Investigating Paraphrase Generation as a Data Augmentation Strategy for Low-Resource AMR-to-Text Generation

Marco Antonio Sobrevilla Cabezudo[♢♣] **Marcio Lima Inácio**[♠]
Thiago Alexandre Salgueiro Pardo[♣]
[♢] Artificial Intelligence Research Group (IA-PUCP)
Pontifical Catholic University of Peru, Perú
[♣] Interinstitutional Center for Computational Linguistics (NILC)
Institute of Mathematical and Computer Sciences, University of São Paulo. São Carlos/SP, Brazil
[♠] CISUC - University of Coimbra, Coimbra, Portugal
msobrevilla@pucp.edu.pe, mlinacio@dei.uc.pt, taspardo@icmc.usp.br

Abstract

Abstract Meaning Representation (AMR) is a meaning representation (MR) designed to abstract away from syntax, allowing syntactically different sentences to share the same AMR graph. Unlike other MRs, existing AMR corpora typically link one AMR graph to a single reference. This paper investigates the value of paraphrase generation in low-resource AMR-to-Text generation by testing various paraphrase generation strategies and evaluating their impact. The findings show that paraphrase generation significantly outperforms the baseline and traditional data augmentation methods, even with fewer training instances. Human evaluations indicate that this strategy often produces syntactic-based paraphrases and can exceed the performance of previous approaches. Additionally, the paper releases a paraphrase-extended version of the AMR corpus.

1 Introduction

Abstract Meaning Representation (AMR) is a widely popular semantic representation. It encodes the whole meaning of a sentence into a labelled directed and rooted graph, including information such as semantic roles, named entities, and coreferences, among others (Banarescu et al., 2013). Moreover, it has been successfully used in diverse applications/tasks such as automatic summarization (Vilca and Cabezudo, 2017), and paraphrase detection (Issa et al., 2018).

Its popularity is partly attributed to its extensive use of mature linguistic resources, like PropBank (Palmer et al., 2005), and its effort to abstract from syntax. Figure 1 illustrates the AMR graph (Subfigure A) and the PENMAN notation (Matthiessen and Bateman, 1991) (Sub-figure B) for the sentence "*The boy must go*" along with other alternative surface forms that, while syntactically and lexically different, convey the same meaning.

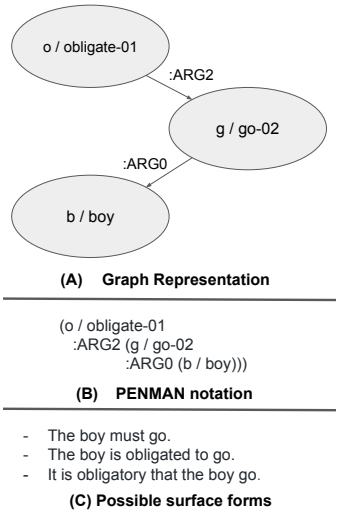

Figure 1: AMR for the sentence "The boy must go."

Interestingly, AMR corpora, as far as we know, include only one reference per AMR graph, not leveraging their syntax-independent nature. In contrast, other semantic representations, such as those in the WebNLG challenge (Gardent et al., 2017) or the E2E dataset (Dušek et al., 2020), typically provide multiple references for each representation. Having multiple references is advantageous for developing Natural Language Generation systems, as it helps them handle potential noise by increasing data diversity (Dušek et al., 2020).

On the other hand, manually creating additional references can be costly. Specifically, the words used in surface forms are tightly connected to the concepts in an AMR graph (Banarescu et al., 2013).

Proceedings of the 17th International Natural Language Generation Conference, pages 663–675
September 23–27, 2024.

Thus, references generated for an AMR graph should ideally include only its concepts in their canonical form or possible derivatives as much as possible. For instance, the concept “boy" in Figure1 should not be replaced with “guy" in a surface form, even if both terms are interchangeable. An alternative to manual annotation is the automatic generation of new references using paraphrase generation models. However, we must still adhere to the aforementioned guideline.

Paraphrase generation has been valuable for data augmentation in various tasks such as natural language understanding (Okur et al., 2022), and task-oriented dialogue systems (Gao et al., 2020). However, to our knowledge, this technique has not yet been explored to enhance AMR-to-Text generation performance or to develop a more robust AMR corpus (apart from the work of Huang et al. (2023)). Moreover, other methods in the literature that utilize AMR parsers to generate new instances (Castro Ferreira et al., 2017; Mager et al., 2020; Ribeiro et al., 2021) might outperform paraphrase generation. Nevertheless, we focus on low-resource scenarios where AMR parsing could negatively impact the AMR-to-Text generation task.

This work seeks to assess the helpfulness of paraphrases in the context of Low-resource AMR-to-text generation for Brazilian Portuguese (BP). More, specifically, we try to answer the question *To what extent can paraphrase generation contribute to improvement of the AMR-to-Text Generation in a Low-resource scenario?* To answer this question, we investigate two approaches for generating paraphrases. The first approach employs a Portuguese paraphrasing model (Pellicer et al., 2022). The second approach uses English as pivot language and is divided into two sub-approaches: one relies solely on machine translation models, while the other also includes an English paraphrase generation model. In addition, we compare this strategy with other well-known data augmentation strategy based on automatic parsing.

Due to the possibility of adding unrelated paraphrases introducing noise into the models, we explore using three selection criteria. These criteria help select a specific number of high-quality paraphrases. Finally, we examine if added paraphrases can benefit when included in the development set in a multi-reference training.

In general, our main contributions are:

- we investigate two paraphrase generation approaches (monolingual and cross-lingual) to generate multiple references in AMR-to-Text generation task;
- we conduct experiments and analysis to prove the helpfulness of paraphrases for Low-resource AMR-to-Text generation;
- we release a paraphrase-focused version of the AMR corpus for Brazilian Portuguese.

2 Paraphrase Generation for producing multiple references

To evaluate the helpfulness of paraphrasing for the Low-Resource AMR-to-Text generation task, we explore generating paraphrases for each reference in the AMR corpus. In particular, we explore two approaches for performing it. The first one assumes the existence of paraphraser models for the target language (in our case, Portuguese). The second one is a cross-lingual approach that tackles the problem under the assumption that there is no paraphraser model for the target language; however, there is a bilingual corpus or a translation model between the target language and another richer-resource language (e.g., English) and, possibly, a paraphrasing model in the richer-resource. This way, we can use this language as a pivot.

