Back-Translation as Strategy to Tackle the Lack of Corpus in Natural Language Generation from Semantic Representations

Marco Antonio Sobrevilla Cabezudo♣ **Simon Mille**♠
Thiago Alexandre Salgueiro Pardo♣
♣ Interinstitutional Center for Computational Linguistics (NILC)
Institute of Mathematical and Computer Sciences, University of São Paulo. São Carlos/SP, Brazil
♠ Universitat Pompeu Fabra. Barcelona, Spain
msobrevillac@usp.br, simon.mille@upf.edu, taspardo@icmc.usp.br

Abstract

This paper presents an exploratory study that aims to evaluate the usefulness of back-translation in Natural Language Generation (NLG) from semantic representations for non-English languages. Specifically, Abstract Meaning Representation and Brazilian Portuguese (BP) are chosen as semantic representation and language, respectively. Two methods (focused on Statistical and Neural Machine Translation) are evaluated on two datasets (one automatically generated and another one human-generated) to compare the performance in a real context. Also, several cuts according to quality measures are performed to evaluate the importance (or not) of the data quality in NLG. Results show that there are still many improvements to be made but this is a promising approach.

1 Introduction

Natural Language Generation (NLG) is the research area that aims to give to the computers the ability to generate texts in human language from some underlying representation of information (Reiter and Dale, 2000). This area has gained relevance in the Natural Language Processing community and in the industry in the last years.

There are several works and efforts in NLG for English.[1] Recently, shared-tasks focused on NLG from semantic representations have gained the attention of the NLG community. Thus, several representations have emerged for attending different contexts. For example, the RDF-based representation presented by Gardent et al. (2017) in its WebNLG challenge, the dialog-act-based representation presented by Novikova et al. (2016), and Abstract Meaning Representation (Banarescu et al., 2013).

There are not as many works for languages other than English: in 2018, the first multilingual surface realization was proposed (Mille et al., 2018). This event proposed two tasks, one focused on reordering a dependency tree and generating inflected words (called shallow track), and the other one focused on generating sentences from a deep-syntax representation similar to a semantic representation (called deep track). It is important to note that while NLG methods were evaluated in corpora for ten different languages in the shallow track, the deep track was limited to evaluating NLG methods on three languages (English, Spanish, and French). The fact that there are less datasets in the deep track is directly related to the higher complexity of the conversion compared to the shallow track, for which a superficial processing (basically order randomization) is sufficient.

Among the efforts to build or adapt semantic representations for non-English languages, it is possible to cite Abstract Meaning Representation (AMR) as an example. Although AMR was not born as an interlingua, several works have tried to use it in that way to annotate sentences in other languages like Chinese and Czech (Xue et al., 2014), Italian, Spanish, and German (Damonte and Cohen, 2018) and Brazilian Portuguese (Anchiêta and Pardo, 2018). Other works have tried to adapt the English AMR guidelines to Spanish and Brazilian Portuguese with some success (Migueles-Abraira et al., 2018; Sobrevilla Cabezudo and Pardo, 2019). However, most of these works report a small number of AMR-annotated sentences (compared to the English corpus) and are restricted to some domains like tales ("The Little Prince"). To the best of our knowledge, the only AMR-annotated corpus comparable (in terms of size) to the English corpus[2] is the

[1] Most of the work may be found at https://aclweb.org/anthology/sigs/siggen/.

[2] Available at https://catalog.ldc.upenn.

Proceedings of the 2nd Workshop on Multilingual Surface Realisation (MSR 2019), pages 94–103
Hong Kong, China, November 3rd, 2019. ©2019 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17

Chinese corpus, containing 10,149 annotated sentences in its first version.[3]

This difficulty to get large corpora with this kind of annotation (due to the difficult and expensive annotation task that it represents) constrains the development of research in other languages. Consequently, it is difficult to achieve the same performance as in English or to replicate state-of-the-art works.

In general, a strategy to overcome the lack of corpora is to translate English corpora to non-English ones. This involves the use of Machine Translation (MT) systems, leveraging the good performance obtained by MT systems that work on English as a source or target language. However, the quality of the translations depends on the language pair. Thus, it is important to filter out some translations according to their quality. This may be accomplished by applying back-translation and performing a quality evaluation (using some quality measures like BLEU or METEOR) in English. In Machine Translation, Back-translation consists of translating a target sentence (in our case, Portuguese) into a source language (in our case, English).

