LLMs and Translation: different approaches to localization between Brazilian Portuguese and European Portuguese

Eduardo G. Cortes[1,2], Ana Luiza Trechel Vianna[1], Mikaela Luzia Martins[1], Sandro Rigo[1], and Rafael Kunst[1]

[1]UNISINOS, Av. Unisinos, São Leopoldo RS, Brazil
{egcortes,alvianna}@edu.unisinos.br {mikaelalm,rigo,rafaelkunst}@unisinos.br
[2]Institute of Informatics, UFRGS, Porto Alegre, Brazil

Abstract

The localization task consists of adapting linguistic and cultural material between different locales. For example, in European Portuguese (EP) the word "autocarro" is used to refer to "bus", while in Brazilian Portuguese (PB) the word "ônibus" is preferred. A precise localization can bring the communication between language variants closer, guaranteeing clear understanding among regional cultures that speak the same language. This study evaluates the effectiveness of Machine Translation approaches to localize sentences considering EP and BP. We assess the extent to which these models tend to paraphrase, quantifying the unnecessary changes made and evaluating the models with a human Multidimensional Quality Metrics (MQM) analysis. We applied a contrastive analysis of the two variants and chose four models (rule-based with a Masked Language Model, pre-trained neural machine translation (NMT), and two GPT-4-based models) to test and analyze. Our results show that the generative Large Language Models (LLMs) consistently delivered superior performance, underscoring their adeptness at grasping EP and BP nuances.

1 Introduction

The task of localizing texts between European Portuguese (EP) and Brazilian Portuguese (BP) holds significant importance. Given the expansive cultural and linguistic influence of both variants, accurate localization can bridge communication gaps, ensuring clarity and resonance with diverse audiences. Furthermore, as globalization intensifies, businesses, academia, and media increasingly seek to engage both European and Brazilian audiences without the expense and inefficiency of creating entirely separate content. However, applying computational models to the task of text localization between these two variants presents a unique challenge compared to conventional Machine Translation (MT). Although this type of localization may require fewer modifications, the choice to adapt or retain specific elements is influenced by context, formality levels, and cultural nuances distinct to each region.

Several models for localization between the two Portuguese variants have emerged over the years, as Ortega et al. (2022); Ruiz Costa-Jussà et al. (2018); Fancellu et al. (2014); Marujo et al. (2011), alongside MT models (Riley et al., 2023; Lakew et al., 2018; Koehn and Knowles, 2017) that perform translations from other languages into Portuguese. However, traditional approaches, such as fine-tuning pre-trained neural machine translation (NMT) models, still face challenges due to the lack of large collections of annotated and high-quality data. This interferes with the development of supervised models that seek to capture contextual nuances, such as formality and regional culture (Koehn and Knowles, 2017). Moreover, rule-based approaches struggle with the extensive load of lexical and grammatical changes, which are often dependent on these contextual nuances. Recently, generative Large Language Models (LLMs) have shown promising results in the domain of MT (OpenAI, 2023; Anil et al., 2023). However, the findings are not yet definitive regarding whether LLMs' scalability and adaptability make them effective at handling the subtle differences between the two Portuguese variants for localization.

In the current literature, multiple datasets exist with paired examples of EP and BP (Tiedemann, 2012; Cettolo et al., 2012; Riley et al., 2023). While some datasets are large, the quality of the available data is often questionable. Specifically, many paired entries show inconsistencies, including added or omitted content. Additionally, the sentences haven't been converted from one Portuguese variant into another. Rather, separate translations of the original English sentences into either BP or EP were added. This leads to considerable adaptations between the paired Portuguese versions, in which many changes don't capture the subtle differences between the two variants. Therefore, during the initial phases of our study, we observed that the models that utilize these for training or as prompt examples tend to paraphrase the input sentences, rather than simply making the essential and appropriate changes for localization, as shown in Table 1.

In this paper, we address the challenge of localizing sentences between EP and BP by introducing and evaluating different approaches, which consider a comparative study of the differences between the two Portuguese variants. Our primary objective is to determine how effectively these models can localize paired sentences, focusing on the necessary modifications. Furthermore, we aim to assess the extent to which these models tend to paraphrase, quantifying the unnecessary changes made. We hypothesize that models that directly integrate information for localization will maintain their performance in making essential adaptations while reducing unnecessary alterations to the input sentence.

To conduct our experiments, we propose four distinct strategies: a rule-based model informed by our contrastive analysis, a pre-trained NMT model focusing on minimal differences between paired sentences, and two GPT-4 based methods, one leveraging localized sentence prompts and another integrating our contrastive findings directly into the prompt. Our experiments rely on the Benchmark FRMT dataset (Riley et al., 2023), comprising handpicked paired sentences from both EP and BP. Our experimental environment involves professional linguists specialized in the target Portuguese variants, manually evaluating the sentences localized by the models using the Multidimensional Quality Metrics (MQM) framework, in conjunction with the application of automatic metrics for evaluation.

