Human Evaluation of the Usefulness of Fine-Tuned English Translators for the Guarani Mbya and Nheengatu Indigenous Languages

Claudio Pinhanez, Paulo Cavalin, Julio Nogima
IBM Research, Brazil
{csantosp,pcavalin,jnogima}@br.ibm.com

Abstract

We investigate how useful are machine translators based on the fine-tuning of LLMs with very small amounts of training data, typical of extremely low-resource languages such as Indigenous languages. We started by developing translators for the Guarani Mbya and Nheengatu languages by fine-tuning a WMT-19 German-English translator. We then performed a human evaluation of the usefulness of the results of test sets and compared them to their SacreBLUE scores. We had a level of alignment around 60-70%, although there were about 40% of very wrong translations. The results suggest the need of a filter for bad translations as a way to make the translators useful, possibly only in scenarios of human-AI collaboration such as writing-support assistants.

1 Introduction

In this paper we present a human evaluation of the usefulness of machine translation (MT) models which we trained to translate sentences from two Brazilian Indigenous Languages (BILs), i.e. Guarani Mbya and Nheengatu, to English. The main goal was to evaluate the end-user usefulness of the MT models based on fine-tuning a pre-trained Transformer-based language model, aka Large Language Model (LLM) (Devlin et al., 2019; Raffel et al., 2020), in the case of extremely low-resource languages.

Our method consisted of fine-tuning the WMT19 model (Ng et al., 2019), trained to translate German sentences to English, with both parallel corpora and language resources, to each of the BILs. Since for both Guarani Mbya and Nheengatu data is quite scarce, we relied on resources such as dictionaries (or lexicons) and educational documents to extract as much parallel data as possible in the training set and to compile a set of parallel sentences for testing. We then measured the performance of the models with SacreBLEU, which is the implementation of the BLEU score (Post, 2018).

It is very difficult to draw conclusions about human usefulness of a translator based only on values from automatic metrics such as SacreBLEU, since they are based on easy-to-perform computations such as word comparison, ignoring often semantic issues. Therefore, to determine the usefulness of the translators, we conducted a human evaluation on the texts generated from the test set inputs. Our analysis consisted of labeling each of the generated outputs in a seven-point scale, ranging to near-perfect quality to very wrong translations.

Results showed that translations were good for only 18% of the Guarani Mbya outputs and 32% for the Nheengatu outputs. Considering content which could be utilized by an user, the results were 35% and 42%, respectively. However, 40% and 42% of the translations were considered very wrong. These results suggest that such translators are more likely to be useful in scenarios of direct human-machine collaboration, such as writing assistants, than of standalone automatic translation. We then compared the human-based usefulness results with the SacreBLEU scores, finding alignments of about 60%.

We believe this work contributes to the understanding of how traditional translation metrics relate to actual end-user usefulness. It also highlights the need of care to use such metrics as system evaluation tools.

2 Datasets

We created two datasets, one for each BIL.

2.1 Guarani Mbya dataset

Sentences from three different sources were used in the construction of the *Guarani Mbya* dataset. The first source was a set of Guarani Mbya short stories with 1,022 sentences, available in both Portuguese and English (Dooley, 1988a,b). The second comprises 245 texts extracted from PDF files with a pedagogical character (Dooley, 1985). The third source was Robert A. Dooley's *Lexical Guarani Mbya Dictionary* (Dooley, 2016), a reference work for the language, from which we extracted 2,230 sentence pairs. The last two sources contained sentence pairs in Guarani Mbya and Portuguese only. We converted them to English using a Portuguese-to-English commercial translation service. We have permission from the author to use this data.

After concatenating the data from the three sources, we cleaned it, removing some non-alphanumeric characters (e.g. *, ≫, •) and normalizing Unicode values. Then, the dataset was split into training and test sets and finalized by removing repeated sentences and cross-contamination between sets, totaling 3,155 and 300 sentence pairs, respectively.

2.2 Nheengatu dataset

The *Nheengatu* dataset used five different sources containing Nheengatu sentences with Portuguese translations. As with the Guarani Mbya dataset, we converted the Portuguese sentences to English using a Portuguese-to-English commercial translation service[1].

The first source is the *Nheengatu lexicon* (Ávila, 2021) with 6846 sentences extracted from the lexicon examples. For that, we processed the original file made available by the author. The second one is *Corpus Lições* (Ávila, 2021), containing 1,665 samples already available in a spreadsheet format. The other sources, which were directly extracted from PDFs, were: *Texto Anônimo* (Navarro, 2011), with 427 samples; *Brilhos na Floresta* (Ishikawa, 2019), with 590 samples; and *Curso LGA* (Navarro, 2016), with a partial extract of 590 samples.

