Tackling neural machine translation in low-resource settings: a Portuguese case study

Arthur T. Estrella[1], **João B. O. Souza Filho**[2]

[1] Electrical Engineering Program (PEE/COPPE), Federal University of Rio de Janeiro
PO Box 68504, RJ 21941-972, Brazil

atelles@coppe.ufrj.br, jbfilho@poli.ufrj.br

Abstract. *Neural machine translation (NMT) nowadays requires an increasing amount of data and computational power, so succeeding in this task with limited data and using a single GPU might be challenging. Strategies such as the use of pre-trained word embeddings, subword embeddings, and data augmentation solutions can potentially address some issues faced in low-resource experimental settings, but their impact on the quality of translations is unclear. This work evaluates some of these strategies on two low-resource experiments beyond just reporting BLEU: errors are categorized on the Portuguese-English pair with the help of a translator, considering semantic and syntactic aspects. The BPE subword approach has shown to be the most effective solution, allowing a BLEU increase of 59% p.p. compared to the standard Transformer.*

1. Introduction

Since the rise of the Neural Machine Translation (NMT) branch, many solutions solely focused on surpassing the state-of-the-art, ignoring the associated computational burden. Thus, most models have been progressively adopting deeper architectures, hugely increasing the number of network parameters and, as a result, the dependence on more extensive datasets. The excessive focus on boosting performance regardless of complexity deviated the researchers from a more profound criticism over how such architectures address the translation task, if more cost-effective models can be proposed, and how to better cope with translation errors.

Low-resource NMT domains are defined as practical development scenarios wherein the GPU memory and the amount of data available to train some model are limited. Some techniques can potentially help under those circumstances, as the prior initialization of neural network embedding weights [Qi et al. 2018] with pre-trained word embeddings, the production of embeddings at a subword level [Sennrich et al. 2015b] or data augmentation with a monolingual dataset [Sennrich et al. 2015a] (also known as back-translation).

To the best of our knowledge, experimental studies discussing and evaluating the cost-effectiveness of strategies aiming to circumvent the practical issues faced with low-resource domains, especially considering the English-to-Portuguese pair, are missing in the literature. Many previous works only focused on optimizing metrics such as BLEU. Despite its usefulness, this index is limited due to only accounting for matches of a fixed number of n-grams, penalizing correct but different lexical translations.

This work[1] uses Transformers [Vaswani et al. 2017] to experimentally evaluate the impact of strategies such as transfer learning (by the use of pre-trained word embeddings), subword modelling, and data augmentation in the translation quality. It considers only one average size GPU and small to medium sized datasets (low-resource), focusing on the English-to-Portuguese pair. Additionally, a qualitative analysis of the translation errors is derived over a sample of sentences by a native translator, considering a multi-dimensional criterion, aiming to evaluate models' performance on a broader scope than BLEU.

This paper is structured as follows: Section II provides a brief coverage of the Transformer architecture, and Section III discusses the main issues to be tackled in low-resource domains, along with some potential strategies that can be exploited in such cases. Section IV provides a brief description of the datasets considered in this work and depicts quantitative and qualitative results for the strategies here considered. Finally, Section VI poses the conclusions and next steps.

2. The Transformer Architecture

Transformers refers to a branch of algorithms based on the seminal work of [Vaswani et al. 2017], representing the state-of-art. Differently from the sequential processing inherent to the Recurrent Neural Networks (RNN) adopted in previous NMT models, the Transformer model processes large sentences in parallel, establishing a richer set of interrelations between source and target sentence words, thus leading to a better inference of the words' context and a higher performance with long sentences. The reader is referred to the original paper for more details about this architecture.