In summary, the scientific contributions of this work are as follows: **1)** A contrastive analysis between EP and BP that identifies the fundamental differences and organizes them into three categories: gerund, pronoun placement, and lexical changes. **2)** The introduction of two new MT models incorporating information about the differences between the two variants. The first uses manual rules to identify patterns in the input text and uses a Masked Language Model (MLM) to help find suitable replacements. The second employs information in GPT-4's prompt related to the contrastive analysis of how to localize the input sentence. **3)** A manual evaluation performed by fluent speakers in the target variant presenting a human perspective on the efficacy of the evaluated models.

This paper is structured in six additional sections. The state of the art is summarized in Section 2. The Contrastive Analysis is present in Section 3. Section 4 introduces the tested models in detail. Section 5 describes our methodology. Section 6 presents the results and analysis. Finally, Section 7 highlights the conclusion and outlines future work.

2 Related Works

Although some traditional tasks in NLP are closer to mapping in EP and BP, such as rewriting, paraphrasing, and lexical substitution, we believe the mapping bears the most similarities with MT and localization since translation between variants requires broader changes than just terminology adjustments (Schäler, 2004; Bendi, 2020), even in the same language (Lopes and Costa, 2008). Furthermore, stylistic conventions and grammatical modifications, among other possibilities, occur in large chunks of texts depending on the context and syntactic structure.

Among the studies that combine NMT with Portuguese variant translations is Lakew et al. (2018), which investigated ways to approach NMT from English into four variant pairs, BP and EP among them. They conclude that the best performance is achieved by training multilingual NMT systems when it comes to the supervised regime. Ruiz Costa-Jussà et al. (2018) investigated the use of NMT techniques to translate directly between the EP and BP and they trained their NMT model using a parallel corpus of subtitles. When compared to an SMT model trained on the same data, their NMT model displayed a performance improvement when translating from both EP to BP and BP to EP. Prior to this, the only two studies concerned with the automatic MT between EP and PB were Marujo et al. (2011), which proposed a rule-based system, and Fancellu et al. (2014) which presented an SMT system trained on parallel data.

Yet, the fact that standard NMT models sometimes have difficulties translating culturally specific information (Yao et al., 2023) and rely on extensive data coverage also opened doors for exploration with LLMs. NMT systems usually overlook the differences between EP and BP. Currently, MT consists of LLMs that can also translate and, at the

	EP (source)	Cerca de 2 mil estudantes estudam em 93 programas de doutoramentos académico.
A	BP	Cerca de 2 mil estudantes estudam em 93 programas de **doutorados acadêmico**.
B	BP	**Aproximadamente** 2 mil **alunos estão inscritos** em 93 programas de **doutorados acadêmico**.
	English	Around 2 thousand students study in 93 academic doctoral programs.

Table 1: Examples of Localization from EP to BP. Blue text indicates essential adaptations and orange text represents optional modifications. Localization **A** demonstrates essential adaptations only, whereas Localization **B** incorporates both essential and optional modifications without altering the meaning.

moment, there is much research going on about this topic (Hendy et al., 2023; Chowdhery et al., 2022; Anil et al., 2023), which aligns with our work. When it comes to prompting LLMs for MT, some studies use sentences from translation memories in the prompt for few-shot learning. However, they selected only the sentences closest to the input sentence and pointed to using LLMs to generate this sterilized data (Lyu et al., 2023; Mu et al., 2023). In the case of translating between EP and BP, prompting seems like a good approach as it should contain fewer variations when compared to two different languages. It is possible that the entry sentence would be very close to the sentences sought from the bank. He et al. (2023) propose a method that offers keywords, topics, and demonstrations without using external knowledge, and the LLMs generate these resources. It has shown the best results compared with traditional fine-tuning NMT models(Liu et al., 2023). The relative position of the input sentence in the prompt and the task instruction is crucial and suggests that it should be allocated to the end, being placed after the input sentence. Studies attested that this strategy provides improvements across common sequence generation tasks, and it has been shown to lead to a higher attention ratio for instructions compared to the baseline (Chen et al., 2023; Liu et al., 2023). When it comes to the evaluation of these tasks, Raunak et al. (2023) investigated how LLM translations differ qualitatively from standard NMT systems and found that LLMs are less literal when translating out of English, especially when the sentences contain idiomatic expressions.

Regarding the Contrastive Analysis, research in this discipline seeks to establish differences and similarities between languages for different purposes. From a computational perspective, studies based on contrastive analysis are linked to second language teaching and learning (Berzak et al., 2015), natural language identification and machine learning (Wong and Dras, 2009; Otomo, 2004). Concerning the use of contrastive analysis for translation purposes, Bennett (2002) discusses how the use of contrastive analysis aimed at translation can help MT researchers, while Korzen and Gylling (2017) use contrastive analysis to work on textualization and textual structure in Italian and Danish. Considering what has already been found, we propose a contrastive analysis between BP and EP in order to map the differences and incorporate them in MT models.