This approach has shown good performance in some classification tasks like Sentiment Analysis and Word Sense Disambiguation (Klinger and Cimiano, 2015; Monsalve et al., 2019). Furthermore, Monsalve et al. (2019) show that despite the introduction of sentences with low quality (according to quality measures), the performance of the classifiers continues improving. Also, this approach has been successful in the context of neural machine translation (Sennrich et al., 2016). In the case of NLG from semantic representations, it would be expected that quality is critical since low-quality sentences may lead to models learning incorrect language. Additionally, other issues that may impact the performance of this task are the translation of the semantic representation and the alignments between language pairs.

In this context, this paper presents an exploratory study that aims to evaluate the usefulness of back-translation in NLG from semantic representations for non-English languages. Specifically, AMR and Brazilian Portuguese (BP) are chosen as semantic representation and language, respectively. Two methods (SMT-based and NMT-based) are evaluated on two datasets (one automatically generated and one human-generated) in order to compare the performance in a real context. Also, several cuts[4] according to quality measures are performed to evaluate the importance (or not) of the data quality in NLG.

This paper is organized as follows: §2 describes some work that applied back-translation to produce corpus in non-English languages. Then, §3 introduces Abstract Meaning Representation (our target representation) and works performed for English and non-English languages on it. Our methodology for generating corpus and the experiments performed are presented in §4. Furthermore, §5 contains the results and a discussion about the results. Finally, the conclusions and future work are presented in §6.

2 Related Work

Several works have proven the usefulness of translating corpora to increase the dataset size and improve the performance of their models. For example, Klinger and Cimiano (2015) used Phrase-based MT and some quality estimation measures to build a corpus with the best translations and use it in Sentiment Analysis. Misu et al. (2012) and Gaspers et al. (2018) explored back-translation in Natural Language Understanding systems using different measures. Misu et al. (2012) showed that BLEU is not a good quality measure and Gaspers et al. (2018) used measures from alignments, machine translation and language models to select the best sentences to be included in the corpus.

Monsalve et al. (2019) also explored some quality measures (BLEU and METEOR) to select the best sentences and build a non-English corpus for Reading Comprehension and Word Sense Disambiguation. Among the results, they showed that despite the introduction of low-quality sentences, the performance is still continually improving. However, their main goal was to get a well-translated corpus and not to get the best results in both tasks.

About the tasks that involve language generation, it is noted that back-translation has been widely, and successfully, used in neural machine translation. The aim was to generate synthetic source sentences to increase the parallel training dataset (Sennrich et al., 2016; Edunov et al.,

edu/LDC2017T10.

[3] Available at https://catalog.ldc.upenn.edu/LDC2019T07

[4] A cut consists of a set of sentences of the corpus with a similar quality.

2018). Also, Prabhumoye et al. (2018) applied back-translation to perform style transfer with good results.

Concerning the described work, a question emerges: How can back-translation influence NLG from semantic representations? It is important to note that not only English sentences will be translated into BP ones, but its corresponding semantic representations will be translated to handle representations for Portuguese. Thus, several issues related to alignments may affect the performance (in addition to the quality translation). The following sections show the influence of back-translation in NLG.

3 Abstract Meaning Representation

Abstract Meaning Representation (AMR) is a semantic formalism that aims to encode the meaning of a sentence with a simple representation in the form of a directed rooted graph (Banarescu et al., 2013). This representation includes information about semantic roles, named entities, spatial-temporal information, and co-references, among other information. AMR-annotated sentences may be represented using logic forms, PENMAN notation, and graphs (Figure 1).

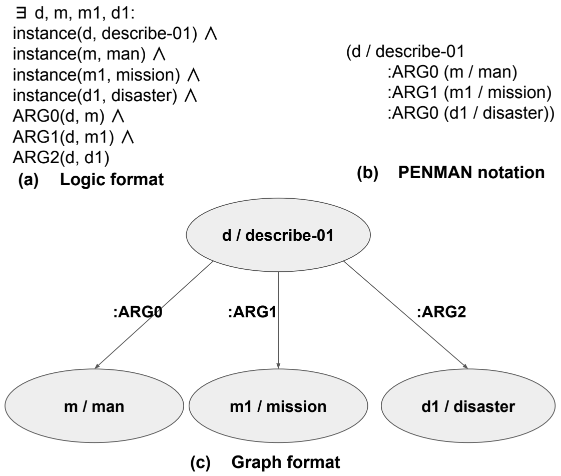


Figure 1: AMR example for the sentence "The man described the mission as a disaster"

AMR has gained relevance in the research community due to its attempt to abstract away from syntactic idiosyncrasies[5] and its wide use of other comprehensive linguistic resources, such as PropBank (Palmer et al., 2005).[6]

The current AMR-annotated corpus for English contains 39,260 sentences. Some efforts have been performed to build a corpus for Non-English languages leveraging the alignments and the parallel corpora that exist and trying to consider AMR an interlingua (Xue et al., 2014; Damonte and Cohen, 2018; Anchiêta and Pardo, 2018). Other works tried to adapt the AMR guidelines to other languages (Migueles-Abraira et al., 2018; Sobrevilla Cabezudo and Pardo, 2019).

