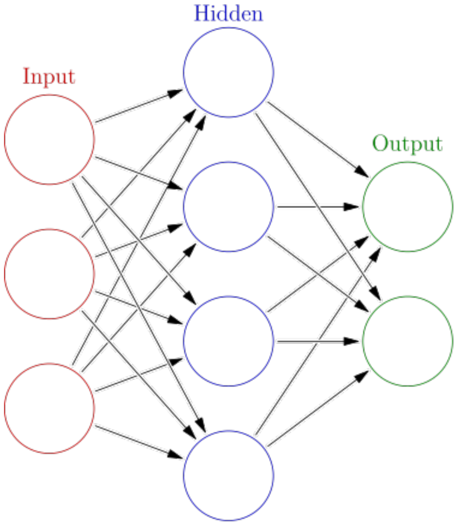


Figure 3. An ANN [3]

context of an ANN there needs to be a way to represent words numerically so that they can be fed into an ANN. To do this, each word in the input sentence is represented as an array of zeros and with a one at the index that corresponds to the word that index represents in a vocabulary Table [8]. The auto-encoder's function is to compress these words in such a way that allows them to be restored/decompressed into their original form. The training for machine translation is essentially for the purpose of creating ANNs that can accurately translate words to vectors and vectors to words. In the case of Bilingual Machine Translation, instead of the Word Embedding being translated back into its original language, it is translated to a target language using a decompressor or "decoder" trained in this target language.

This brings us to the encoder-decoder architecture that is the standard for NMT. In Figure 2, we can see an sequence of English words being compressed into Word Embeddings and then being stored within a "hidden vector". The vertical arrows represent the Word Embeddings and the horizontal arrows represent each loop in the Recurrent Neural Network (RNN) sequence. After looping through each ANN in the sequence, the result is finally represented in the output layer with a vector of fixed length and then is fed again through an decoder which is similar in structure to the encoder but works backwards (fixed vector to target language sentence). When this vector-decoded sentence is compared with a human translated version of the input sentence, the weights of the encoder-decoder ANNs can be subsequently adjusted to better emulate the human translation. Training data is important for this because it can improve the accuracy of the ANN model by influencing the weight of the different connections to produce a more accurate output.

This is how most NMT systems are generally constructed. That being said, what is most important is understanding how each operates with different ranges of parallel and monolingual data.

2.2 Supervised NMT

Supervised NMT is a probabilistic model much like SMT, attributing a weighted score to each subsequent translated word to guess the correct sequence. But unlike the SMT, it can observe the entire input sentence, word-to-word, as opposed to breaking down into phrases. In this sense I mean that SMT looks at the probability of a word being represented as a part of a sentence in a particular order. As supervised NMT models are dependent on comparatively large data sets, it precludes the application of it to LRL. What will be discussed in the following sections is what techniques are employed in semi-supervised and unsupervised scenarios.

2.3 Semi supervised NMT

Semi supervised NMT models differ from Supervised NMT in that they do not possess an abundance of parallel data to be trained on, but do contain a considerable amount of monolingual data. Thus the methods used for improving their translation quality rely on the availability of monolingual data from both languages. The main technique for using this data includes using a language model for the decoder-side of the architecture.

A language model in the context of NMT is essentially a scoring system that is implemented on the decoding side of the architecture. While the translation (NMT) model is trained on what parallel data exists, the language model is trained on the monolingual data of the target language for the decoder. There are two types of language model implementation: shallow fusion and deep fusion. As the names imply, they are related to how fundamentally they impact the NMT structure.

In shallow fusion the language model is not a part of the architecture of the decoder model. It simply takes each word after it has been converted from its vector form by the TM ANN and then assigns priority based on its probabilistic model to a list of potential words that fit better in the target language. It is only used in training and is not looped into the decoder RNN. With deep fusion, the LM is integrated into the decoder with a RNN and is used in every subsequent loop, or with every word embedding in the sequence that is being processed.