3 Contrastive Analysis

In Linguistics, one way to study language is by comparing or contrasting two or more languages. From this perspective, Contrastive Analysis aims to contrast languages to analyze and establish the similarities or differences between them (Ke, 2019; Krzeszowski, 2011; James, 1980). This discipline is composed of two levels: theoretical and practical. The theoretical level seeks to find models or theoretical frameworks to compare and establish basic notions of similarity and equivalence between the languages. In this sense, it is assumed that there are universal features between languages, or within a pair of languages, and such universal categories are applied to specific linguistic systems. The practical level, on the other hand, aims to apply the findings of theoretical contrastive analyses to practical purposes, such as in second language teaching and learning, translation, terminology, and lexicography (Ke, 2019).

For this study, regarding the theoretical part, we analyze previous materials (Djajarahardja, 2020; Castilho, 2013; Hříbalová, 2010; KATO, 2006; Teyssier and Cunha, 1982; Aco, 2014) that focused on describing the differences and similarities between BP and EP and considerations related to the Portuguese Orthographic Agreement. The practical part is to establish sixteen categories related to the differences between the Portuguese variants, such as numerals, variable accentuation, verbs and prepositions, reflexive pronouns, double negation, contrastive case in noun complement, combinations with oblique pronouns, article omission, among others. Considering the formal language

Category	EP example	BP example	English
Gerund	A verdade é que **estás a vencer** na vida que tens.	A verdade é que **está vencendo** na vida que tens.	The truth is that you are winning in the life you have.
Pronoun Placement	E esse significado **deu-me** esperança.	E esse significado **me deu** esperança.	And that meaning gave me hope.
Lexical Changes	Eles saíram logo depois do **pequeno-almoço**.	Eles saíram logo depois do **café da manhã**.	They left right after breakfast.

Table 2: Examples of the difference between each category from the contrastive analysis. The words in blue are differences between the EP and PB variants.

register and if the category is mandatory and not just an optional change, for this study, we select the three main differences between them, which are: **Gerund**, **Pronoun Placement**, and **Lexical Changes**. Regarding the **Gerund** category, in BP, the gerund form is more used, that is, auxiliary verb + verb in the gerund. In EP, the gerund is not used, instead, the following structure is applied: auxiliary verb + preposition + infinitive verb. The **Pronoun Placement** category is related to the use of the pronouns next to the verb. In BP, proclisis is commonly used, that is, the pronoun goes before the verb, and, in EP, the pronoun goes after the verb (enclisis). However, it is important to mention that, in BP, enclisis is also used at the beginning of a sentence. The last category, **Lexical Changes**, is related to lexical differences between the Portuguese variants that are mandatory. Table 2 exemplifies the differences between each category presented.

4 Proposed Models

We proposed different models, mainly based on approaches found in the literature that claim abilities to provide few-shot or zero-shot controllable translations. Among these are two standard methods: a rule-based model and a pre-trained NMT model for localization. In addition to that, some models incorporate information from categories of the differences identified in the contrastive analysis.

4.1 Rule-based + MLM Model

This model is a rule-based approach combined with the Masked Language Model (MLM) Albertina PT-* fine-tuned for Portuguese variants (Rodrigues et al., 2023). Specifically, the MLM is employed for handling candidate terms within the **Lexical Changes** category. The model aims to control when to make changes in the input sentence by identifying patterns implemented through manual rules. Once a pattern is identified, the MLM is then employed. Unlike fixed substitutions, the MLM allows for dynamic selection of the most suitable substitute terms based on the specific context in which they will be applied. This adds a layer of flexibility and contextual understanding to the text modification process, making the substitutions more coherent and contextually relevant.

We create three rules considering the categories identified during the contrastive analysis. For the **Lexical Changes** category, we use a lookup table with 306 lexical variants between EP and BP, to identify terms that can be localized. This table is formed by observations from various parallel data sources, including the OPUS OpenSubtitles dataset (Lison and Tiedemann, 2016), linguistic articles and books related to this topic, and several literary books translated into both variants. Upon exact matching of a term's base form in the lookup table, we identified the optimal substitution by evaluating the probabilities of each candidate in context using MLM. The MLM returns a matrix of logits for each token position across the entire vocabulary. We apply the softmax function to the logits corresponding to the masked position to obtain a probability distribution. The candidate's probability is then computed by averaging the probabilities of its constituent tokens from this distribution:

$$P(c|S') = \frac{1}{|c|}\sum_{i=1}^{|c|} P(c_i|S')$$

where $|c|$ is the number of tokens in candidate c, the i-th token is represented by c_i, and $P(c_i|S')$ the probability of token c_i from the softmax-transformed logits of the masked sentence S'. In this context, the candidates refer to the alternative terms listed in the lookup table, along with the original term. In the case of **Gerund** category, we employ a graph to map potential patterns in a sentence, taking both lexical and syntactic attributes into account. When a pattern is recognized, the tokens, denoted by nodes, are substituted as dictated by the rule associated with that node. As for the

Pronominal Placement category, a regular expression is harnessed to spot the pattern and execute the substitution directly.