Figure 2 shows an example of both approaches. The sub-figure A corresponds to the first approach, whereas the other two (B and C) correspond to the cross-lingual approach. In B, we only use machine translation models, whereas in C, we also use a paraphrasing model for the pivot language.

2.1 Portuguese Paraphrase Generation

This strategy uses a paraphraser model for Portuguese to generate the candidate paraphrases for reference. In particular, we use the model proposed by Pellicer et al. (2022) (named PTT5-Paraphraser), which was obtained by fine-tuning PTT5 (Carmo et al., 2020) on the Portuguese subset from TaPaCo corpus (Scherrer, 2020).

2.2 English-pivot Paraphrase Generation

Back-translation It is a simple way to generate paraphrases that consists of using a translation model that translates the reference into a pivot language (e.g., English) and another model that does the inverse process. This strategy has successfully been used in tasks such as machine translation

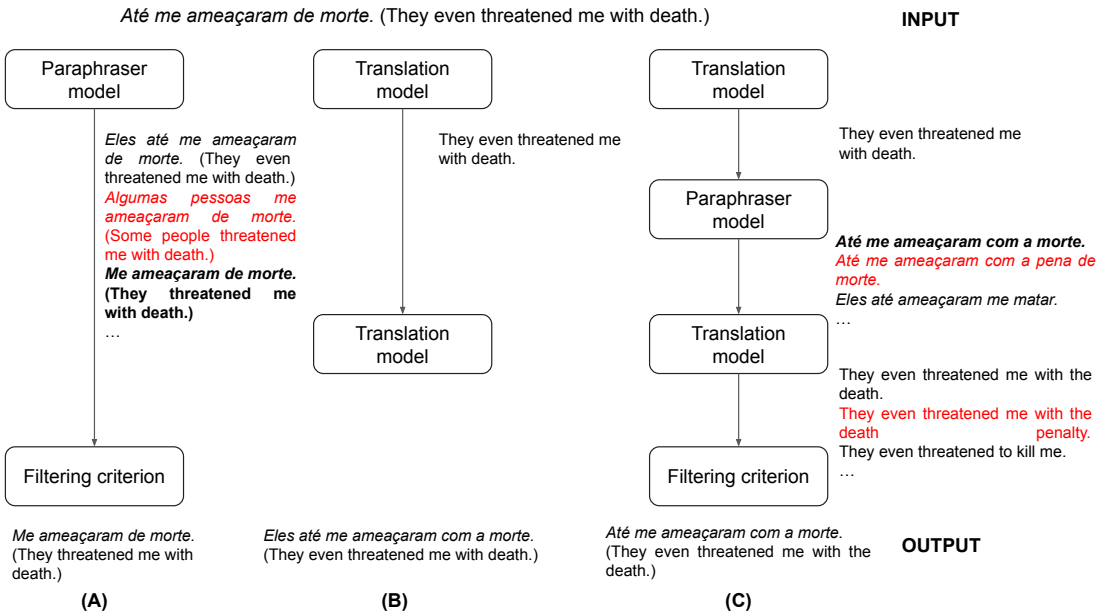

Figure 2: Pipeline Example for Paraphrase Generation. (A) Portuguese approach: A sentence written in Brazilian Portuguese (BP) is given to a Portuguese paraphrase model, and it generates the paraphrases. (B) English-pivot approach: A sentence written in BP is given to a machine translation model that generates the corresponding translation and then passes it to another translation model (back-translation) that generates a paraphrase of the original sentence. (C) English-pivot approach: Similar to (B), but translation is passed into an English paraphrase model to generate the paraphrases that are given to the back-translation model. In addition, a filtering criterion is used to select the best paraphrases.

(Edunov et al., 2020) and data-to-text generation (Sobrevilla Cabezudo et al., 2019).

We explore two ways of applying back-translation. The first one consists of generating only one output for each translation step. In this way, we only generate one paraphrase for each instance. The second one consists of generating only one output in the first translation step and n outputs in the second step (back-translation step).

Translations are generated by two translation models (*Portuguese-to-English* and vice-versa) provided by MariaNMT (Junczys-Dowmunt et al., 2018) and available at HuggingFace[1]

Back-translation + English Paraphrase Generation Similar to the previous strategy, it generates only one output in the first translation step. However, the second step aims to generate "n" paraphrases for the translation obtained previously by using a paraphraser model in the pivot language. Finally, another translation step converts the "n" paraphrases into the target language.

The paraphraser model for English is similar to the one proposed by Pellicer et al. (2022), which is obtained by fine-tuning T5 (Raffel et al., 2020) on the PAWS corpus (Zhang et al., 2019)[2].

One of the main drawbacks of all the proposed strategies is that the paraphrases generated can differ from the source reference in lexical terms due to translation and paraphraser models. Therefore, we explore some widely-used metrics used in paraphrase evaluation for ranking and selecting the best paraphrases for a target reference (Zhou and Bhat, 2021). In particular, we use BLEU (Papineni et al., 2002), METEOR (Lavie and Agarwal, 2007)[3] and TER (Snover et al., 2006).

3 Experimental Setup

3.1 Dataset

We conduct experiments on the AMRNews, which includes the journalistic section of the AMR-PT corpus (Inácio et al., 2022)[4]. The AMRNews corpus comprises 870 sentences from Brazilian news texts manually annotated following the

[1]Available at Helsinki-NLP/opus-mt-ROMANCE-en and Helsinki-NLP/opus-mt-en-ROMANCE.

[2]Available at https://huggingface.co/Vamsi/T5_Paraphrase_Paws.

[3]In experiments, we only use the stem and the exact similarity.

[4]AMRNews is available at https://github.com/nilc-nlp/AMR-BP/tree/master/AMRNews.

AMR guidelines for Brazilian Portuguese (Sobrevilla Cabezudo and Pardo, 2019). The corpus is split into 402, 224, and 244 instances for training, development, and test sets.

3.2 Settings

We evaluate different criteria such as the number of paraphrases per instance added to the training set (1-10), the metric used for selecting the best paraphrases (BLEU, TER, and METEOR), and the use of the paraphrases in two ways:

- Only-Train (T): We add paraphrases into the training set, i.e., we use it as a paraphrase-based data augmentation strategy.
- Train-Dev (B): We add paraphrases into the training/development sets to verify if increasing diversity in the development set can lead to better performance. Besides, this approach aims to create a multi-reference AMR corpus.