The Nheengatu dataset contains 7,281 samples, with a random split of 241 samples (10% of the data from all sources except Nheengatu lexicon) for testing and 6,804 samples for training.

3 Machine Translation Models

We trained two models by fine-tuning a pre-trained Transformer-based Language Model to translate from Guarani Mbya and Nheengatu to English. That was done by fine-tuning the parameters of the `WMT19` model (Ng et al., 2019), a 315M-parameter German-to-English machine translator pre-trained with about 28M pairs of translated sentences and more than 500M back-translated sentences. Both models were fine-tuned for 10 epochs, considering a batch size of 32, learning rate of 2.10^{-5} decaying to 2.10^{-6} according to a cosine function.

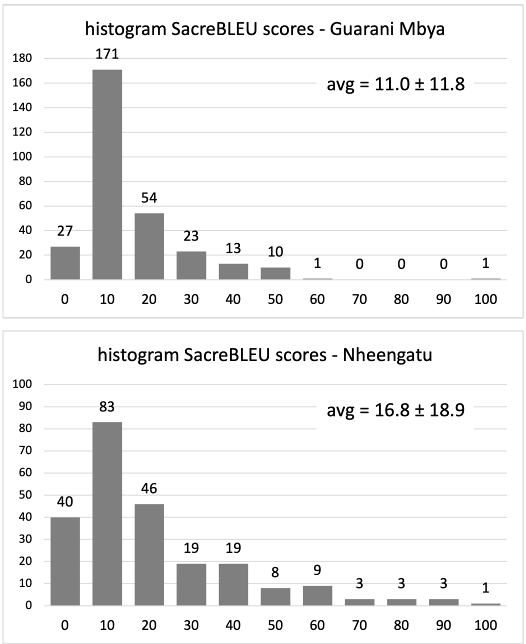


Figure 1: Histograms of the SacreBLEU scores of the Guarani Mbya and Nheengatu translators.

4 SacreBLEU Evaluation

To evaluate the results we relied on the the *SacreBLEU* metric which is the BLEU score computed with the SacreBLEU Python package (Post, 2018). We computed `sentence-level` scores and considered the average of those scores for system-level evaluation.

We observe slightly higher scores from the Nheengatu translator, with a SacreBLEU score of 16.8 against 11.0 from Guarani Mbya. We see, however, that the Nheengatu model resulted in higher standard deviation, with 18.9 against 11.8 of the Guarani Mbya model, which made us question the distribution of the data and compute the histograms of the scores for each test set which are shown in Figure 1. What we observe is a skinnier distribution for Guarani Mbya which may explain the higher standard deviation of Nheengatu.

5 Human Evaluation of Usefulness

In this section we present the results of a human-based evaluation of all the test set outputs of the translators which we conducted to understand the

[1] IBM Watson Language Translation v9.0.0

usefulness of the sentences generated by them. We had two goals in doing a comprehensive manual evaluation of the translations: first, to help us to determine how far the translators are from an actual deployment; and second, to understand how much standard ML metrics can be relied on as a predictor of success in actual human tasks.

The evaluation was performed by one of the authors of the paper by comparing, for all sentences of the test sets, the translation to English from the test set to the generated output. The evaluator did not know both languages but had access to the original text in the Indigenous languages for inspection purposes. Through a process similar to what is used for *thematic networks* (Attride-Stirling, 2001), the categories and their meanings were developed by an iterative process of evaluating sentences, refining the categories, and re-evaluating the sentences until saturation was reached. From that point on, all entries were then evaluated. This process led to the following categories and labels of the usefulness of the translations:

***very wrong*:** the output was completely unrelated to the expected translation or had gross mistakes such as repetitions, words from the source language, or it was empty;

***incorrect*:** no blatant mistakes but there was no relation with the expected text;

***mostly incorrect*:** one or two correct words but mostly of the rest was useless;

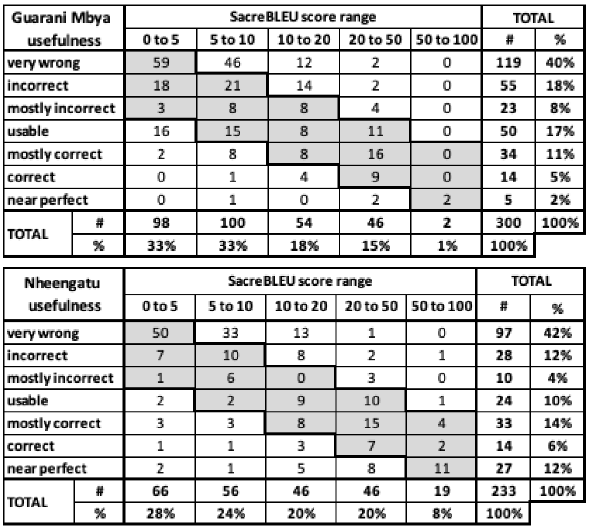
***usable*:** the output could be used as a starting point for a translation because it had two or three correct words or it resembled the structure of the expected sentence;