The drawbacks to using a language model is that both the NMT and language model need to be separately trained [10], which could potentially cause a disparity within the NMT model to form, like an under-trained language model assigning a poor score to a highly trained NMT's proposed word.

2.4 Unsupervised NMT

Generally, in the context of machine learning algorithms, "Unsupervised" means the algorithm is learning from unlabeled data sets. In this context, however, unsupervised NMT are exemplified by the lack of monolingual/parallel data for both languages. Unsupervised NMT architectures make use of Generative Adversarial Networks (GANs), defined later in this section, to "bridge the gap" between the monolingual corpora of two languages. This is achieved through initialization, back translation, and the discriminative classification.

The first part of this processes involves initialization, or creating a language map that overlays two languages over each other, to put it simply. This is accomplished through word-embedding schemes. Word embeddings are essentially vector representations of words mapped to a vector space and each language has a vector space with word embeddings in it. The logic is that the same context must exist for all languages as we exist in the same physical world. For example, the word "tiger" must have a similar embedding space containing the adjacent words "striped", "predator", or "feline", in the same area relative to "tiger", no matter what language. Inversely, if words such as "large", "gray", and "animal" are all adjacent to one another in the vector space, the word "elephant" must be adjacent as well. Figure 5 helps illustrate how this process works, with English being mapped over French in the same common space.

This initialization is then fed through an auto-encoder. This process continually translates the source language to the target language, and then the target language into the source to reduce the amount of noise between translation. A similar example would be to take an English sentence and repeatedly translate back-and-forth to French until the results are consistent; this is for the purpose of generating more training data. Given that these two languages share a vector space, translated sentences can be reconstructed by the model regardless of the presence of substandard machine translated sentences [10].

Finally, the system is wrapped in a generator-discriminator frame. The way GANs work is that the generator creates plausible "fake" translations from the source-target data and the discriminator is tasked with detecting these fake translations. Eventually, through the iterative process, the generator will become so adapt at creating translations that the discriminator will be unable to differentiate between the samples and the "fakes".

2.5 Multilingual NMT

These are NMT systems that are designed around multiple language pairs, as opposed to two. Google translate is an example of a prominent multilingual NMT system. Multilingual NMT (MNMT) have shown to perform better than their bilingual counterparts [10] due to their ability to analyze the shared relationship between multiple language groups and the much larger data sets that those language groups need.

3 LRL Techniques

This section covers two methods for improving the accuracy of NMT for LRL pairs: Data Augmentation, which does not require changing the internal architecture of the NMT model, and Transfer Learning, which does.

3.1 Data Augmentation

Data augmentation is a technique used for various NMT architectures that synthetically increases the monolingual and bilingual corpora for a given language to offer better training to the NMT models. This technique, in and of itself, does not influence the internal structure of the translation model. There are three methods that are employed for this technique: word replacement based augmentation, back translation based augmentation, and parallel corpus mining.

3.1.1 Word Replacement. Word replacement is a method that involves taking a sample collection of sentences from a monolingual corpus and syntactically replacing the words and phrases of that sentence to artificially generate more sentences for the corpus. An example of this would be to replace the word "quick" with its synonym "nimble" in the phrase: "The quick, brown fox jumps over the lazy dog". We could also swap the adjectives so that the sentence becomes "The lazy fox jumps over the quick, brown dog". This method can be accomplished through manual translation or through training a language model, similar to the fusion method,

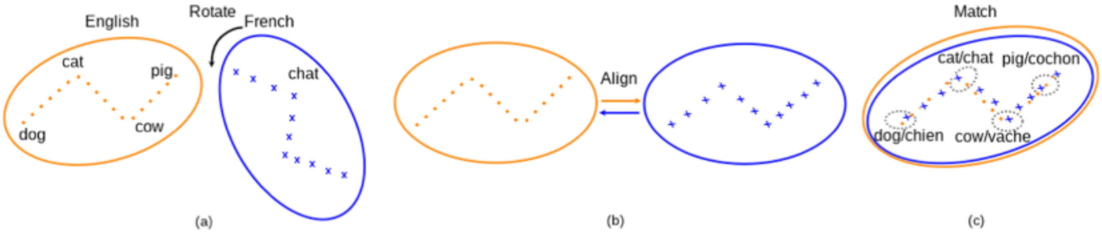


Figure 4. Diagram of a data initialization [10]

either with deep or shallow fusion, seen in Section 2.3. The amount of sentences that can be generated from the method can help pad the monolingual corpus of languages in zero-shot scenarios (where there is no training data of any kind).