For Brazilian Portuguese, there are two AMR-annotated corpora, one automatically built from the alignments between the sentences of the "The Little Prince" book in English and Portuguese (Anchiêta and Pardo, 2018), and the other one that contains news texts sentences manually annotated using an adaptation of the AMR guidelines (Sobrevilla Cabezudo and Pardo, 2019). The lexical resource used to annotate some concepts in both corpora was the Verbo-Brasil (Duran and Aluísio, 2015), which is analogous to the PropBank lexical repository.

Concerning the Little Prince corpus, the style of the sentences reflects a rather unusual genre (tales) and the vocabulary is restricted to the story. Also, this corpus only contains 1,527 annotated sentences. In relation to the second corpus, although annotated sentences belong to news texts, the corpus size is still small, containing 299 annotated sentences. Besides, only the sentences that contain lexical units found in Verbo-Brasil were annotated, excluding those that are not represented in it. As a result, the current limitations of the corpora in terms of genre, size and richness of annotations hinders the development or adaptation of methods that target general purpose and semantics-oriented NLG tasks.

4 Methodology

In order to deal with the lack of corpus in the AMR-to-Text generation task, firstly, a corpus generation process was developed to build an AMR dataset for Brazilian Portuguese (BP) from an English one. This process involved back-translation and some MT measures to select the high-quality BP sentences that are comprised in the dataset. Secondly, several experiments using well-known methods for AMR-to-Text generation were used to evaluate the performance of each method, measure the influence of the qual-

[5] In Figure 1, there are other possible sentences like "The man's description about the mission: a disaster" that could generate the same representation despite syntactic difference.

[6] In Figure 1, the frameset "describe-01" belongs to the PropBank lexical repository.

ity of the translated sentences, determine the most useful MT measure to select high-quality BP sentences, and verify if the results obtained with the translated datasets are comparable with a curated dataset (gold dataset).

4.1 Corpus generation

The corpus generation was divided in two phases: the first one focused on filtering and splitting the original English corpus and the second one focused on translating the concepts of the AMR graph according to the alignments between English and Portuguese tokens in the sentences.[7]

4.1.1 Corpus Filtering and Splitting

The corpus filtering phase consisted of the following steps:

- select the sentences in the English corpus. This step focused on selecting English sentences which have a similar size to those annotated in the BP corpus, i.e., 23 tokens maximum. The number of sentences after this step was 27,464.

- apply the back-translation. This strategy consisted of translating English sentences into BP sentences and then translating those BP sentences into English sentences to measure the quality of the translation in Portuguese via English (since the Portuguese references did not exist). To achieve this goal, the Machine Translation model provided by Google Translate API was used;[8]

- evaluate the sentences according to automatic quality measures. In the same way as Monsalve et al. (2019), F[9] and METEOR were used to automatically measure the quality of the sentences. The quality scores of BP sentences were calculated applying the quality measures to their respective English sentences. This generated a dataset for each quality measure (F and METEOR), where each instance of each dataset comprised the BP sentence and its respective quality score, aiming to define some sets.

- define the development and test sets.[10] To achieve this step, firstly, a set of sentences with a quality higher than the mean plus one standard deviation of all sentences according to each quality measure was selected, generating two sub-sets. Secondly, the sentences included in the intersection of the sub-sets were selected in order to filter the highest-quality sentences. Finally, the development and test sets were defined as 25% of the sentences in the intersection. In total, 1,073 sentences were used for development and test sets, respectively.

- define cuts according to quality measures. Finally, the remaining sentences in the translated BP datasets were sorted decreasingly according to each quality measure. Then, five cuts of 5,000 sentences each were performed for each quality measure, thus, the first cut contained the 5,000 best sentences according to one quality measure and the last cut contained the 5,000 worst sentences. Table 1 shows the mean and standard deviation (std) of each cut for each dataset (for quality measure). These datasets and cuts constitute the training set.

Measure		1	2	3	4	5
F	mean	0.92	0.74	0.60	0.32	0.00
	std	0.07	0.04	0.04	0.20	0.00
METEOR	mean	0.98	0.58	0.48	0.41	0.30
	std	0.03	0.09	0.01	0.01	0.08

Table 1: Statistics of all cuts performed in the AMR Corpus

4.1.2 Target Corpus Generation

In order to get the AMR-annotated corpus in Brazilian Portuguese (BP), it was also necessary to convert the English AMR graphs into Portuguese ones.