3. Tackling low-resource settings

Low-resource constraints refer to limitations on dataset quality/size and computational resources available. Small datasets can be strongly biased in specific contexts, which may induce the predictions produced by the decoder model to move away from the reference or even turn the training process unstable, reducing the final model performance. In turn, computational aspects are primarily related to the number of GPUs available and their standalone memory. Memory constraints directly affect the definition of training and model hyperparameters, such as the batch size, the number of hidden layers, the embedding size, and the size of the attention mechanisms. To avoid an experiment failure due to an out of memory error, one should first consider the largest sentence size, a common batch size limiting factor. Too small batches may lead to unstable and biased training, increasing the epoch time and resulting in sub-optimal translation quality. Moreover, the estimation a priori of a minimum quantity of sentences for an adequate translation is also a challenging task, severely depending on the complexity of the application domain. Hopefully, the following strategies may mitigate the need for abundant data and GPU memory in the translation task:

3.1. Transfer learning methods

Transfer learning exploits other application parameters as initial values for the NMT model training (warm-start). A typical example refers to using pre-trained word embed-

[1]ACKNOWLEDGMENT - This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil(CAPES) – Finance Code 001.

dings in the embedding layer of neural models instead of the default random parameters' initialization. This procedure accelerates and improves training since these embeddings already carry out some semantic word meanings to be subsequently refined to a particular translation task (fine-tuning).

The use of pre-trained words can be pretty effective in low-resource scenarios, as pointed in the analysis conducted in [Qi et al. 2018]. According to the authors, there is a "sweet spot" for dataset size, according to which this strategy is more effective. Similarly, the impact over more related language pairs is often higher, such as Spanish and Portuguese.

3.2. Subword methods

Subword embeddings represent a useful strategy to reduce the out of vocabulary (OOV) occurrences. The central idea is decomposing words into sub-parts (character groups), which are common to many words, turning the model less susceptible to the vocabulary content and its size. This technique, referred to as BPE (Byte Pair Encoding) - a compression algorithm, was introduced by [Sennrich et al. 2015b]. Roughly, BPE brokes the words in a corpus into smaller parts (the small BPE unit is a character); some of them subsequently merged, with the number of merge operations being the main hyperparameter to be tuned. The BPE drawback is the impossibility of defining a maximum vocabulary size a priori. A clear advantage of moving from word-level to subword NMT using BPE is reported in [Sennrich and Zhang 2019]: an increase of BLEU score from 7.2 to 16.6 in an ultra-low-resource setting as well as a consistent rise in the BLEU values for a wide range of application scenarios.

Pre-trained word embeddings also followed the subword trend with the proposition of the Fast Text algorithm [Bojanowski et al. 2016]. This technique treats words as a bag of character n-grams and adds tokens to distinguish among prefixes, suffixes, and other character sequences. In practice, each word is assigned to a given number of n-grams, typically $3 \leq n < 6$. Finally, the word is represented by the sum of the vector representations (embeddings) associated with the n-grams composing it.

3.3. Data Augmentation

Large-scale parallel corpora is not a common resource for most existing language pairs, unlike monolingual corpora. However, is it possible to exploit the abundant and large monolingual datasets widely available these days for data augmentation? The answer is yes and this technique is referred to as Back-Translation (BT) [Sennrich et al. 2015a]. The idea is quite simple: let us consider the development of a translation model from a language A to B. New pairs of sentences can be simply synthesized by training an inverse model, i.e., a translator from the language B to A, with the same sentence pairs. Once completed this training, this auxiliary model can be fed with corpora from a similar or a different context domain to produce new sentence pairs. BT has shown to be a simple but effective method to address low data availability in many domains, as shown in [Poncelas et al. 2018], often increasing translation performance.

4. Experimental Setup

Datasets with low to medium complexity levels were carefully picked for the proposed analysis: Tatoeba [Tiedemann 2020] and TED Talks [Cettolo et al. 2012]. Both repre-

sent a low-resource scenario due to the scarce number of sentences, in alignment with some references [Sennrich and Zhang 2019] [Zoph et al. 2016]. News Commentary v16 [Tiedemann 2012] is a different domain monolingual dataset used for BT and includes a rich range of sentences in terms of content and complexity.