3.1.2 Back Translation. Back Translation is a method of Data Augmentation that involves taking a target language monolingual corpus and using a Machine Translation model to translate the corpus into the source language for the creation of a synthetic parallel corpus. The resulting corpus is then subsequently filtered of all synthetic sentences containing noise (sub-standard automatic translations) to improve the training of the NMT system. This is done through a function that compares the original source text to the text that is "double-translated". There is an issue with this method, however, as it assumes that there is already an MT system that exists between the pair.

3.1.3 Parallel Corpus Mining. Parallel Corpus Mining is the method of using comparable corpora to increase the size of a parallel corpus. An example of this would be the use of the same Wikipedia article in different languages for the use of sentence-pairing. A prominent technique for mining this information is through the use of multilingual sentence embedding schemes. In these schemes, a similar encoder-decoder matrix, much like the one described in Section 2.1, is used to create multilingual sentences embeddings to be used in extraction processes. Sentence embeddings are ways in which a sentence can be translated to a vector that an encoder or decoder can readily process. The extracted sentences are then appraised of their similarity to each other and stored in the corpus. This approach is limited however if the target language is underrepresented in the pre-trained model.

3.2 Transfer Learning

A somewhat adhoc approach to solving the issue of poor machine translation is the method of pivoting. Pivoting is the simple method of using an intermediary, high-resource language to improve the translation quality between two languages. An example of this is would be to translate from Hindi to English to Spanish, as opposed to just Hindi to Spanish. While this method can work, it has a potential to propagate translation errors between models and as both models need to be trained, the time complexity increases [10].

Transfer Learning (TL) is a process that seeks to avoid these issues through using a "parent" NMT model to train a "child" NMT model by using the common, applicable knowledge that the parent gained through learning. The amount of parallel data the child model has determines whether the transfer process will be a warm-start (sufficient training data) or a cold start (little training data). The cold start is generally preferred [10] as it is more true to a real-life scenario. Figure 5 shows a simplified illustration of the process, with a source/target parent languages being used in the training of source/target child languages. When recalling, in the previous chapter, how an NMT model can be trained, we can see how a parent NMT's parameters can be transferred to a child model. There are multiple methods for how a parent model's parameters can be transferred to a child model. The first involves a training a model with exclusively HRL pairs, then fine-tuning it with a parent-child corpus, and then finally only using child data.

Even though this method is generally successful, the parent side often needs to be fine-tuned before the child can utilize it. If the parent is not trained, then this is called "freezing" the parent [10]. There are numerous approaches to freezing that can entail freezing none, a part, or all of the hidden layers of a parent NMT to get the desired result [10].

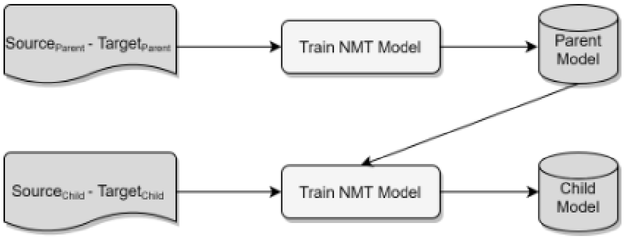

$\text{Source}_{\text{Parent}}$ - $\text{Target}_{\text{Parent}}$
Train NMT Model
Parent Model
$\text{Source}_{\text{Child}}$ - $\text{Target}_{\text{Child}}$
Train NMT Model
Child Model