Finally, the new multi-reference AMR corpus comprises AMR graphs, corresponding sentences, and paraphrases (one per line). For training, each input consists of a prefix and an AMR graph in the PENMAN notation (eliminating the frameset numbers). We use the expression "*gerar texto desde amr:*" ("Generate text from amr:") as the prefix for each instance, and the output is the corresponding sentence or paraphrase.

3.3 Baselines

Fine-tuning on AMRNews To evaluate the effectiveness of paraphrasing in increasing the number of references, we establish the baseline model by fine-tuning PPT5 (Carmo et al., 2020) on the original AMRNews, which includes only one reference.

Data augmentation by Parsing We explore another data augmentation strategy. Specifically, we train an end-to-end AMR parser and use it to annotate a subset from the corpus Bosque (Afonso et al., 2002)[5] in a similar way to existing literature (Castro Ferreira et al., 2017; Mager et al., 2020). The parser is trained by fine-tuning PTT5 on the AMRNews. The source side comprises the sentences, and the target one comprises the AMR graphs in PENMAN notation; however, we remove the variables from the PENMAN notation and use the actual concepts in the coreferences.

This approach suffers from problems such as the lack of parentheses or coreferences. This way, we use the tool proposed by van Noord and Bos (2017)[6] to restore the AMR graphs. In total, we add 4,126 instances to the training set.

4 Results and Discussion

Table 1 shows the overall results for the models on the test set from the original AMR corpus[7]. We report the results for each approach and each paraphrase selection criterion, training the models under the setting T. In general, we report BLEU (Papineni et al., 2002), METEOR (Lavie and Agarwal, 2007), chrF++ (Popović, 2017), and BERTScore (Zhang et al., 2020)[8][9].

Overall, we can see that all the paraphrase-based models surpass the baseline in all the metrics, with the largest difference of 3.81 for BLEU, 0.04 points for METEOR, 0.05 points for chrF++ and 0.02 points for BERTScore[10], proving the helpfulness of this strategy.

Regarding the paraphrase generation strategy, we observed that, as expected, paraphraser models (both for Portuguese and English-pivot approaches) produce better results than translation models alone. Additionally, METEOR appears to yield slightly better performance when using the paraphrase-based approach and there are mixed results in translation-based approaches.

We also note that all approaches outperform the results obtained by the classic data augmentation approach (Bosque-Augmented in Table 1), requiring fewer instances to achieve better performance. For example, the Portuguese approach only needs approximately 2,000 instances to achieve higher performance. Surprisingly, we can see that even adding only one paraphrase per instance (BACK-TRANSLATION 1-1 experiment in Table 1) achieves comparable results.

The main drawback is that performance does not improve with more than 8 paraphrases and may even decrease (see Figure 4 and Figure 6 in Appendix A). It is suggested to evaluate whether increasing instances in the classic data augmentation

[5]Available at https://www.linguateca.pt/Floresta/corpus.html.

[6]Available at https://github.com/RikVN/AMR.

[7]The model for each criterion is selected according to the best metrics obtained in the development set

[8]We execute four runs for each experiment and show the mean and standard deviation.

[9]Metrics are calculated by using the code available at https://github.com/WebNLG/GenerationEval.

[10]We note that the last three metrics are reported in the range 0.00-1.00.

APPROACH		CRITERIA	BLEU	METEOR	chrF++	BERTScore
BASELINE			10.39 ± 0.48	0.29 ± 0.01	0.41 ± 0.01	0.82 ± 0.00
BOSQUE-AUGMENTED			11.35 ± 0.64	0.29 ± 0.01	0.43 ± 0.01	0.82 ± 0.00
PORTUGUESE	PARAPHRASE	BLEU	13.01 ± 0.45	0.32 ± 0.01	0.44 ± 0.01	0.83 ± 0.00
		METEOR	14.20 ± 0.41	0.33 ± 0.01	0.46 ± 0.01	0.84 ± 0.01
		TER	14.02 ± 1.48	0.33 ± 0.02	0.44 ± 0.01	0.84 ± 0.01
ENGLISH-PIVOT	BACK-TRANSLATION 1-1		11.28 ± 0.87	0.29 ± 0.01	0.42 ± 0.02	0.82 ± 0.01
	BACK-TRANSLATION 1-N	BLEU	14.00 ± 1.22	0.32 ± 0.01	0.44 ± 0.01	0.84 ± 0.01
		METEOR	13.46 ± 1.16	0.32 ± 0.01	0.44 ± 0.01	0.83 ± 0.00
		TER	11.89 ± 0.61	0.31 ± 0.01	0.43 ± 0.01	0.83 ± 0.01
	BACK-TRANSLATION + PARAPHRASE	BLEU	13.43 ± 1.63	0.32 ± 0.01	0.44 ± 0.02	0.83 ± 0.00
		METEOR	14.22 ± 0.54	0.33 ± 0.01	0.45 ± 0.01	0.83 ± 0.00
		TER	14.30 ± 1.03	0.33 ± 0.01	0.45 ± 0.01	0.84 ± 0.01

Table 1: Overall results on setting T. We show the best models for each selection criterion. BOSQUE-AUGMENTED is the method of parsing to incorporate more instances into the training set. BACK-TRANSLATION 1—1 represents the method that generates one translation and then uses it to generate the corresponding back-translation. On the other hand, BACK-TRANSLATION 1—N represents that one that generates one translation and uses it to generate multiple possible back-translations. BACK-TRANSLATION + PARAPHRASE represents the method that uses English paraphrase generation in the middle of the translation and back-translation steps.

approach could lead to better results or simply introduce more noise (due to the extremely low-resource setting), potentially harming performance.

To conduct a deep analysis, we answer some questions about the number of paraphrases, the paraphrase selection criteria, and the setting used for augmenting data (T or B).

How many paraphrases are helpful? Regarding setting T (where instances are only added to the training set), Figures 4 and 6 illustrate the changes in performance on the development set based on the number of paraphrases used for data augmentation.