4.2 Pre-trained NMT model

The foundation of this model draws inspiration from conventional translation methodologies, where a pre-trained NMT model undergoes fine-tuning using parallel datasets. Specifically, we leverage the multilingual model mBART-50 described in Tang et al. (2020). This model is noteworthy as it is not just pre-trained but also simultaneously fine-tuned for multiple languages, encompassing Portuguese. For our fine-tuning process, we utilize the EP-BP parallel data available in the OPUS OpenSubtitles dataset (Lison and Tiedemann, 2016). This dataset is a collection of multilingual subtitle data gathered from various sources and offers a vast array of parallel sentences, making it particularly suitable for our study.

However, during our exploration, we notice a trend of substantial paraphrasing within the sentence pairs. This paraphrasing often extended beyond the necessary modifications for standard localization. To counteract this, we measure the similarity between these sentence pairs using cosine similarity, subsequently handpicking 100,000 examples that exhibited high similarity scores. Our training process incorporates a batch size of 4 and spanned over 10 epochs. We set the learning rate to 5×10^{-6} and a weight decay coefficient at 0.01.

4.3 GPT-4 + Examples

This model is inspired by recent studies that have achieved state-of-the-art results in translation tasks by utilizing prompt-based strategies with generative LLMs. To adopt these methodologies, we employ GPT-4 (OpenAI, 2023). Our prompt structure employs in-context learning and is based on literature results that enhanced the prompt by using examples of translations (Lyu et al., 2023; Mu et al., 2023; Brown et al., 2020). Specifically, we begin by providing 10 example sentences demonstrating localization to set the context for in-context learning. After establishing this context, we clarify that the subsequent task is one of localization. Concluding the prompt, we instruct the model to localize the given input sentence, transitioning it from the source Portuguese variant to the target variant. For our settings, we maintain the temperature at zero and incorporate ten random localization sentence pairs, specifically drawn from the "example" bucket of the FRMT dataset.

4.4 GPT-4 + Categories

In this approach, we meticulously design a prompt strategy that details each category identified in the contrastive analysis. For every category, illustrative examples showcasing the necessary modifications are provided. To conclude the prompt, we direct the model to translate the given input sentence from the source to the target Portuguese variant. Particularly for the **Lexical Changes** category, our method extends beyond static examples. Inspired by the findings of Yao et al. (2023), which showed an improvement when including dictionary examples in the prompt, we enrich the prompt with additional examples that are directly extracted from terms present in the input sentence. These terms have a reference point in our lookup table, which enumerates the lexical disparities between EP and BP.

5 Methodology

This section outlines the methodology of this study, which aimed to assess the ability of the proposed models to make localization between EP and BP. Therefore, the evaluation method sought to isolate or disregard essential text adaptations from optional change, allowing us to measure model performance based solely on overall localization quality and optional changes. For this reason, the FRMT benchmark (Riley et al., 2023) was used as the evaluation set, as its sentences capture region-specific linguistic differences between EP and BP variants. Both manual human evaluation and automatic metrics were employed for the assessment.

The results from our study cannot be directly equated with those of the FRMT benchmark (Riley et al., 2023). In our research, we focus on the direct localization between EP and BP. Conversely, the FRMT benchmark is designed for the task of translating English into a target language while accounting for regional nuances.

5.1 Dataset

The FRMT dataset contains a set of paired sentences between EP and BP. Sentences for each variant are translations from English sentences performed by translators specialized in the respective Portuguese variants. Importantly, the FRMT dataset curators specifically selected original English sentences that would require distinct, non-optional translations into each Portuguese variant.

For example, if the English sentence has the word "bus" it should be translated to "ônibus" in BP and "autocarro" in EP. For this study, we selected 300 random instances from the FRMT test set for evaluation due to cost constraints and used the sentences from the example set in prompts for the generative models.

5.2 Human Evaluation

Traditional automatic methods of MT evaluation are sensitive to the linguistic styles generated by the sentence translator, often underrepresenting minor yet crucial changes through automated metric values (Mariana, 2014). Therefore, this study's manual evaluation aims to precisely and humanely assess localization quality, seeking to identify types of errors that automated metrics might not capture.