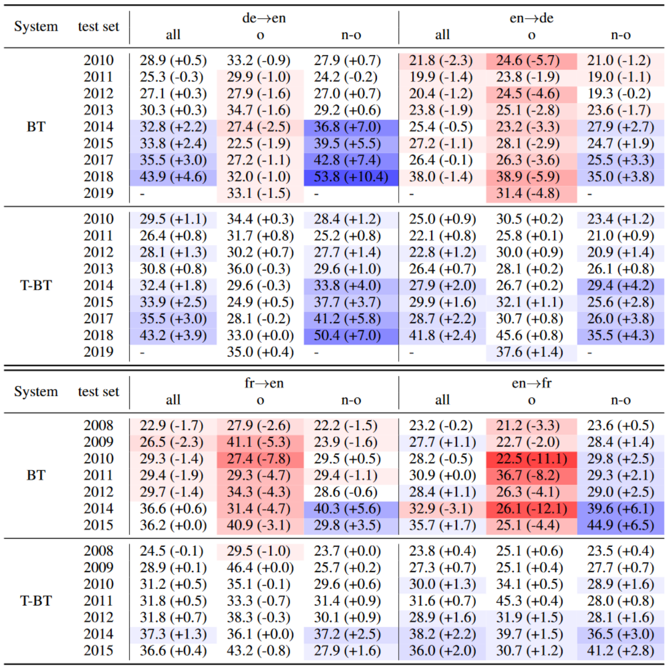
Figure 5. Diagram of a simple transfer protocol [10]

4 Application/Results

To demonstrate the effectiveness of LRL techniques, Figure 2 shows an example of an application of using a Back Translation (BT) Data Augmentation method for a significant improvement to the initial BLEU scores, which measure translation accuracy. The methodology that is being tested here is the effectiveness of using BT data with noise and using BT data that tags the potentially noisy data. The data below shows the increase the BLEU scores, which are a metric for determining the accuracy for machine translations, for both English-French and English-German translations in comparison to the original model. While this testing uses HRLs such and English and French the researchers deliberately choose to limit the corpus of the two to 200K words for the purpose of modeling a low resource environment [9]. Non-tagged BT systems are shown to have no notable impact on the BLEU scores. However, when a special tag is added to back-translated data (T-BT) so that the model can differentiate, there is a noticeable increase in score with the 2018 German to English test set producing a 10.4 increase in the number of BLEU points.

The two rows in Figure 6 denoted by BT (untagged) and T-BT (tagged) represent the impact on BLEU scores in their data sets when they were applied. The numbers in parentheses shows the difference in BLEU score from not using the BT at all and the blue/red color coding indicates the improvement/deterioration of BLEU scores, respectively. The columns "de ⟶ en" and "en ⟶ de" represent the translation task. The sub-columns are the different types of data that were being used for training. The column "o" denotes training data that is from articles, movies, TV shows, and other media that extracted for translation. The column "n-o" is training data that has been hand-translated for the express purpose of back-translation, and is generally simpler in syntactic structure. The column "all" represent the average improvement in BLEU scores of the two. The same applies to the lower table, which shows the translation task between French and English. This table suggests to us that tagging back-translated data prevents the deterioration that is seen in the "o" column of both language pairs.

This shows the merit of tagged back translation as a method that can improve the translation quality of NMT. Here are other papers that provide results that show the effectiveness of other techniques: Language Models [5] and Transfer Learning [4].