Overall, the best performance is achieved by adding a few paraphrases (up to 5-6) for the Portuguese paraphrasing approach. However, for the English-pivot approaches, more paraphrases (7-9) are needed. This may be due to a trade-off between quantity and quality: while English-pivot approaches may produce lower-quality paraphrases, the increased diversity from adding more paraphrases can enhance performance.

Another important point is that the back-translation + paraphrasing strategy presents the steepest decline in all metrics when more data is added, especially with 10 paraphrases. This indicates the need for careful selection of instances when using this strategy. Conversely, other approaches show a gentler decline, with BERTScore being the least affected metric. The semantic nature of BERTScore likely explains its resilience to synonyms and paraphrases in the outputs.

Additionally, the standard deviation for most metrics rises with the addition of more paraphrases, particularly impacting the BLEU score. This is expected, as BLEU is a more restrictive metric. A plausible explanation is that incorporating more paraphrases in training makes the model more likely to produce diverse paraphrases.

Figures 5 and 7 illustrate the results when models are trained under setting B. Different from experiments on setting T (where 5-6 paraphrases are enough), adding 7-9 paraphrases yields better results. However, adding 10 paraphrases results in a performance drop, with both the Portuguese and the English-pivot back-translation + paraphrasing strategies being the most affected.

What are the best paraphrase selection criteria? In setting T (Figures 4 and 6), the behavior varies based on the paraphrase generation approach. For the Portuguese method, METEOR metric perform better when fewer paraphrases (5-6 paraphrases) are added, but performance declines with more paraphrases. This is likely because this metric quickly select the best instances when paraphrases are of good quality, assuming the Portuguese approach introduces less noise.

For English-pivot approaches, results along the three metrics are similar. In particular, TER produces different trends. However, in test it shows a drop with back-translation alone but comparable results to the Portuguese approach when English paraphrase generation is included, proving useful in the absence of non-English paraphrase models.

In setting B, the Portuguese approach shows different results, with BLEU and TER as the best selection criteria but high standard deviations. Evaluating models on the test set reveals that while TER achieves high performance in development, it de-

creases in test set BLEU scores, reflecting TER's nature of not prioritizing exact words/n-grams. For English-pivot approaches in setting B, similar behavior to setting T is observed, with BLEU and METEOR producing the best results.

How much does the paraphrase's quality affect the performance? To assess how paraphrase quality impacts AMR-to-Text performance, we trained a model using one of the best settings but replaced the best paraphrases with the worst ones. We used the Portuguese approach, the METEOR criterion, and 5 paraphrases. In the case of the worst ones, we select the worst 5 paraphrases from the experiment with 10 paraphrases.[11]

Table 2 shows the development set results and similarity metrics between the paraphrases and original training instances. The metrics include cosine similarity and the three selection metrics from the experiments (BLEU, TER, and METEOR). All similarity metrics showed a significant drop, with cosine similarity being the least affected due to its ability to handle synonyms and related words.

The overall performance decreased across all metrics, with BLEU being less affected (a drop of 0.34 points). Conversely, its standard deviation doubled. It might confirm the hypothesis that paraphrase generation serves as an oversampling strategy in which some infrequent words/n-grams become easier to decode because they become more frequent but, at the same time, it introduces some noise coming from less-related or nonsense words.

How much does including paraphrases in the development set contribute? Given the current corpus has only one reference per instance, we created a multi-reference version of the test set. This was done by applying a successful previous strategy: using a Portuguese-based model trained with five paraphrases per instance and METEOR as the selection criterion. The resulting multi-reference test set contains 1-6 references per instance.

Table 3 shows the performance of the Portuguese-based model trained in both settings (T and B) for each selection criterion, evaluated on both one-reference and multi-reference test sets. In the one-reference evaluation, adding paraphrases to the development set yielded mixed results, increasing standard deviation and affecting the BLEU score the most. This suggests the strategy can be helpful but also introduces noise and instability. BLEU was the most beneficial selection criterion, improving performance by 1.24 points (from 13.01 to 14.25), while TER caused a small BLEU performance drop, correlating to previous analysis that suggests TER is more prone to generate different words/synonyms, keeping the meaning (as the other metrics remain almost the same).

In the multi-reference evaluation, we confirm that TER tends to produce more diverse outputs and may not harm the output quality as the performance in both settings (T and B) is almost the same (differently from the one-reference evaluation) in terms of BLEU and better in terms of METEOR and chrF++. On the other hand, the performance difference for the BLEU and METEOR selection criteria is similar to the obtained in the one-reference evaluation.

5 Manual Revision

To gain insights into some results, we conduct a manual revision. We select 112 instances from the development set to identify the primary mistakes and phenomena generated by the models.

We define two categories in the evaluation: valid and invalid outputs. Valid outputs are further divided into three sub-categories: "equivalent", where the system output and the reference are the "same" (with minor modifications such as the use of determiners); "semantic", where the system output is equivalent to the reference but uses different words or non-syntax paraphrases; and "syntactic", where the output is equivalent to the reference but exhibits some syntax differences (e.g., changing from active to passive voice).

Invalid outputs include 3 sub-categories: "missing", when the system output is similar to the reference, but omitted a few words; "partial hallucination", when the output contains part of the reference and part of extra information not related to the input/reference; and "total hallucination", when the output is totally unrelated to the reference.

The analyzed approaches include the baseline, the data augmentation by parsing approach, the Portuguese paraphrasing approach (under the setting T and B), and the two English-pivot sub-approaches under the setting T. More details about the selected models are described in A.3.

Table 4 shows the percentage of valid and invalid outputs according to the distribution of their subcategories. In general, non-paraphrase approaches, i.e., the baseline and the Bosque-augmented ones,

[11] It is worth noting that we set a beam size of 20 during experiments. This way, the experiment represents the best of the worst scenarios.

	SIMILARITY				EVALUATION			
	COSINE	BLEU	TER	METEOR	BLEU	METEOR	chrF++	BERTScore
BEST	0.91 ± 0.09	54.87 ± 19.17	29.33 ± 28.35	0.73 ± 0.15	15.73 ± 0.59	0.37 ± 0.01	0.46 ± 0.01	0.84 ± 0.00
WORST	0.86 ± 0.11	40.55 ± 17.42	42.35 ± 40.10	0.59 ± 0.17	15.39 ± 1.28	0.35 ± 0.01	0.45 ± 0.01	0.83 ± 0.00

Table 2: Results for the Portuguese approach when the best 5 paraphrases (BEST) and the worst 5 paraphrases (WORST) are added to the training set. The Portuguese approach uses the METEOR selection criteria for this experiment. In addition, models are evaluated on the development set.