This conversion was performed leveraging the alignments between English and BP sentences and the alignments between the English sentences and the AMR graphs provided by the corpus. Thus, Fast Align (Dyer et al., 2013) was applied to obtain the alignments between the sentences. Then,

[7] In this work, the LDC2016E25 corpus was used to perform all experiments.

[8] Google Translate API was used due to the good results obtained in Machine Translation. Eventually, other MT systems could be used. Available at `https://cloud.google.com/translate/`.

[9] In this work, F measure is defined as the harmonic mean of BLEU and ROUGE scores.

[10] In this step, both the use of the mean plus one standard deviation and the 25% of the intersection were used as a threshold empirically defined.

these alignments were used to change the alignments (the numbers) in the AMR graph and to replace the English concepts by their respective BP concepts.

It is worth noting that not all concepts in the AMR graph were changed as some of them were not aligned in the corpus. Also, some concepts belonging to PropBank (PropBank framesets) were replaced by their corresponding framesets in BP using Verbo-Brasil (Duran and Aluísio, 2015). PropBank concepts (framesets) that could not be mapped to Verbo-Brasil framesets were kept in their English version. In general, 825 of 3,965 framesets were translated, representing 20.81% of the framesets. All other aligned English concepts were replaced by their corresponding BP ones in the sentence-alignments, excepting AMR-defined framesets, modal verbs, and AMR-defined entities. Besides, some rules were applied to change some concepts like ly-adverbs.

Concerning the alignment types, we note that there were some issues in "1-n" and "n-1" alignments. In the case of "n-1" alignments ("n" English tokens corresponding to 1 BP token), all "n" concepts were replaced by the same one concept, and in the case of "1-n" alignments, the one English concept was replaced by the concatenation of all "n" BP concepts. Figure 2 shows the pipeline of the AMR graph translation. Tokens and numbers in bold are the ones which were translated.

4.2 Experiments

Experiments were performed using the Statistical Machine Translation (SMT) and Neural Machine Translation (NMT) methods provided by Castro Ferreira et al. (2017) to compare how each method behaved in the evaluated context.

The SMT method used the same parameters proposed by Castro Ferreira et al. (2017) and a 5-gram language model trained on the BP corpus provided by Hartmann et al. (2017). Also, the AMR graph pre-processing comprised a compression and a pre-ordering step without delexicalization (described as -Delex+Compress+Preorder in the original paper) as this configuration got one of the best results.

The NMT method used similar parameters to Castro Ferreira et al. (2017). The encoder was bidirectional RNN with GRU, each with a 1024D hidden unit. Source and target word embeddings were 300D each and both were trained jointly with the model. Also, the vocabulary was shared. All weights were initialized using a Xavier uniform, which draws samples from a uniform distribution within a range. The decoder RNN also used GRU with an attention and a copy mechanism (Bahdanau et al., 2015).

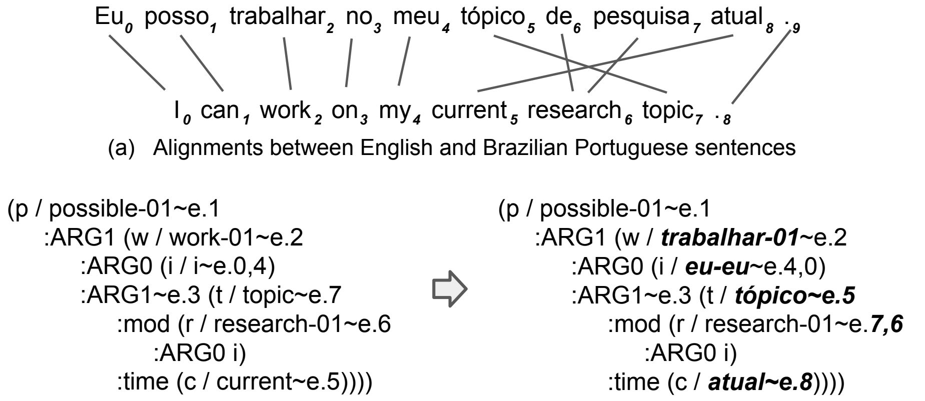
We applied dropout with a probability of 0.3. Models were trained using the Adadelta optimizer with a learning rate of 1.0 and a learning rate decay of 0.7 every 5 epochs, and mini-batches of size 64. We applied early stopping for model selection based on accuracy and perplexity scores so that if a model does not improve on the development set for more than 25 epochs, training is halted.

Besides, the AMR graph pre-processing was composed of a delexicalization and a pre-ordering step without compression (described as +Delex-Compress+Preorder in the original paper) as this configuration got one of the best results.