The reduced Tatoeba dataset contains 143.8k small and basic to intermediate English level sentences, posing a low complexity challenge for the NMT task. It includes 26.3k unique words in PT and 15.3k in EN. TED Talks is a medium-size database covering a range of subjects, including from low to high complex sentences.

In the experiments, 10% of Tatoeba was held out for testing, while the remaining data was split into 10% for validation and 90% for training, using a seed equal 0. Despite TED disposing predefined training, test, and validation sets, the original validation set is too small (906 sentences), leading us to move the last 20 talks (2081 sentences) from the training set to this set. As a result, the training set contains 236.1k (1918 talks) sentences, randomly sampled to defining training batches using a seed equal to 157, and the test set includes 11.4k sentences. Additionally, all text was pre-processed to eliminate all XML enclosed sentences and tags, except for the ones related to title and description.

The experiments were performed on a single GPU, using Google Colaboratory and Kaggle infrastructure. Typically such environments dispose of NVIDIA GPUs like Tesla P100, Tesla K80 or Tesla T4, with a GPU memory ranging from 12GB to 16GB.

5. Results

Regarding the Transformers, the parameters adopted were $d_{model} = 256$, $d_{ff} = 256$, 8 attention heads, and Q, K and V square matrices with dimension 64. The pre-trained Fast Text-based models, which employed embeddings described in [Hartmann et al. 2017], are exceptions, considering $d_{model} = 300$ and 6 attention heads. All variants adopted the Adam optimizer with $\beta \in (0.9, 0.98)$ and $\epsilon = 10^{-8}$, a learning rate of 10^{-4} and the beam search considered a beam with size 3. The early stopping criterion was based on the validation perplexity behaviour for ten epochs, halting the training in case of performance stagnation.

Sacrebleu [Post 2018] and NLTK [Loper and Bird 2002] are two BLEU variants used for assessing the performance of the models. The major difference between them resides in a stronger Sacrebleu's penalization over cases where the translated and reference sentences differ in length.

5.1. Effects of restricting dataset content

To shed light on the possible effects of limited data on the performance of NMT models, we considered a hypothetical experimental scenario where only a fraction of TED and Tatoeba training sets were used in training, according to the following percentages: 33.3%, 50%, 66.6%, 83.3%. Table 1 summarizes the results.

Results show that the Sacrebleu scores for Tatoeba were about twice the achieved with TED, corroborating with the much higher complexity of the latter. The BLEU metrics for both datasets have shown a monotonic behaviour, with exceptions to TED in two cases: Sacrebleu (100% $\times$ 83.3%) and NLTK (100% $\times$ 83.3 and 83.3% $\times$ 66.6 %). The reasons for such findings may include: (1) the possible use of synonyms in the translations, an aspect ignored by any BLEU metric; (2) a higher incidence of repetition errors

Table 1. Data augmentation scores

Fraction of the Dataset	Tatoeba				TED			
	Sacrebleu	NLTK BLEU	Batch Size	Epochs	Sacrebleu	NLTK BLEU	Batch Size	Epochs
33.3%	48.64	67.09	512	76	24.7	57.65	30	40
50%	52.53	70.12	512	65	25.18	56.46	30	40
66.6%	55.3	72.12	512	58	26.22	**56.81**	29	36
83.3%	56.24	73.18	512	58	**26.74**	56.57	28	30
100%	**57.99**	**74.07**	512	58	25.24	55.36	28	30

due to data quality issues (to be discussed further in the following section); (3) the more complex and richer TED content, which might have led to a wider subject coverage in the training set, reducing model accuracy, a hypothesis deserving a future investigation. Finally, models developed with a fraction of the original training datasets (66.6%) performed surprisingly well.