***mostly correct*:** at least two thirds of the generated text were correct but it still had mistakes which needed human correction;

***correct*:** the generated text was an acceptable translation of the original sentence although it could fail to capture completely the meaning of the expected text;

***near perfect*:** the output was almost a literal repetition of the expected text.

The rightmost columns of Table 1 depict the number of sentences evaluated into those different categories for the 300 outputs of the Guarani Mbya test set and the 233 of the Nheengatu test set. For the Guarani Mbya translator, we see about 40% of all outputs are in the *very wrong* category and 26% in the *incorrect* and *mostly incorrect* categories. Of the remaining 34%, about 28% are sentences which need some level of human intervention to be used (categories *usable* and *mostly correct*) and only 7% would be suitable in an automatic translation scenario. The numbers of the Nheengatu translator are better, with 42% in the *very wrong* category but only 16% in the *incorrect* and *mostly incorrect* categories. Of the remaining 42%, 24% would need human correction to be usable and 18% would be suitable for an automatic translation scenario.



Guarani Mbya usefulness		SacreBLEU score range					TOTAL	
		0 to 5	5 to 10	10 to 20	20 to 50	50 to 100	#	%
very wrong		59	46	12	2	0	119	40%
incorrect		18	21	14	2	0	55	18%
mostly incorrect		3	8	8	4	0	23	8%
usable		16	15	8	11	0	50	17%
mostly correct		2	8	8	16	0	34	11%
correct		0	1	4	9	0	14	5%
near perfect		0	1	0	2	2	5	2%
TOTAL	#	98	100	54	46	2	300	100%
	%	33%	33%	18%	15%	1%	100%	

Nheengatu usefulness		SacreBLEU score range					TOTAL	
		0 to 5	5 to 10	10 to 20	20 to 50	50 to 100	#	%
very wrong		50	33	13	1	0	97	42%
incorrect		7	10	8	2	1	28	12%
mostly incorrect		1	6	0	3	0	10	4%
usable		2	2	9	10	1	24	10%
mostly correct		3	3	8	15	4	33	14%
correct		1	1	3	7	2	14	6%
near perfect		2	1	5	8	11	27	12%
TOTAL	#	66	56	46	46	19	233	100%
	%	28%	24%	20%	20%	8%	100%	

Table 1: Results of the human evaluation of usefulness the Guarani Mbya and Nheengatu translators and their relationship with SacreBLEU scores (alignment regions marked with a grey background).

Next, we examined how the human evaluation of usefulness of the generated translations related to the SacreBLEU scores. Table 1 also depicts the number of sentences of each category in relation to 5 ranges of SacreBLEU scores, which follow a log-like distribution. If the two methods of evaluation were aligned, we would expect the majority of the sentences to be along the main diagonal of the tables. However, there is a good amount of spread and to quantify it we divided the table in two areas: the cells close to the main diagonal (depicted with a grey background on Table 1) and those in the left-bottom and right-top triangles.

In the results of the Guarani Mbya translator, the main diagonal contains 186 (62%) of all outputs while the non-aligned areas comprise 114 (38%). In the Nheengatu translator, there are 125 (71%) outputs on the main diagonal and 51 (29%) on the non-aligned areas. In general, in about one third of the cases, for both translators, the SacreBLEU score does not seem to be not a good predictor of

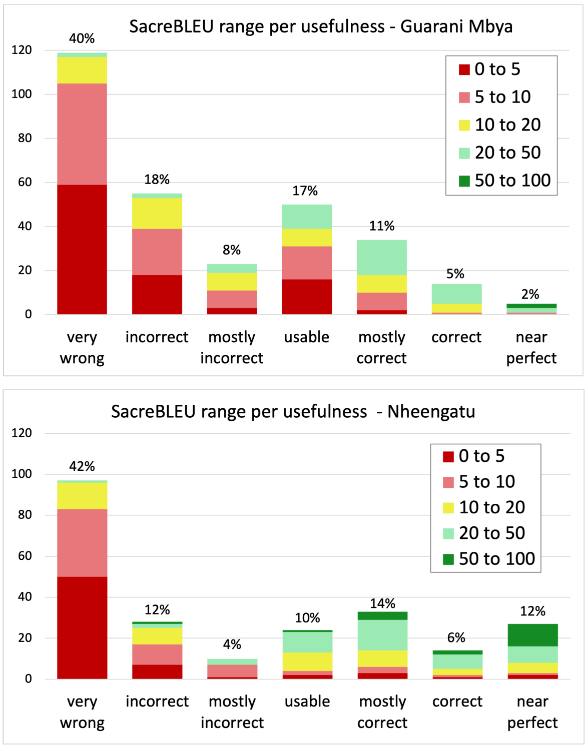


Figure 2: Distribution of SacreBLEU scores to each of the qualitative evaluation categories for the Guarani Mbya and Nheengatu translators.

the usefulness of a translation.