These methods were trained according to two configurations and evaluated using the automatically generated development set described in §4.1.1. The two configurations are described as follows:

- training on each cut described in §4.1.1 independently. It was expected that the performance decreases in each cut as the cut quality decreases as well.
- training on cut 1 plus each cut included progressively. At the beginning, the training set was composed by the cut 1. Then, a lower quality cut was added to the training set at each training phase until all the cuts were included. The goal of this experiment was to evaluate how the method performance varied when lower quality data was inserted into the training set.

It is worth noting that each configuration was performed using the cuts generated by F and METEOR to evaluate the quality measure in the corpus selection task. The test was performed on the automatically generated test set described in §4.1.1. In order to compare the results in a real context, the methods were also evaluated on the AMR-annotated BP dataset described in §3. Similar to Castro Ferreira et al. (2017), we used BLEU (Papineni et al., 2002), METEOR (Lavie and Agarwal, 2007) and TER (Snover et al., 2006) as metrics to evaluate fluency, adequacy and post-editing efforts of the models, respectively.


Eu0 posso1 trabalhar2 no3 meu4 tópico5 de6 pesquisa7 atual8 .9
I0 can1 work2 on3 my4 current5 research6 topic7 .8
(p / possible-01~e.1
:ARG1 (w / work-01~e.2
:ARG0 (i / i~e.0,4)
:ARG1~e.3 (t / topic~e.7
:mod (r / research-01~e.6
:ARG0 i)
:time (c / current~e.5))))
(p / possible-01~e.1
:ARG1 (w / trabalhar-01~e.2
:ARG0 (i / eu-eu~e.4,0)
:ARG1~e.3 (t / tópico~e.5
:mod (r / research-01~e.7,6
:ARG0 i)
:time (c / atual~e.8))))

(a) Alignments between English and Brazilian Portuguese sentences

(b) Conversion of AMR graph from English to Brazilian Portuguese according to alignments

Figure 2: Pipeline for the translation of the AMR corpus

5 Results and discussion

5.1 Overview

Figures 3, 4, and 5 show the results obtained by the NMT and SMT approaches using cuts generated by F and METEOR and evaluated on the development, test and gold test sets for each metric (BLEU, METEOR, and TER). Bars show the results of the first configuration (each cut independently) and lines represent the results of the second experiment (training on cut 1 plus each cut included progressively).

In general, results show that the performance on development and test sets increased while more data (no matter that was of lower quality) was incorporated (except on the last cut). On the other hand, the performance decreased when a lower quality cut was used as training data. Also, results on the curated test[11] (also called gold test) showed that there are many improvements to perform in order to achieve similar results to the development and test sets. In this set, BLEU and TER were the most affected metrics as values between 0.02 and 0.04 were obtained for BLEU (Figure 3), and 0.73 and 0.92 were obtained for TER (Figure 5).

5.2 Discussion

Quality or Quantity? At first glance, quantity seemed to be more important than quality. Also, in the case of NMT, quantity seemed to be still more important than in the case of SMT. A detail to note is that the increase in the performance was lower when the latest cuts (with lower quality) were incorporated into the training set. Besides, the performance decreased when the latest cuts were incorporated in some cases (cut 5 in Figure 3). Thus, a different analysis is required to check if the quantity is more important than quality as the size of the training set could hide some problems caused by the lower quality cut.

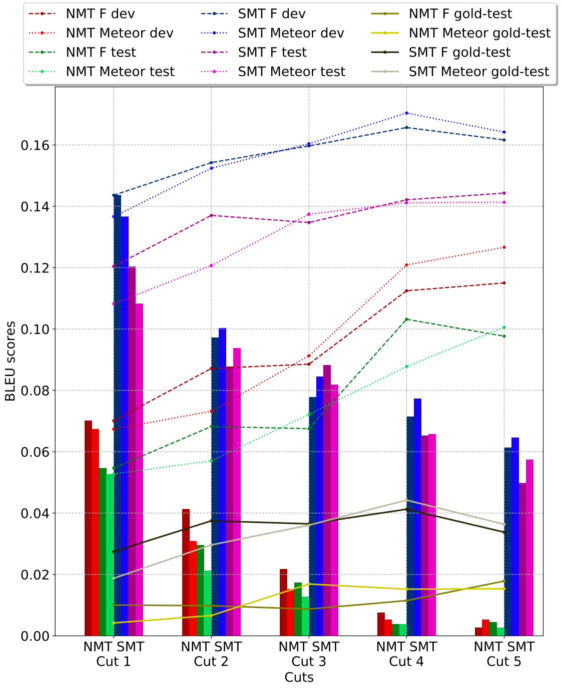


Figure 3: BLEU scores

In order to perform this analysis, four training sets were built. Each training set was composed by the cut 1 and another different cut (from highest to lowest quality cuts). Figures 6, 7, and 8 show the results of this experiment for each metric. Bars show the results on the development and test sets, and lines represent the results on the gold test set.