5.2. Effects of transfer learning and subword embeddings strategies

Aiming to evaluate the leveraging effects of pre-trained Fast Text and BPE [Sennrich et al. 2015b] strategies in low-resource NMT tasks, BPE models were implemented in the Texar framework [Hu et al. 2019] (PyTorch version). In contrast, the alternative models considered a customized PyTorch [Paszke et al. 2019] solution. Table 2 exhibits these results, reproducing the last line of Table 1 to allow an easier comparison of the results.

Table 2. Transfer learning and subword embeddings translation results

Technique applied	Tatoeba				TED			
	Sacrebleu	NLTK BLEU	Batch Size	Epochs	Sacrebleu	NLTK BLEU	Batch Size	Epochs
None	57.99	74.07	512	58	25.24	55.36	28	30
Fast text	56.96	69.91	512	50	24.07	51.69	30	45
Subword BPE	**66.63**	**83.02**	512	40	**40.26**	**72.20**	32	40

Curiously, the use of Fast Text embeddings is associated with an unexpected performance drop for both datasets. Conversely, the gains observed with BPE, which also exploits word embeddings, were impressive. One hypothesis for the bad Fast Text performance is a possible overspecialization to other text domains, since it was produced with content mined by a crawler [Hartmann et al. 2017]. The higher BPE gain in TED (15.02) compared to Tatoeba (8.64) signalizes the effectiveness of BPE in dealing with more complex NMT scenarios, especially regarding a more diverse vocabulary, avoiding OOV occurrences.

5.3. Effects of the Back-Translation (BT) strategy

The BT experiments were restricted to the TED dataset. Data augmentation was performed with synthetic sentences produced with the own TED (using its left out sentences) and with the News dataset. These experiments aimed to verify if data augmentation can result in higher BLEU scores under low-resource constraints.

A single EN-PT Transformer was trained with the entire TED dataset to generate the synthetic sentences, reaching 27.73 and 63.8 points for the Sacrebleu and NLTK, respectively. The subset of back-translated sequences appended to the training sets was randomly sampled using the following seeds: 157 (TED) and 0 (News).

Table 3. TED Talks back-translation results

Technique applied	Batch size	Epochs Trained	Sacrebleu	NLTK BLEU
None (Original TED)	30	27	25.24	55.36
Reduction of TED to 50%	40	30	25.18	**56.46**
BT (50% of News synthetic examples)	34	33	21.80	51.34
BT (50% of TED synthetic examples)	34	28	**25.95**	56.42
Reduction of TED to 66.6%	36	29	26.22	56.81
BT (33.3% of News synthetic examples)	34	27	24.12	53.77
BT (33.3% of TED synthetic examples)	34	27	**27.54**	**58.95**
Reduction of TED to 83.3%	28	30	26.74	56.57
BT (16.6% of News synthetic examples)	34	29	31.28	63.30
BT (16.6% of TED synthetic examples)	34	27	**34.62**	**64.61**

Table 3 exhibits the results. For a more severe restriction on the dataset size (50%), using other domain synthesized sentences is harmful to model performance, while own-domain synthesis resulted in a marginally better BLEU score. However, for a lower percentage of synthetic data, positive effects start to appear. Considering an intermediate restriction ($\approx 33\%$), using the same domain sentences in back-translation led to a mild increase in both BLEU values compared to the Original TED, signalizing that such "noisy" sentences may contribute to increasing translation quality. Finally, considering a small restriction ($\approx 16.6\%$), both domain approaches are quite effective, resulting in models that largely surpasses the model developed over original data.

5.4. Subjective evaluation

This analysis focused on two dimensions: sentence complexity and error patterns. Random samples were selected from TED, analysed by a human translator, and stratified according to the CEFR scale [Council]. Due to dataset characteristics, this study was restricted to sentences classified as A1, A2, B1 and B2. Ten sentences from each level were presented to two models: the Transformer trained over a fraction of 66% of original data and the BPE variant developed over the entire dataset. The idea here was twofold: first, evaluate the effects of restricting the dataset size over the error patterns; and second, assessing qualitatively the translations produced by the model of best performance, and thus the impact of eliminating <unk> occurrences, generating custom words, and switching from word to subword level.