The expert-based Multidimensional Quality Metrics (MQM) evaluation framework was employed (Freitag et al., 2022), chosen for its high fidelity to human assessment and its ability to individually evaluate different characteristics. The human evaluators were experienced linguists with training in translation and demonstrated knowledge of the language pair. They agreed on how to use MQM metrics, what linguistic aspects to take into consideration when evaluating each section of the translations, and how to attribute value to the identified mistakes. Evaluators were presented with a set of instances, each containing the source sentence, a reference sentence - which was used only in cases when the models' outputs were confusing or ambiguous to prevent evaluation biases -, and the model-generated translation to be evaluated. The selected metrics and MQM application methodology followed the recommendations of Freitag et al. (2022, 2021). Additionally, we introduced a custom metric specifically designed to count all optional changes made in the input sentence. Unlike obligatory changes, these optional alterations are not translation errors. Rather, they modify the style and to some extent paraphrase the sentence. Two evaluators were used for each instance, all of whom were experts in the target variant.

5.3 Automatic Metrics Evaluation

In addition to manual evaluation using MQM, which can be resource-intensive and not always feasible, we also employed standard MT metrics for a more scalable evaluation. These include BLEU (Bilingual Evaluation Understudy), which is specifically based on the FRMT benchmark and measures the overlap of token n-grams between the generated and reference text[1](Papineni et al., 2002). BLEU assesses the quality of generated text by comparing it with a reference one, quantifying how many words and phrases in the generated text match the reference one. TER (Translation Edit Rate) is designed to evaluate translations at the word level(Snover et al., 2006; Post, 2018). This metric calculates the number of edits (insertions, deletions, substitutions) required to change a generated text into the reference one. CharacTER, on the other hand, focuses on character-level edit distances (Wang et al., 2016). It measures the number of character-level edits (insertions, deletions, substitutions) needed to change the generated text into the reference one.

The inclusion of these automatic metrics facilitates comparisons across different studies and complements the in-depth, qualitative analysis provided by MQM. Their utilization offers a more comprehensive understanding of machine translation performance, encompassing both high-level fluency and fine-grained linguistic accuracy.

5.4 Lexical Accuracy

Lexical accuracy is an evaluation method focused on assessing the necessary and known lexical changes between Portuguese variants (Riley et al., 2023). For this purpose, we use the mapped lexical differences from FRMT lexical evaluation method consisting solely of words that must be adapted, regardless of context. For instance, "doutoramento" (EP) should be adapted to "doutorado" (BP). For each term pair, the number of sentences containing the correct variant (N_{match}) and the number of sentences with an incorrect variant (N_{mismatch}) were calculated with $Accuracy = N_{\text{match}}/(N_{\text{match}} + N_{\text{mismatch}})$.

5.5 Limitations

The scope of the experiments is focused on the localization of EP to BP. The human evaluators selected for this study have expertise in the BP variant, leading to a focused localization of the EP for the BP. This choice is informed by the insights from the FRMT results (Riley et al., 2023). The study showed that evaluators are more likely to assign higher rankings to sentences that are in their native variant. Therefore, our methodology aligns with this natural bias among evaluators, ensuring a

[1] nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1

more consistent localization to BP.

It is worth noting that the FRMT dataset derives its sentences from a predetermined, manually curated list of lexical differences. While meticulous, this approach may not cover the full range of regional linguistic variations due to the dynamic nature of language. Additionally, the FRMT dataset is not limited to minimal pairs. As a result, our study might analyze sentences that don't provide a direct one-to-one comparison between the two Portuguese variants.

6 Results

Our results are divided into three parts. First, the ones from a human evaluation using the MQM framework, followed by the results with automatic metrics and finalized with the lexical accuracy.

In Figure 1, the human evaluation results using MQM for the models delineated in Section 4 are presented. The overall quality measures the performance of the models in localizing 300 sentences, each evaluated by two different reviewers, emphasizing both necessary adjustments and broader linguistic characteristics such as spelling and grammar. On the other hand, the optional changes during localization present the quantification by human evaluators of unnecessary changes, presenting a value that seeks to represent how much the model is paraphrasing information during the localization process. These two metrics are independent and serve distinct evaluative purposes.

Notably, the GPT-4 + Examples configuration yielded the most promising results concerning overall localization quality, closely followed by the GPT-4 + Categories. This is consistent with the trends in state-of-the-art translation, where generative LLMs utilizing prompt strategies outperform rule-based and fine-tuned NMT approaches. Moreover, the model leveraging localization examples outperformed the one using category information in the prompt by a margin of 42%. This underscores the potential of few-shot settings in enabling the model to discern the nuances differentiating EP from BP and preserving the quality of localized sentences. Contrastingly, the Rule-based + MLM model's performance was 65% inferior to the second-best model, signifying that a strategy focusing merely on essential sentence aspects may not yield localizations of comparable quality to a generative LLM. However, it's important to note that the pre-trained NMT model was outperformed by the Rule-based + MLM approach, which showed a 17% improvement in performance. This indicates that the pre-trained NMT model encountered more significant challenges in this localization task.