In this case, results on development set did not show a decrease in performance. However, results

[11] The curated test was composed by the manually-annotated 299 BP sentences.

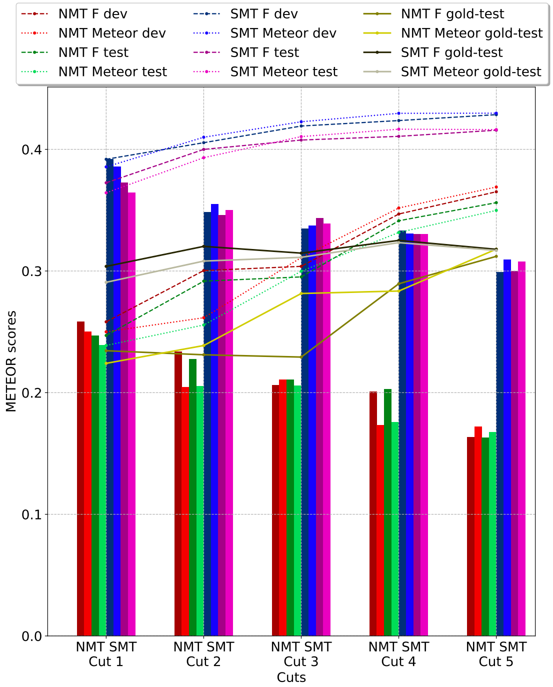


Figure 4: METEOR scores

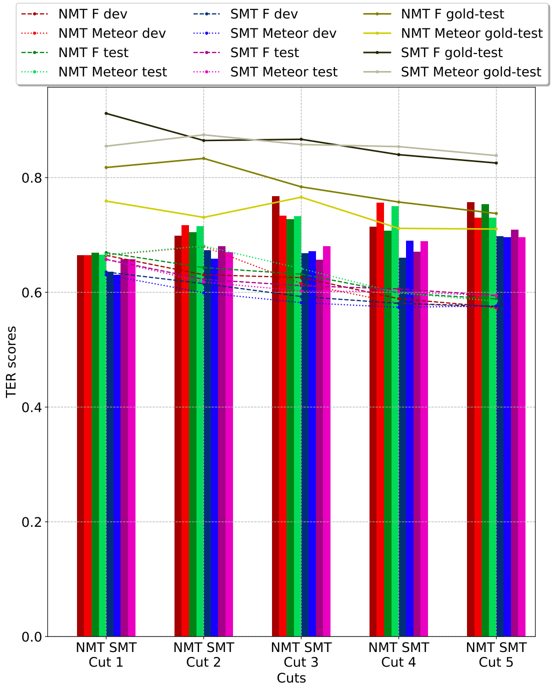


Figure 5: TER scores

on test set showed that the performance decreased when lower quality sets were incorporated (see cut 1 + cut 4 and cut 1 + cut 5 in Figures 6 and 7). In the case of the gold test set, results showed slight increases and decreases in performance, hindering the analysis. Similarly, TER results showed slight increases and decreases in performance. A possible explanation to the slight (or no) variation in the results obtained was that Google Translate API usually produced good translations, and, although some translations could show low scores in terms of F or METEOR, they could be paraphrases or sentences with synonyms of some words of the original sentences. Thus, it is expected that in cases of languages where machine translation systems present worse performance, this analysis will show more useful information to select better cuts.

Finally, from a quality perspective, it is important to note that it would be useful considering cuts with higher quality to perform better corpus analysis. However, another problem emerges in the context of semantic representations. Alignments between English and BP sentences may not be "1-1" and this could make the correct generation of semantic representations for some sentences more difficult. Thus, an interesting research would consist in evaluating how alignments may affect the performance of the methods in this context.