REF.	SETTING		TEST			
	SET	CRITERIA	BLEU	METEOR	chrF++	BERTScore
One	T	BLEU	13.01 ± 0.45	0.32 ± 0.01	0.44 ± 0.01	0.83 ± 0.00
		METEOR	14.20 ± 0.41	0.33 ± 0.01	0.46 ± 0.01	0.84 ± 0.01
		TER	14.02 ± 1.48	0.33 ± 0.02	0.44 ± 0.01	0.84 ± 0.01
	B	BLEU	14.25 ± 1.61	0.33 ± 0.01	0.45 ± 0.02	0.83 ± 0.01
		METEOR	14.75 ± 1.35	0.33 ± 0.02	0.46 ± 0.01	0.84 ± 0.00
		TER	13.77 ± 1.14	0.33 ± 0.01	0.45 ± 0.01	0.84 ± 0.00
Multi	T	BLEU	20.91 ± 1.02	0.38 ± 0.01	0.47 ± 0.01	0.85 ± 0.00
		METEOR	21.76 ± 0.32	0.39 ± 0.01	0.49 ± 0.01	0.86 ± 0.01
		TER	22.80 ± 1.82	0.39 ± 0.01	0.48 ± 0.01	0.85 ± 0.01
	B	BLEU	22.19 ± 1.69	0.38 ± 0.02	0.49 ± 0.02	0.85 ± 0.01
		METEOR	22.36 ± 1.54	0.39 ± 0.02	0.50 ± 0.01	0.86 ± 0.00
		TER	22.83 ± 0.84	0.40 ± 0.01	0.50 ± 0.01	0.86 ± 0.00

Table 3: Best results on the test for the Portuguese approach on setting T and B using one reference and multi-references set. The results are shown for each criteria.

produce more equivalent outputs (up to 15.18%). However, they are more prone to generate total hallucinations (up to 64.29%). In the case of the Bosque-Augmented, it is expected since the AMR quality of the augmented instances can add more noise to the training.

Concerning the paraphrase approaches, we note that the Portuguese one produces the best results, generating more semantic and syntax-based paraphrases than all remaining approaches. In particular, we can see that the percentage of syntactically equivalent outputs surpasses the same percentage on the Bosque-augmented approach by 8.03% (five times). Furthermore, this approach also gets more valid outputs in general (26.78%), beating the previously mentioned approach (20.54%).

On the other hand, English-pivot approaches are also promising to generate syntactic-based paraphrases; however, they are unsuitable for generating equivalent outputs, being overcome by the Bosque-augmented approach almost twice (7.14%). In addition, we note that the overall percentage of valid outputs is lower than the obtained by the baseline and the Bosque-augmented approach (19.64% and 18.76% vs 22.32% and 20.54%), showing that automatic metrics can hide some undesirable behaviour as English-pivot approaches gets better results in automatic evaluation. It could be explained by the fact that generating more diverse (and less related) paraphrases during training can add noise, thus being prone to generate more hallucinations.

Analyzing the invalid outputs, we see that Paraphrase approaches tend to omit some words in the outputs, particularly Portuguese ones. This way, some models generate "*Ele ficou só*" ("He was alone.") instead of the reference "*Ele ficou literalmente só*" ("he was literally alone."), omitting the word "*literalmente*" ("literally").

Concerning the hallucinations, it is worth noting that all approaches produce a high number of hallucinations (47.32%-64.29%). This can be produced by the limited size of the original dataset and the high relation/node sparsity, however, more research should be done to confirm this hypothesis. About the approaches, paraphrase approaches are less prone to generate total hallucinations, being the best Portuguese approach and the worst English-pivot approach that applies Back-translation and Paraphrase generation. We can see an example in Figure 3.

As we can see in Figure 3, paraphrase approaches produce outputs more related to the reference, demonstrating the effectiveness of the approach. Another interesting finding we found is that the major gain of this approach raises in the ability to produce the tokens included in the AMR representation, i.e., paraphrase approach helps to better identifying concepts but not relations between them. We analyze this by using a sample that comprises only totally hallucinated outputs in the baseline model and verifying to what class (valid/invalid) they belong after applying the paraphrase approach. The results show that 13.23% of the outputs are fixed in the paraphrase approach, but 17.65% and 17.65% are classified as missing and partial hallucination classes, respectively.

Finally, we find the occurrence of partial hallucinations in the outputs produced by the paraphrase approach. Even though models can be better than the baseline, they are more prone to generate additional expressions to the original one. For instance, the model generates "*outro problema* ***político*** *tem um fundo político.*" ("another **political** problem has a political background.") when the reference

		VALID				HALLUCINATIONS	
		EQUIVALENT	SEMANTIC	SYNTACTIC	MISSING	PARTIAL	TOTAL
BASELINE		15.18	0.00	7.14	9.82	8.93	60.72
BOSQUE-AUGMENTED		15.18	2.68	2.68	8.04	10.71	64.29
PORTUGUESE	PAR (T)	12.50	3.57	10.71	15.18	16.96	47.32
	PAR (B)	10.71	3.57	8.93	17.86	14.29	50.00
ENGLISH-PIVOT	BT 1-N (T)	8.04	0.89	10.71	12.5	10.71	58.04
	BT + PAR (T)	8.93	1.79	8.04	9.82	11.61	61.61

Table 4: Human analysis for the outputs provided by the different models (in %). PAR(T) represents the model that uses paraphrases only in the training set. PAR (B) represents the model that uses paraphrases in both training and development sets. BT 1—N (T) represents the model that follows the BACK-TRANSLATION 1—N strategy and BT + PAR (T) represents the model that follows the BACK-TRANSLATION + PARAPHRASE strategy described in in Sub-section 2.2 and Table 1.