Regarding the identification of error patterns, a multidimensional evaluation in eight categories was considered: similar word choice, omission, out of context, verb tense, sentence choice (the translation is OK, but the outcome is entirely different from the reference), insertion, repetition and <unk> errors [2]. Table 4 shows the number of errors committed by each model, stratified by sentence complexity and error category. Considering the limitations of such analysis, such as the reduced sample and the analysis of

[2] A detailed error description and some evaluation samples can be found at `https://github.com/Art31/pt-nmt-low-resource.git`.

only one translator, both models performed quite similarly regarding the ”similar word choice” occurrence. Nonetheless, the BPE produced fewer errors related to ”omission” (levels A1 and B1), ”sentence choice” (B1), ”insertion” (A1, A2 e B1), ”repetitions” (all), and ”<unk> errors” (all), performing worse regarding ”out of context” and ”verb tense”.

Table 4. Class-error ratios per dataset and sentence complexity.

Model Name	Comple-xity	Similar word choice	Omis-sion	Out of context	Verb tense	Sentence choice	Inser-tion	Repeti-tion	<unk> error
66% TED	A1	$^{2}/_{10}$	$^{3}/_{10}$	$^{0}/_{10}$	$^{0}/_{10}$	$^{1}/_{10}$	**$^{6}/_{10}$**	$^{3}/_{10}$	$^{2}/_{10}$
	A2	$^{7}/_{10}$	$^{6}/_{10}$	$^{1}/_{10}$	$^{1}/_{10}$	$^{3}/_{10}$	$^{4}/_{10}$	$^{2}/_{10}$	$^{2}/_{10}$
	B1	$^{7}/_{10}$	$^{5}/_{10}$	**$^{3}/_{10}$**	$^{3}/_{10}$	$^{3}/_{10}$	$^{2}/_{10}$	$^{2}/_{10}$	$^{3}/_{10}$
	B2	**$^{8}/_{10}$**	$^{7}/_{10}$	**$^{2}/_{10}$**	**$^{7}/_{10}$**	**$^{5}/_{10}$**	$^{5}/_{10}$	**$^{6}/_{10}$**	**$^{5}/_{10}$**
	Average	60.0%	52.5%	15.0%	27.5%	30.0%	**42.5%**	**32.5%**	**30.0%**
BPE	A1	$^{2}/_{10}$	$^{0}/_{10}$	$^{0}/_{10}$	$^{1}/_{10}$	$^{1}/_{10}$	$^{2}/_{10}$	$^{0}/_{10}$	$^{0}/_{10}$
	A2	$^{7}/_{10}$	**$^{6}/_{10}$**	**$^{3}/_{10}$**	$^{4}/_{10}$	**$^{3}/_{10}$**	$^{1}/_{10}$	$^{0}/_{10}$	$^{0}/_{10}$
	B1	$^{4}/_{10}$	$^{3}/_{10}$	$^{1}/_{10}$	$^{3}/_{10}$	$^{1}/_{10}$	$^{0}/_{10}$	$^{0}/_{10}$	$^{0}/_{10}$
	B2	**$^{8}/_{10}$**	$^{5}/_{10}$	**$^{3}/_{10}$**	**$^{5}/_{10}$**	$^{2}/_{10}$	**$^{6}/_{10}$**	**$^{1}/_{10}$**	$^{0}/_{10}$
	Average	52.5%	35.0%	17.5%	32.5%	17.5%	**22.5%**	**2.5%**	**0.0%**

Results from Table 4 underwent a Multiple Fisher test to evaluate if the differences observed between the error ratios of the two models are statistically significant. This analysis considered multiple 2x2 tables (one to each class of error), with rows defining the model and columns associated with the occurrence or not of some class of error. The significance level was set to 5%; thus, the null hypothesis was rejected whenever the p-value was lower than 0.05, representing a statistically significant difference. This analysis concluded that the ”repetition error” ($p = 0.0002$), the ”<unk> error” ($p = 0.0001$) and the ”insertion error” ($p = 0.0001$) are indeed less frequent in BPE than 66% TED.