Figure 2 provides a visual rendition of the data on Table 1 which shows more clearly that the Nheengatu SacreBLEU scores seem to be better correlated with the usefulness evaluation than the scores of the Guarani Mbya translator.

6 Final Discussion

In this paper we explored two forms of evaluation of two translators from Guarani Mbya and Nheengatu languages to English. The first evaluation method, totally automatic, used the traditional SacreBLEU metric for translators, resulting in sentence averages of 11.0 ± 11.8 and 16.8 ± 18.9 respectively. The second form of evaluation was based on a human-created scale of usefulness established through an iterative process based on thematic networks. Results indicated that, for the Guarani Mbya translator, 40% of all generated sentences were totally useless, 26% had too many mistakes to be usable, about 28% could sometimes help knowledgeable end-users, and 7% were ready to be used. The Nheengatu translator had a better performance with 42% useless, 16% almost useless, 24% usable, and 18% with no errors. The two metrics had about 62% and 71% of alignment, respectively. These results seem to indicate that the translators, at this stage, can be only used in demos and initial prototypes.

It is very rare to find any kind of human-based usefulness testing of ML language systems, and even more in ML translator systems. We believe this kind of evaluation is particularly important in contexts of extremely low-resource languages such as the ones studied in this paper, since small amounts of data may impact the quality of traditional human-free ML metrics. Moreover, in this work we developed an evaluation which was specific to the task and based on the characteristics of the actual data, making it more ecologically valid. Unlike other works, we did not use human beings to validate a ML metric but instead we developed a more comprehensive metric which is directly related to the intended use. We see this as a major contribution of this work.

The work has important limitations which should be highlighted. First and foremost, only one human evaluator was used, a non-speaker of both languages. We plan to do, in future works, studies with multiple and language-knowledgeable evaluators to further validate our results. Another issue is that the Nheengatu translator was built with more than twice the number of training samples of the Guarani Mbya, what may explain its superior results.

Beyond those issues, the results of the human evaluation suggest more focused ways to improve the end-user performance which go beyond the traditional focus on simply increasing overall accuracy. In particular, around 40% of the outputs were very wrong, in a way that possibly they can be filtered out by a simple ML detector built directly with the data. Notice that, as shown in Table 1, only half of those outputs are easily detectable by the SacreBLEU score (0 to 5), and therefore a simplistic focus on improving the scores may not be enough to fix the problem.

Finally, we want to underscore the importance of ML developers to explore and have direct contact with the output data. During the evaluation process we could notice some other issues and errors which suggested readily available opportunities for improvement. This is an important benefit of human-based evaluations, which are often shun by developers as wasteful and time-consuming. Manually exploring and evaluating the output should be, in our opinion, a fundamental process in the construction of machine translation systems.

References

Jennifer Attride-Stirling. 2001. Thematic networks: an analytic tool for qualitative research. *Qualitative research*, 1(3):385–405.

J. Devlin, M. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of ACL'19*.

Robert Dooley. 1985. Nhanhembo'e aguã nhandeayvu py [1-5].

Robert A. Dooley. 1988a. Arquivo de textos indígenas – guaraní (dialeto mbyá) [1].

Robert A. Dooley. 1988b. Arquivo de textos indígenas – guaraní (dialeto mbyá) [2].

Robert A. Dooley. 2016. Léxico guarani, dialeto mbyá: Guarani-português.

Noemia Kazue Ishikawa. 2019. *Brilhos na Floresta*. Editora Valer; Editora Inpa, Manaus.

Eduardo de Almeida Navarro. 2011. Um texto anônimo, em língua geral amazônica, do século xviii. *Revista USP*, (90):181–192.

Eduardo de Almeida Navarro. 2016. *Curso de língua geral (nheengatu ou tupi moderno): a língua das origens da civilização amazônica*, 2 edition. Edição do Autor, São Paulo.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR's WMT19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319, Florence, Italy. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

M. T. Ávila. 2021. *Proposta de dicionário nheengatu-português*. Tese de doutorado, Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo. Recuperado em 2023-12-27, de `www.teses.usp.br`.