AMR Graph

```
(q / quantity
    :quant 20000
    :time (d / date-entity
        :year 2017))
```

Reference	*Foram 20 mil em 2017* (There were 20 thousand in 2017).
Baseline	*o que é 20000 ?* (what is 20000?)
Bosque-augmented	*a partir de 2017 , serão oferecidas 20 mil passagens .* (As of 2017, 20,000 tickets will be offered.)
Portuguese approach (T)	*em 2017 , serão 20000 .* (in 2017, it will be 20000)
Portuguese approach (B)	*em 2017 , o número é de 20000 .* (in 2017, the number is 20000.)
English-pivot (T) (Back-translation + Paraphrase Generation)	no total , 20000 serão gastos em 2017 . (in total 20000 will be spent in 2017.)
English-pivot (T) (Back-translation 1-N)	*em 2017 , serão 20000 000 .* (in 2017 , it will be 20000 000 .)

Figure 3: Output comparison between the reference, the baseline, the Bosque-augmented approach and the best models for each approach (including one that is trained on setting B). The first lines for each model are the sentences generated in Brazilian Portuguese, and the next ones are the corresponding English translations. Non-related n-grams are highlighted in red and a difference in verb tense is highlighted in blue.

is "*outro problema tem fundo político.*" ("Another problem has a political background.").

Models are expected to produce hallucinations as they are trained on a tiny corpus (402-4020 instances); however, generating bad paraphrases can exacerbate this behaviour. For example, we show the paraphrases generated by one approach for the reference "*teve chance suficiente para se salvar .*":

- *teve chance suficiente para se salvar .* (he had enough chance to save himself.) - original
- *você tem oportunidade suficiente para se salvar* (you have enough opportunity to save yourself)
- *você teve uma chance de se salvar* (you had a chance to save yourself)
- *para que você tenha uma chance de se salvar* (so that you have a chance to save yourself)

As we can see, most paraphrases are valid ones; however, the last one is not related to the original reference. We also show another example of the approach that generates a non-related paraphrase for the "*entra em cena a comida*".

- *entra em cena a comida .* (food comes into play.) - original
- *a comida está no local .* (the food is on the spot.)

6 Related Work

Paraphrase Generation has been widely studied in Natural Language Understanding tasks such as dialogue systems (Quan and Xiong, 2019; Okur et al., 2022), intent classification (Rentschler et al., 2022) and slot filling (Hou et al., 2021). For Natural Language Generation (NLG), we have found that using multiple references leads to a more robust evaluation (Gardent et al., 2017; Dušek et al., 2020). Besides, it has been successful in neural translation tasks (Zheng et al., 2018).

In the case of Low-Resource NLG, as far as we know, there are few works. Gao et al. (2020) proposes a paraphrase-augmented response generation framework that jointly trains paraphrasing and response generation models to improve dialogue generation. Besides, the authors describe a strategy to generate paraphrase training sets. On the other hand, Mi et al. (2022) proposes a target-side paraphrase-based data augmentation method for low-resource language speech translation.

7 Conclusion and Further Work

This study investigates the effectiveness of paraphrases for the AMR-to-text generation task in Brazilian Portuguese. Two paraphrase generation strategies were explored: one using a model trained on Brazilian Portuguese and the other using English as a pivot. The quality of generated paraphrases was evaluated using three automatic criteria, and the impact of the number of paraphrases on model performance was examined. Experiments were conducted in two settings: adding paraphrases only to the training set and adding them to both the training and development sets.

Key findings include that paraphrase generation is a powerful data augmentation strategy, outperforming the baseline and traditional data augmentation in low-resource settings. However, not all metrics respond equally, and careful selection of paraphrases is crucial. The paraphrase-extended AMR corpus showed slight improvement, with better performance seen when more paraphrases per instance were added. Regarding human evaluation, Portuguese-based models generated more valid outputs but also omitted words, while English-pivot models had lower performance and were more prone to hallucinations.

As future work, we plan to curate the AMR corpus with paraphrases and to explore new methods for generating syntax-focused paraphrases. This study acknowledges that its approach can only add a limited number of paraphrases and suggests combining it with classical data augmentation methods to expand the AMR corpus. Finally, the AMR corpus for Brazilian Portuguese and the associated code will be made publicly available at `https://github.com/msobrevillac/amr-paragen`.

8 Acknowledgments

This work was carried out at the Center for Artificial Intelligence of the University of São Paulo (C4AI - http://c4ai.inova.usp.br/), with support by the São Paulo Research Foundation (FAPESP grant 2019/07665-4) and by the IBM Corporation. This research also had the support of Coordination for the Improvement of Higher Education Personnel (CAPES) and the OPINANDO project (PRP 668).

References

Susana Afonso, Eckhard Bick, Renato Haber, and Diana Santos. 2002. Floresta sintá(c)tica: A treebank for portuguese. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*. European Language Resources Association (ELRA).

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Diedre Carmo, Marcos Piau, Israel Campiotti, Rodrigo Nogueira, and Roberto Lotufo. 2020. Ptt5: Pretraining and validating the t5 model on brazilian portuguese data. *arXiv preprint arXiv:2008.09144*.

Thiago Castro Ferreira, Iacer Calixto, Sander Wubben, and Emiel Krahmer. 2017. Linguistic realisation as machine translation: Comparing different MT models for AMR-to-text generation. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 1–10, Santiago de Compostela, Spain. Association for Computational Linguistics.

Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. 2020. Evaluating the state-of-the-art of end-to-end natural language generation: The e2e nlg challenge. *Computer Speech Language*, 59:123–156.

Sergey Edunov, Myle Ott, Marc'Aurelio Ranzato, and Michael Auli. 2020. On the evaluation of machine translation systems trained with back-translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2836–2846, Online. Association for Computational Linguistics.

Silin Gao, Yichi Zhang, Zhijian Ou, and Zhou Yu. 2020. Paraphrase augmented task-oriented dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 639–649, Online. Association for Computational Linguistics.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. Creating training corpora for NLG micro-planners. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 179–188, Vancouver, Canada. Association for Computational Linguistics.

Yutai Hou, Sanyuan Chen, Wanxiang Che, Cheng Chen, and Ting Liu. 2021. C2c-genda: Cluster-to-cluster generation for data augmentation of slot filling. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):13027–13035.

Kuan-Hao Huang, Varun Iyer, I-Hung Hsu, Anoop Kumar, Kai-Wei Chang, and Aram Galstyan. 2023.

ParaAMR: A large-scale syntactically diverse paraphrase dataset by AMR back-translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8047–8061, Toronto, Canada. Association for Computational Linguistics.