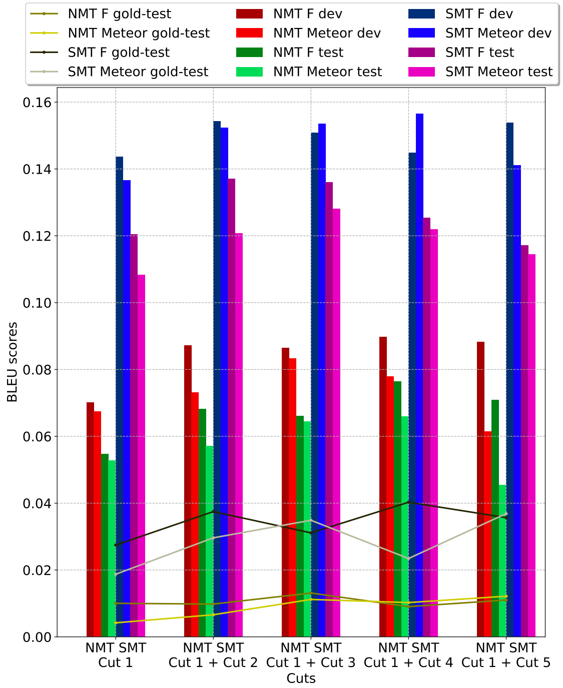


Figure 6: BLEU scores for the cut 1 plus the other cuts

What is the best quality measure? Following the idea that Google Translate API generates paraphrases or sentences with synonyms of some words of the original sentence, it would be expected that METEOR shows better results (due

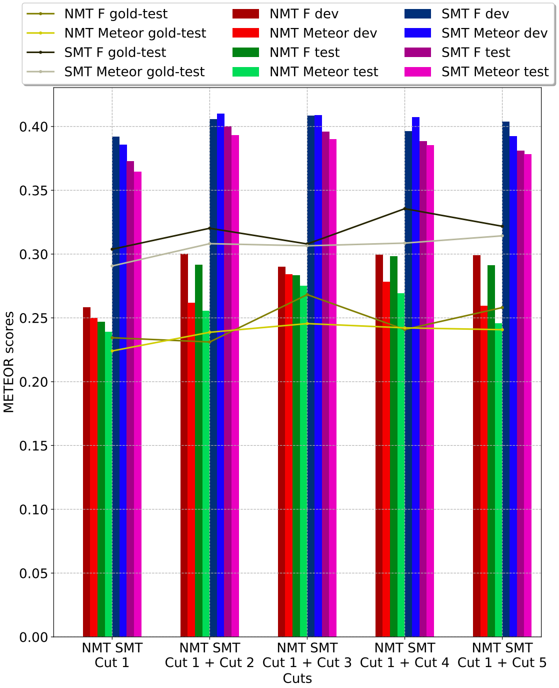


Figure 7: METEOR scores for the cut 1 plus the other cuts

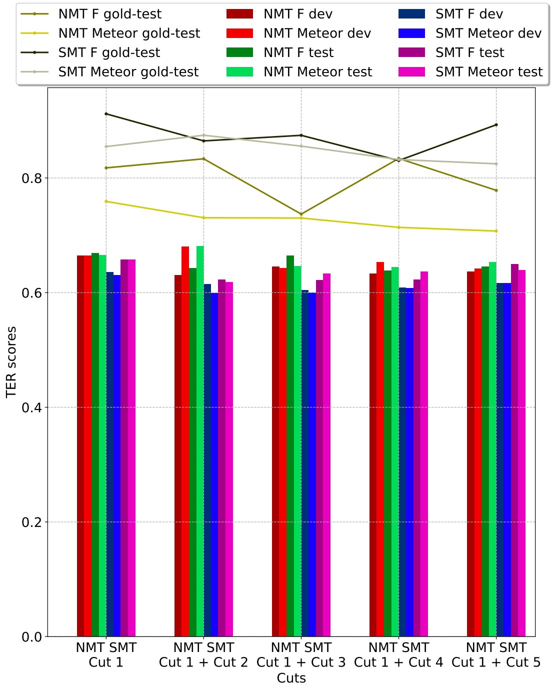


Figure 8: TER scores for the cut 1 plus the other cuts

to the fact that METEOR considers synonyms and stems). However, analysing the test set, it is possible to see that F produced stable and better results in BLEU and METEOR metrics (see Figure 6 and 7). In the case of TER, both F and METEOR produced mixed results (Figure 8). Besides, in the gold test set, F also produced better results than METEOR, excepting in the TER metric (Figure 8).

How is each approach affected? As expected, SMT outperformed NMT on the three sets in most cases. In the case of TER, NMT outperformed SMT on the gold test set (Figure 5). In the case of development and test sets, the difference between results was small and decreased while more data was incorporated into the training set, regardless of their quality. Also, the tendency of TER values to vary was lower than for METEOR and BLEU. On the other hand, it is important to highlight the greater trend of NMT to increase when more data was incorporated.

Are the results comparable in curated datasets? In general, the results in the BP corpus (gold-test set) were quite worse than in the test and development sets for all metrics, excepting METEOR. Although the METEOR values were low, the difference between these values and the values obtained in the development and test sets was not as big (principally considering NMT) as the other metrics. Also, the values were close to the ones obtained with the NMT approach in the last cut (Figure 4).