Concerning optional changes, the Rule-based + MLM model showcased superior performance with a notable 61% reduction in unnecessary changes compared to the next best model. This underscores the high precision of the rules incorporated in the Rule-based + MLM model, which seems to pinpoint the essential linguistic aspects requiring modification adeptly. Following closely, the Pre-trained NMT model registered a significantly reduced level of unwarranted paraphrasing compared to the LLMs. This is likely attributable to its training on data exhibiting high similarity between sentence pairs. Also, between the two GPT-4 variants, the GPT-4 model augmented with Categories overcomes the performance of its counterpart supplemented with Examples. This engages with our hypothesis that explicit infusion of contrastive information detailing localization nuances can steer the model toward minimizing superfluous modifications.

Model	BLEU	TER	CharacTER
Pre-trained NMT	33.98	50.23	0.3701
Rule-based + MLM	34.62	46.62	0.3506
GPT-4 + Examples	**40.65**	**44.74**	**0.3281**
GPT-4 + Categories	38.93	45.93	0.3321

Table 3: Result with automatic metrics BLEU (↑), TER (↓) and CharacTER (↓) for each model. The sentences from the test set of the FRMT dataset were used. The reference sentence is a translation produced by a human translator from English to BP.

In accordance with human evaluation, the automatic metrics, as enumerated in Table 3, reveal that the GPT-4 models outperformed the other methods, consistent with findings from human assessments. Notably, the GPT-4 with Examples model surpassed its counterpart, GPT-4 with Contrastive Information, exhibiting a 4.2% augmentation in the BLEU score. In contrast, the Rule-based + MLM approach lagged by 11% in comparison to the second best model but managed to exceed the Pre-trained NMT model by a modest margin of 1.9%. The TER and CharacTER metrics further reinforced these observations, underscoring the nuanced yet tangible differences between the methodologies tested.

Table 4 presents the lexical accuracy results, that measure the performance of the models in adapting

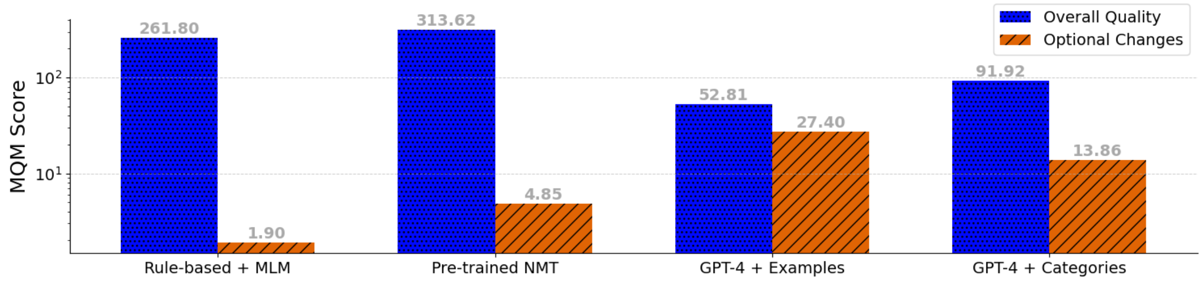

Figure 1: MQM (↓) scores for the localization from EP to BP with the models in logarithm scale. The dotted blue bar indicates the overall performance over the required changes in the sentence, indicating that the GPT-4 models achieved superior results. The orange stripe bar indicates the metric of optional changes made by the model, indicating that the Rule-based + MLM and Pre-trained NMT models tend to perform less paraphrasing than GPT-4.

Model	Accuracy (%)
GOLD	98.6
Pre-trained NMT	52.4
Rule-based + MLM	77.1
GPT-4 + Examples	96.1
GPT-4 + Categories	**97.4**

Table 4: Lexical accuracy on FRMT test. GPT-4 outperforms other models, with a small advantage for Categories information prompt strategy. GOLD is the human performance of sentences translated from English to BP by a human translator, taken from the FRMT dataset.

terms known to be different between the variants. Again, the generative LLMs demonstrated superior performance, with only a marginal 1.2 percentage point deficit to human performance (GOLD). Notably, the GPT-4 + Categories outperformed others, holding a 1.3 percentage point lead over its closest competitor, GPT-4 + Examples. The Rule-based + MLM model secured the third position, trailing the GPT-4 with Examples by 19 percentage points. The Pre-trained NMT model lagged further, underperforming the Rule-based + MLM model by 24 percentile point worse than GPT-4 + Examples.

In general, the generative LLMs consistently delivered superior performance, underscoring their adeptness at grasping the nuanced linguistic variations between EP and BP. Yet, these models also displayed a higher tendency for paraphrasing. However, the GPT-4 + Categories, when enhanced with explicit contrastive localization instructions, manifested reduced paraphrasing relative to the GPT-4 + Examples.