6. Conclusion

This paper focused on dealing with low-resource NMT scenarios, considering low and medium complexity Portuguese-English datasets (TED and Tatoeba). It experimentally evaluated the impact of transfer learning (pre-trained word embeddings), subword embeddings (BPE), and Back-Translation strategies (using the same and different domains data) over BLEU performance. In addition, this work presented a qualitative analysis conducted by a human translator over the outcomes of some best performing models, considering a specifically designed multidimensional evaluation criteria, for a sample constituted by a total of 40 sentences, equally stratified in four CEFR levels.

The BPE was the most effective technique for dealing with a low-resource setting, attaining the highest BLEU values and the lower error rates in six from eight error categories defined by the qualitative analysis. Same domain data augmentation has also led to exciting results when synthesising only a small portion of the original training set (16.6%).

Future works include evaluating models exploiting both BPE and BT and considering more sentences, as well as CEFR levels, in the qualitative analysis, possibly bringing a clearer view of error patterns and enlighting the practical effects of each strategy in objective and subjective translation quality aspects.

References

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching word vectors with subword information. *CoRR*, abs/1607.04606.

Cettolo, M., Girardi, C., and Federico, M. (2012). WIT3: Web inventory of transcribed and translated talks. *Proc. of EAMT*, pages 261–268.

Council, E. The CEFR levels - council of europe (coe). `https://tinyurl.com/cefrlcoe`. Accessed: 2021-08-12.

Hartmann, N., Fonseca, E. R., Shulby, C., et al. (2017). Portuguese word embeddings: Evaluating on word analogies and natural language tasks. *CoRR*, abs/1708.06025.

Hu, Z., Shi, H., Tan, B., et al. (2019). Texar: A modularized, versatile, and extensible toolkit for text generation. In *ACL 2019, System Demonstrations*.

Loper, E. and Bird, S. (2002). Nltk: The natural language toolkit. In *Proc. of the ACL Workshop on Effective Tools for Teaching Natural Language Processing.*

Paszke, A., Gross, S., Massa, F., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Proc. Systems 32*, pages 8024–8035. Curran Associates, Inc.

Poncelas, A., Shterionov, D. S., Way, A., et al. (2018). Investigating Back-translation in neural machine translation. *CoRR*, abs/1804.06189.

Post, M. (2018). A call for clarity in reporting BLEU scores. In *Proc. of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Assoc. for Comp. Linguistics.

Qi, Y., Sachan, D., Felix, M., et al. (2018). When and why are pre-trained word embeddings useful for neural machine translation? pages 529–535, New Orleans, Louisiana. Assoc. for Comp. Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2015a). Improving neural machine translation models with monolingual data. *CoRR*, abs/1511.06709.

Sennrich, R., Haddow, B., and Birch, A. (2015b). Neural machine translation of rare words with subword units. *CoRR*, abs/1508.07909.

Sennrich, R. and Zhang, B. (2019). Revisiting low-resource neural machine translation: A case study. In *Proc. of the 57th Annual Meeting of the Assoc. for Comp. Linguistics*, pages 211–221, Florence, Italy. Assoc. for Comp. Linguistics.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In *Proc. of the Eight International Conf. on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Assoc. (ELRA).

Tiedemann, J. (2020). The Tatoeba translation challenge - realistic data sets for low resource and multilingual MT. *CoRR*, abs/2010.06354.

Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. *CoRR*, abs/1706.03762.

Zoph, B., Yuret, D., May, J., and Knight, K. (2016). Transfer learning for low-resource neural machine translation. In *Proc. of the 2016 Conf. on Empirical Methods in Natural Language*, pages 1568–1575, Austin, Texas. Assoc. for Comp. Linguistics.