Marcio Lima Inácio, Marco Antonio Sobrevilla Cabezudo, Renata Ramisch, Ariani Di Felippo, and Thiago Alexandre Salgueiro Pardo. 2022. The amr-pt corpus and the semantic annotation of challenging sentences from journalistic and opinion texts. *SciELO Preprints*.

Fuad Issa, Marco Damonte, Shay B. Cohen, Xiaohui Yan, and Yi Chang. 2018. Abstract Meaning Representation for paraphrase detection. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 442–452, New Orleans, Louisiana. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the 2nd Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics.

Manuel Mager, Ramón Fernandez Astudillo, Tahira Naseem, Md Arafat Sultan, Young-Suk Lee, Radu Florian, and Salim Roukos. 2020. GPT-too: A language-model-first approach for AMR-to-text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1846–1852, Online. Association for Computational Linguistics.

Christian Matthiessen and John A. Bateman. 1991. *Text Generation and Systemic-Functional Linguistics: Experiences from English and Japanese*. Pinter Publishers.

Chenggang Mi, Lei Xie, and Yanning Zhang. 2022. Improving data augmentation for low resource speech-to-text translation with diverse paraphrasing. *Neural Networks*, 148:194–205.

Eda Okur, Saurav Sahay, and Lama Nachman. 2022. Data augmentation with paraphrase generation and entity extraction for multimodal dialogue system. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4114–4125, Marseille, France.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Lucas Francisco Amaral Orosco Pellicer, Paulo Pirozelli, Anna Helena Reali Costa, and Alexandre Inoue. 2022. Ptt5-paraphraser: Diversity and meaning fidelity in automatic portuguese paraphrasing. In *Computational Processing of the Portuguese Language: 15th International Conference, PROPOR 2022, Fortaleza, Brazil, March 21–23, 2022, Proceedings*, page 299–309, Berlin, Heidelberg. Springer-Verlag.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Jun Quan and Deyi Xiong. 2019. Effective data augmentation approaches to end-to-end task-oriented dialogue. In *2019 International Conference on Asian Language Processing (IALP)*, pages 47–52.

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.

Sophie Rentschler, Martin Riedl, Christian Stab, and Martin Rückert. 2022. Data augmentation for intent classification of German conversational agents in the finance domain. In *Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)*, pages 1–7, Potsdam, Germany. KONVENS 2022 Organizers.

Leonardo F. R. Ribeiro, Martin Schmitt, Hinrich Schütze, and Iryna Gurevych. 2021. Investigating pretrained language models for graph-to-text generation. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 211–227, Online. Association for Computational Linguistics.

Yves Scherrer. 2020. TaPaCo: A corpus of sentential paraphrases for 73 languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6868–6873, Marseille, France. European Language Resources Association.

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.

Marco Antonio Sobrevilla Cabezudo, Simon Mille, and Thiago Pardo. 2019. Back-translation as strategy to tackle the lack of corpus in natural language generation from semantic representations. In *Proceedings of the 2nd Workshop on Multilingual Surface Realisation (MSR 2019)*, pages 94–103, Hong Kong, China. Association for Computational Linguistics.

Marco Antonio Sobrevilla Cabezudo and Thiago Pardo. 2019. Towards a general abstract meaning representation corpus for Brazilian Portuguese. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 236–244, Florence, Italy. Association for Computational Linguistics.

Rik van Noord and Johan Bos. 2017. Neural semantic parsing by character-based translation: Experiments with abstract meaning representations. *Computational Linguistics in the Netherlands Journal*, 7:93–108.

Gregory César Valderrama Vilca and Marco Antonio Sobrevilla Cabezudo. 2017. A study of abstractive summarization using semantic representations and discourse level information. In *Text, Speech, and Dialogue*, pages 482–490. Springer-Verlag.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.

Renjie Zheng, Mingbo Ma, and Liang Huang. 2018. Multi-reference training with pseudo-references for neural translation and text generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3188–3197, Brussels, Belgium. Association for Computational Linguistics.

Jianing Zhou and Suma Bhat. 2021. Paraphrase generation: A survey of the state of the art. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5075–5086, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Appendix

A.1 Model Hyperparameters

Training Models are generated by fine-tuning the Portuguese T5 (PTT5)[12] on our diverse paraphrase-based corpora. We use AdamW optimizer with a learning rate of 5e-4, a maximum source and target length of 120 and 80 tokens, respectively, a batch size of 8, and a gradient accumulation of 4. The model trains by 12 epochs and is evaluated after each epoch. We use perplexity as evaluation criteria, and the training is halted if the model does not improve after 4 epochs.

Decoding For the paraphrase generation, we use a batch size of 32 and a beam size of 20. Also, we use a top_k of 120 and a top_p of 0.98, and early stopping with a maximum length of 80 tokens. For text generation, we use a beam size of 5, a maximum target length of 80 with early stopping, an n-gram length that can be repeated is set to 1, a repetition penalty of 2.5, and a length penalty of 1.0.

A.2 Results

Figures 4 and 5 show the performance changes for BLEU selection criteria when more paraphrases per instance are added in T and B setting, respectively.

Figures 6 and 7 presents the results for METEOR, chrF++ and BERT scores per selection criterion and per number of selected paraphrases in the T and B settings. The results reported are obtained on the development set.

A.3 Models for Human Evaluation

- Data augmentation by Parsing (Bosque-augmented in Table 1)
- Portuguese approach (T): We select one of the best models for setting T. In particular, the selected one uses METEOR as criterion selection and 5 paraphrases.
- Portuguese approach (B): We select one of the best models on the setting B. The selected one includes METEOR as criterion selection and 9 paraphrases.
- English-pivot approach (Back-translation): We select one of the best models for the setting T. The selected one includes TER as criterion selection and 8 paraphrases.

[12] Available at https://huggingface.co/unicamp-dl/ptt5-base-portuguese-vocab.