System	test set	de→en			en→de		
		all	o	n-o	all	o	n-o
BT	2010	28.9 (+0.5)	33.2 (-0.9)	27.9 (+0.7)	21.8 (-2.3)	24.6 (-5.7)	21.0 (-1.2)
	2011	25.3 (-0.3)	29.9 (-1.0)	24.2 (-0.2)	19.9 (-1.4)	23.8 (-1.9)	19.0 (-1.1)
	2012	27.1 (+0.3)	27.9 (-1.6)	27.0 (+0.7)	20.4 (-1.2)	24.5 (-4.6)	19.3 (-0.2)
	2013	30.3 (+0.3)	34.7 (-1.6)	29.2 (+0.6)	23.8 (-1.9)	25.1 (-2.8)	23.6 (-1.7)
	2014	32.8 (+2.2)	27.4 (-2.5)	36.8 (+7.0)	25.4 (-0.5)	23.2 (-3.3)	27.9 (+2.7)
	2015	33.8 (+2.4)	22.5 (-1.9)	39.5 (+5.5)	27.2 (-1.1)	28.1 (-2.9)	24.7 (+1.9)
	2017	35.5 (+3.0)	27.2 (-1.1)	42.8 (+7.4)	26.4 (-0.1)	26.3 (-3.6)	25.5 (+3.3)
	2018	43.9 (+4.6)	32.0 (-1.0)	53.8 (+10.4)	38.0 (-1.4)	38.9 (-5.9)	35.0 (+3.8)
	2019	-	33.1 (-1.5)	-	-	31.4 (-4.8)	-
T-BT	2010	29.5 (+1.1)	34.4 (+0.3)	28.4 (+1.2)	25.0 (+0.9)	30.5 (+0.2)	23.4 (+1.2)
	2011	26.4 (+0.8)	31.7 (+0.8)	25.2 (+0.8)	22.1 (+0.8)	25.8 (+0.1)	21.0 (+0.9)
	2012	28.1 (+1.3)	30.2 (+0.7)	27.7 (+1.4)	22.8 (+1.2)	30.0 (+0.9)	20.9 (+1.4)
	2013	30.8 (+0.8)	36.0 (-0.3)	29.6 (+1.0)	26.4 (+0.7)	28.1 (+0.2)	26.1 (+0.8)
	2014	32.4 (+1.8)	29.6 (-0.3)	33.8 (+4.0)	27.9 (+2.0)	26.7 (+0.2)	29.4 (+4.2)
	2015	33.9 (+2.5)	24.9 (+0.5)	37.7 (+3.7)	29.9 (+1.6)	32.1 (+1.1)	25.6 (+2.8)
	2017	35.5 (+3.0)	28.1 (-0.2)	41.2 (+5.8)	28.7 (+2.2)	30.7 (+0.8)	26.0 (+3.8)
	2018	43.2 (+3.9)	33.0 (+0.0)	50.4 (+7.0)	41.8 (+2.4)	45.6 (+0.8)	35.5 (+4.3)
	2019	-	35.0 (+0.4)	-	-	37.6 (+1.4)	-

System	test set	fr→en			en→fr		
		all	o	n-o	all	o	n-o
BT	2008	22.9 (-1.7)	27.9 (-2.6)	22.2 (-1.5)	23.2 (-0.2)	21.2 (-3.3)	23.6 (+0.5)
	2009	26.5 (-2.3)	41.1 (-5.3)	23.9 (-1.6)	27.7 (+1.1)	22.7 (-2.0)	28.4 (+1.4)
	2010	29.3 (-1.4)	27.4 (-7.8)	29.5 (+0.5)	28.2 (-0.5)	22.5 (-11.1)	29.8 (+2.5)
	2011	29.4 (-1.9)	29.3 (-4.7)	29.4 (-1.1)	30.9 (+0.0)	36.7 (-8.2)	29.3 (+2.1)
	2012	29.7 (-1.4)	34.3 (-4.3)	28.6 (-0.6)	28.4 (+1.1)	26.3 (-4.1)	29.0 (+2.5)
	2014	36.6 (+0.6)	31.4 (-4.7)	40.3 (+5.6)	32.9 (-3.1)	26.1 (-12.1)	39.6 (+6.1)
	2015	36.2 (+0.0)	40.9 (-3.1)	29.8 (+3.5)	35.7 (+1.7)	25.1 (-4.4)	44.9 (+6.5)
T-BT	2008	24.5 (-0.1)	29.5 (-1.0)	23.7 (+0.0)	23.8 (+0.4)	25.1 (+0.6)	23.5 (+0.4)
	2009	28.9 (+0.1)	46.4 (+0.0)	25.7 (+0.2)	27.3 (+0.7)	25.1 (+0.4)	27.7 (+0.7)
	2010	31.2 (+0.5)	35.1 (-0.1)	29.6 (+0.6)	30.0 (+1.3)	34.1 (+0.5)	28.9 (+1.6)
	2011	31.8 (+0.5)	33.3 (-0.7)	31.4 (+0.9)	31.6 (+0.7)	45.3 (+0.4)	28.0 (+0.8)
	2012	31.8 (+0.7)	38.3 (-0.3)	30.1 (+0.9)	28.9 (+1.6)	31.9 (+1.5)	28.1 (+1.6)
	2014	37.3 (+1.3)	36.1 (+0.0)	37.2 (+2.5)	38.2 (+2.2)	39.7 (+1.5)	36.5 (+3.0)
	2015	36.6 (+0.4)	43.2 (-0.8)	27.9 (+1.6)	36.0 (+2.0)	30.7 (+1.2)	41.2 (+2.8)