The Rule-based + MLM model exhibited minimal paraphrasing, signaling its precision in discerning vital changes in the input. However, this same precision might be a double-edged sword. The stringent adherence to rules possibly made it less adaptable, thus compromising its overall localization quality. Thus, it is possible to improve the model's performance at the cost of efforts to create new manual rules or through more flexible rules, which may come at the cost of increasing the level of paraphrases.

Insights of our empirical observations from the Pre-trained NMT model reveal a propensity to mirror the input sentence, and we believe that its training strategy is the reason behind it. By emphasizing sentence pairs with substantial similarity, which seeks to reduce paraphrasing, the model seems inadvertently biased, resulting in the least satisfactory performance in overall quality among the evaluated methods.

Interestingly, while GPT-4 + Examples got the best results for overall localization quality, GPT-4 + Categories triumphed in lexical accuracy. The strategy of dynamically including examples of lexical replacement in the prompt extracted directly from the input sentence seems pivotal to this achievement. These insights pave the way for innovative generative LLM approaches, leveraging few-shot paradigms combined with descriptive localization cues from contrastive analysis. Such model can capture nuances of the regional context through examples, and achieve greater precision in changes through the descriptions of the contrastive analysis categories. However, prospective methodologies should be conscious of the model's token constraints, as this approach might necessitate an ample token budget for effective prompting.

7 Conclusion

In this study, we addressed the challenges of localization between EP and BP when using different approaches and determined how effectively the models can perform the task. We carried out a contrastive analysis, which identified the most relevant differences between EP and BP, and integrated this information into the models. Our experiments relied on the dataset from the Benchmark FRMT (Riley et al., 2023), and we also based our evaluation on the feedback provided by professional linguists specialized in the target variant. The results show that generative GPT-4 delivered superior performance, which is consistent with the trends in state-of-the-art translation, where generative LLMs utilizing prompt strategies got the best results. Also, we showed the ability of the models to perform localization avoiding paraphrasing the input, where the results showed that the rule-based approach makes fewer unnecessary changes, compared to LLMs.

These findings open doors for novel generative LLM techniques with prompts, which utilize few-shot models along with descriptive cues from contrastive analysis. This approach could allow the model to understand regional subtleties via examples and enhance accuracy in modifications using insights from the contrastive analysis classifications. Moreover, we intend to tailor causal language models with techniques such as prompts that include task-specific knowledge in order to further experiment with this task.

8 Acknowledgments

This study was financed by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

We thank Bianka Buschbeck, Miriam Exel, and Christian Lieske (SAP SE) as well as Douglas Hermann, Maitê Dupont, Carolina Scheuermann, Felipe Costa, and Ariel Azzi (SAP Labs Latin America) for their valuable contributions to our work and valuable feedback on draft versions of this paper.

References

2014. *Acordo Ortográfico da Língua Portuguesa: Atos Internacionais e Normas Correlatas*, 2 edition. Senado Federal, Coordenação de Edições Técnicas, Brasília. Conteúdo: Dispositivos constitucionais pertinentes – Acordo Ortográfico da Língua Portuguesa – Outros atos internacionais – Anexo: acordos ortográficos anteriores – Normas correlatas – Informações complementares.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report.

Merouan Bendi. 2020. The reception of localized content: A user-centered study of localized software in the algerian market. *When Translation Goes Digital*.

Paul Bennett. 2002. Teaching contrastive linguistics for MT. In *Proceedings of the 6th EAMT Workshop: Teaching Machine Translation*, Manchester, England. European Association for Machine Translation.

Yevgeni Berzak, Roi Reichart, and Boris Katz. 2015. Contrastive analysis with predictive power: Typology driven estimation of grammatical error distributions in ESL. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 94–102, Beijing, China. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020.

Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Ataliba Teixeira de Castilho. 2013. A hora e a vez do português brasileiro. *Museu da Língua Portuguesa. São Paulo*, 24.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit3: Web inventory of transcribed and translated talks. In *Proceedings of the Conference of European Association for Machine Translation (EAMT)*, pages 261–268.

Yijie Chen, Yijin Liu, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. 2023. Improving translation faithfulness of large language models via augmenting instructions.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Natalia Djajarahardja. 2020. Aspectos da variação entre o pe e o pb: guia para a adaptação linguística entre as duas variedades. Master's thesis.

Federico Fancellu, Andy Way, and Morgan O'Brien. 2014. Standard language variety conversion for content localisation via SMT. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*, pages 143–149, Dubrovnik, Croatia. European Association for Machine Translation.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2023. Exploring human-like translation strategy with large language models.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation.