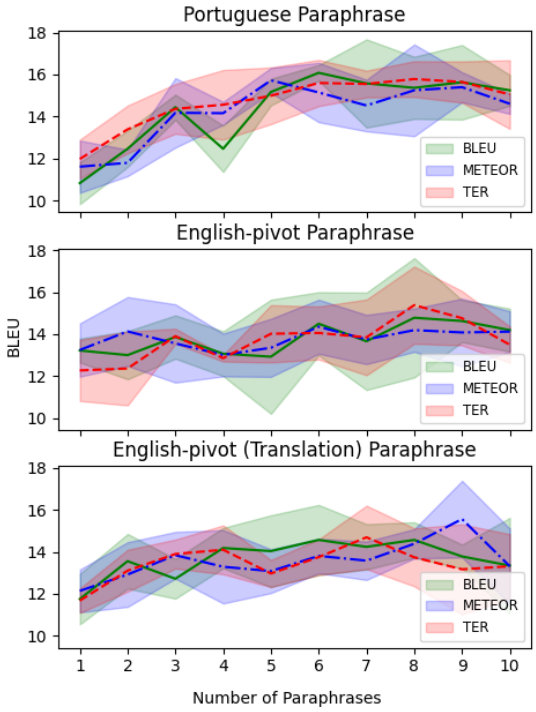

Figure 4: BLEU scores per selection criterion and per number of selected paraphrases in the T setting. Results are shown on the development set.

- English-pivot approach (Back-translation + Paraphrase): We select one of the best models for setting T. The selected one includes METEOR as criterion selection and 9 paraphrases.

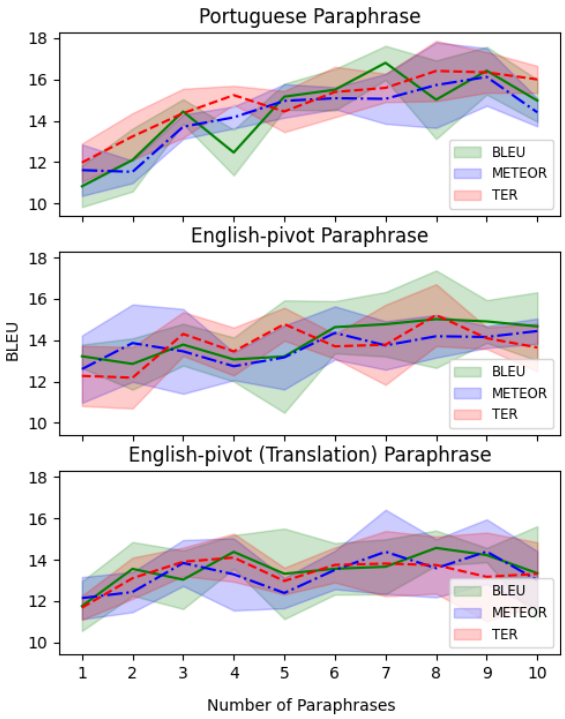

Figure 5: BLEU scores per selection criterion and per number of selected paraphrases in the B setting. Results are shown on the development set.

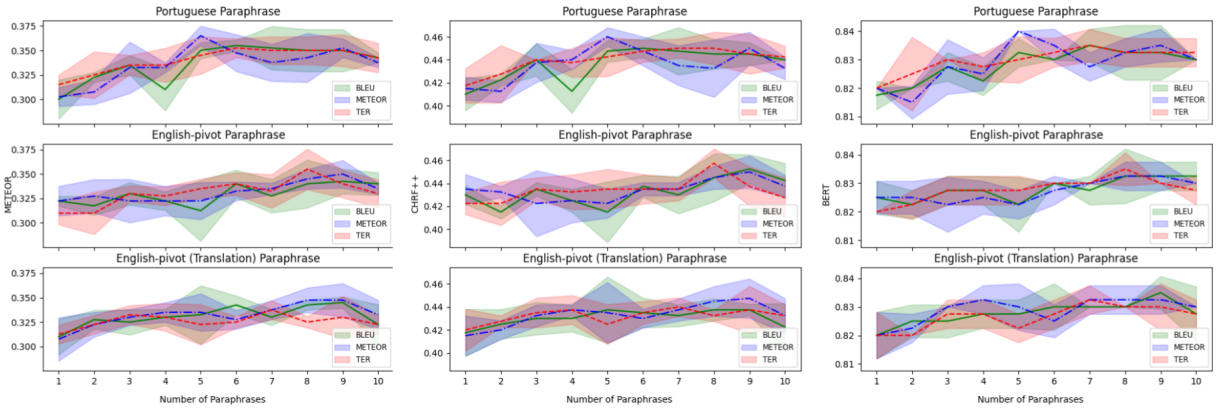

Figure 6: METEOR, chrF++ and BERT scores per selection criterion and per number of selected paraphrases in the T setting. Results are shown on the development set.

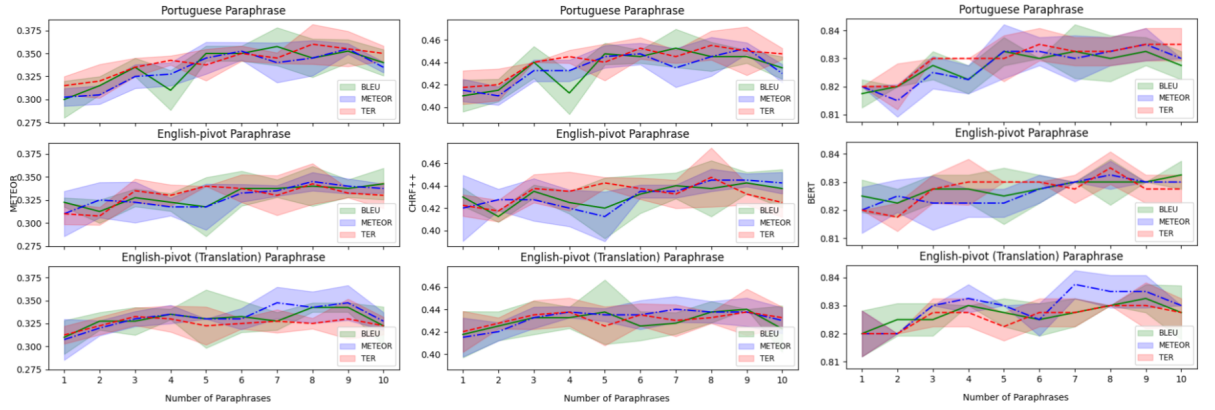

Figure 7: METEOR, chrF++ and BERT scores per selection criterion and per number of selected paraphrases in the B setting. Results are shown on the development set.