Figure 6. Tagged and Back-Tagged Data [9]

5 Conclusion

Differences in translation quality are still prevalent in NLP technology even as it continues to play a larger role in international communication. However, this paper has surveyed there are a few prominent methods of translation to bridge this gap. Data Augmentation, a method that influences the data directly, and Transfer Learning, a method that involves manipulation of an NMT internal thresholds. As LRL-NMT has made such considerable advancements, the key issue currently is understanding what NMT technique is well-fitted for a particular data setup. Nevertheless, the ever-growing increase in research publications for LRL-NMT [10] show that there is good reason to be optimistic for better translation standards for those who speak under-represented languages.

Acknowledgments

I would like to thank both my advisor Elena Machkasova and Peter Dolan for their guidance during the drafting of this paper.

References

[1] [n. d.]. Opus. https://opus.nlpl.eu/

[2] [n. d.]. Wikipedia. https://en.wikipedia.org/wiki/BLEU

[3] [n. d.]. Wikipedia. https://en.wikipedia.org/wiki/Machine_learning

[4] Raj Dabre, Atsushi Fujita, and Chenhui Chu. 2019. Exploiting Multilingualism through Multistage Fine-Tuning for Low-Resource Neural Machine Translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 1410–1416. https://doi.org/10.18653/v1/D19-1146

[5] Çaglar Gülçehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loïc Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. On Using Monolingual Corpora in Neural Machine Translation. *CoRR* abs/1503.03535 (2015). arXiv:1503.03535 http://arxiv.org/abs/1503.03535

[6] Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting*

of the Association for Computational Linguistics. Association for Computational Linguistics, Online, 6282–6293. https://doi.org/10.18653/v1/2020.acl-main.560

[7] Philipp Koehn and Rebecca Knowles. 2017. Six Challenges for Neural Machine Translation. In *Proceedings of the First Workshop on Neural Machine Translation*. Association for Computational Linguistics, Vancouver, 28–39. https://doi.org/10.18653/v1/W17-3204

[8] Quinn Lanners. 2019. Website Title. https://towardsdatascience.com/neural-machine-translation-15ecf6b0b

[9] Benjamin Marie, Raphael Rubino, and Atsushi Fujita. 2020. Tagged Back-translation Revisited: Why Does It Really Work?. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 5990–5997. https://doi.org/10.18653/v1/2020.acl-main.532

[10] Surangika Ranathunga. 2022. Neural Machine Translation for Low Resource languages: A Survey. *Comput. Surveys* 55 (2022), 1–37. Issue 11. https://doi.org/10.1145/3567592

[11] Rita Tse, Silvia Mirri, Su-Kit Tang, Giovanni Pau, and Paola Salomoni. 2020. Building an Italian-Chinese Parallel Corpus for Machine Translation from the Web. In *Proceedings of the 6th EAI International Conference on Smart Objects and Technologies for Social Good* (Antwerp, Belgium) *(GoodTechs '20)*. Association for Computing Machinery, New York, NY, USA, 265–268. https://doi.org/10.1145/3411170.3411258