Linda Hříbalová. 2010. *Diferenças entre o português europeu eo português brasileiro*. Ph.D. thesis, Masarykova univerzita, Filozofická fakulta.

Carl James. 1980. Contrastive analysis. Research report. ERIC Number: ED202229; 208 pages.

Mary KATO. 2006. Comparando o português da américa com o português de portugal e com outras línguas. *Língua Portuguesa, Museu da Língua Portuguesa*.

Ping Ke. 2019. *Contrastive Linguistics*, 1 edition. Peking University Linguistics Research. Springer, Singapore.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation.

Iørn Korzen and Morten Gylling. 2017. Chapter 3 text structure in a contrastive and translational perspective : On information density and clause linkage in italian and danish iørn korzen.

Tomasz P Krzeszowski. 2011. *Contrasting languages: The scope of contrastive linguistics*, volume 51. Walter de Gruyter.

Surafel Melaku Lakew, Aliia Erofeeva, and Marcello Federico. 2018. Neural machine translation into language varieties. *CoRR*, abs/1811.01064.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles.

Yijin Liu, Xianfeng Zeng, Fandong Meng, and Jie Zhou. 2023. Instruction position matters in sequence generation with large language models.

Nuno G. Lopes and Carlos J. Costa. 2008. Erp localization: exploratory study in translation: European and brazilian portuguese. In *ACM International Conference on Design of Communication*.

Chenyang Lyu, Jitao Xu, and Longyue Wang. 2023. New trends in machine translation using large language models: Case examples with chatgpt.

Valerie R Mariana. 2014. *The Multidimensional Quality Metric (MQM) framework: A new framework for translation quality assessment*. Brigham Young University.

Luis Marujo, Nuno Grazina, Tiago Luis, Wang Ling, Luisa Coheur, and Isabel Trancoso. 2011. BP2EP - adaptation of Brazilian Portuguese texts to European Portuguese. In *Proceedings of the 15th Annual Conference of the European Association for Machine Translation*, Leuven, Belgium. European Association for Machine Translation.

Yongyu Mu, Abudurexiti Reheman, Zhiquan Cao, Yuchun Fan, Bei Li, Yinqiao Li, Tong Xiao, Chunliang Zhang, and Jingbo Zhu. 2023. Augmenting large language model translators via translation memories.

OpenAI. 2023. Gpt-4 technical report.

John E Ortega, Iria de Dios-Flores, Pablo Gamallo, and José Ramom Pichel. 2022. A neural machin e translation system for galician from transliterated portuguese text. In *Proceedings of the Annual Conference of the Spanish Association for Natural Language Processing. CEUR Workshop Proceedings*, volume 3224, pages 92–95.

Asako Otomo. 2004. A contrastive study of function verbs in English and Japanese : Cut and kiru. In *Proceedings of the 18th Pacific Asia Conference on Language, Information and Computation*, pages 235–242, Waseda University, Tokyo, Japan. Logico-Linguistic Society of Japan.

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

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, Belgium, Brussels. Association for Computational Linguistics.

Vikas Raunak, Arul Menezes, Matt Post, and Hany Hassan Awadalla. 2023. Do gpts produce less literal translations?

Parker Riley, Timothy Dozat, Jan A. Botha, Xavier Garcia, Dan Garrette, Jason Riesa, Orhan Firat, and Noah Constant. 2023. FRMT: A benchmark for few-shot region-aware machine translation. *Transactions of the Association for Computational Linguistics*, 11:671–685.

João Rodrigues, Luís Gomes, João Silva, António Branco, Rodrigo Santos, Henrique Lopes Cardoso, and Tomás Osório. 2023. Advancing neural encoding of portuguese with transformer albertina pt-*.

Marta Ruiz Costa-Jussà, Marcos Zampieri, and Santanu Pal. 2018. A neural approach to language variety translation. In *COLING 2018: The 27th International Conference on Computational Linguistics: Proceedings of the Conference*, Stroudsburg, PA. Association for Computational Linguistics. Conference held on August 20-26, 2018, Santa Fe, New Mexico, USA.

Reinhard Schäler. 2004. Language resources and localisation. In *Proceedings of the Second International Workshop on Language Resources for Translation Work, Research and Training*, pages 18–25, Geneva, Switzerland. COLING.

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

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning.

Paul Teyssier and Celso Ferreira da Cunha. 1982. História da língua portuguesa. *(No Title)*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. 2016. CharacTer: Translation edit rate on character level. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 505–510, Berlin, Germany. Association for Computational Linguistics.

Sze-Meng Jojo Wong and Mark Dras. 2009. Contrastive analysis and native language identification. In *Proceedings of the Australasian Language Technology Association Workshop 2009*, pages 53–61, Sydney, Australia.

Binwei Yao, Ming Jiang, Diyi Yang, and Junjie Hu. 2023. Empowering llm-based machine translation with cultural awareness.