There were two reasons that we hypothesize that could lead to these results. Firstly, the number of words in the gold test set that were not in the training vocabulary. Even though the BP AMR corpus and the original AMR corpus were focused on general domains, it is necessary to analyze the overlap between them. The other problem was related to alignment types. There were several translated sentences in the corpus that present alignments “1-n”, “n-1”, or “1-n and n-1” and the generation of their respective semantic representations presented some issues like the concatenation between two tokens (token “eu-eu” in Figure 2). This could generate more sparsity and decrease the performance of the methods.

6 Conclusions and Future Work

This paper presented an exploratory study that aimed to evaluate the usefulness of back-translation in NLG from semantic representations. The followed pipeline showed how to perform a simple back-translation process in an NLG context and this may be applied to any language. Results showed that quantity is important when Machine Translation systems are good enough. However, quality may be critical in the context of low-resource languages, when translations may be too poor.

It is worth noting that the selection of cuts to be included in the training set has to be performed carefully. In this study, we proposed to analyze the performance considering 5 cuts and the last cut did not contribute positively to the performance (due to the poor quality scores). However, a deep analysis of the use of cuts may be performed to better determine the number of cuts that allow for filtering out the worst instances in order to improve the performance of the models and provide a high-quality translated dataset.

On the other hand, there are several improvements to be made to achieve similar results in real (curated) datasets. It is necessary to analyze the alignments and out-of-vocabulary words. Thus, a research direction is to analyse how these issues affect the NLG task in non-English languages. Also, we plan to explore the text generation in a curated dataset as a domain adaptation problem.

Acknowledgments

The authors are grateful to CAPES and USP Research Office for supporting this work, and would like to thank NVIDIA for donating the GPU. This work is part of the OPINANDO project (more details can be found in https://sites.google.com/icmc.usp.br/opinando/), and has been partly supported by the European Commission in the framework of the H2020 Programme via contracts to UPF, with the numbers 779962-RIA, 700475-IA, 7000024-RIA, and 645012–RIA.

References

Rafael Anchiêta and Thiago Pardo. 2018. Towards AMR-BR: A SemBank for Brazilian Portuguese language. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, pages 974–979, Miyazaki, Japan. European Languages Resources Association.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceddings of the 3rd International Conference on Learning Representations*, San Diego, CA, USA.

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.

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.

Marco Damonte and Shay B. Cohen. 2018. Cross-lingual abstract meaning representation parsing. 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 1146–1155, New Orleans, Louisiana. Association for Computational Linguistics.

Magali Sanches Duran and Sandra Maria Aluísio. 2015. Automatic generation of a lexical resource to support semantic role labeling in Portuguese. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 216–221, Denver, Colorado. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. Creating training corpora for 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.

Judith Gaspers, Penny Karanasou, and Rajen Chatterjee. 2018. Selecting machine-translated data for quick bootstrapping of a natural language understanding system. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 137–144, New Orleans - Louisiana. Association for Computational Linguistics.

Nathan Hartmann, Erick Fonseca, Christopher Shulby, Marcos Treviso, Jéssica Silva, and Sandra Aluísio. 2017. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In *Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology*, pages 122–131, Uberlândia, Brazil. Sociedade Brasileira de Computação.

Roman Klinger and Philipp Cimiano. 2015. Instance selection improves cross-lingual model training for fine-grained sentiment analysis. In *Proceedings of the 19th Conference on Computational Natural Language Learning*, pages 153–163, Beijing, China. 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.

Noelia Migueles-Abraira, Rodrigo Agerri, and Arantza Diaz de Ilarraza. 2018. Annotating abstract meaning representations for Spanish. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*, pages 3074–3078, Miyazaki, Japan. European Languages Resources Association.

Simon Mille, Anja Belz, Bernd Bohnet, Yvette Graham, Emily Pitler, and Leo Wanner. 2018. The first multilingual surface realisation shared task (SR'18): Overview and evaluation results. In *Proceedings of the 1st Workshop on Multilingual Surface Realisation*, pages 1–12, Melbourne, Australia. Association for Computational Linguistics.

Teruhisa Misu, Etsuo Mizukami, Hideki Kashioka, Satoshi Nakamura, and Haizhou Li. 2012. A bootstrapping approach for SLU portability to a new language by inducting unannotated user queries. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4961–4964. IEEE.

Fabricio Monsalve, Kervy Rivas Rojas, Marco Antonio Sobrevilla Cabezudo, and Arturo Oncevay. 2019. Assessing back-translation as a corpus generation strategy for non-English tasks: A study in reading comprehension and word sense disambiguation. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 81–89, Florence, Italy. Association for Computational Linguistics.

Jekaterina Novikova, Oliver Lemon, and Verena Rieser. 2016. Crowd-sourcing NLG data: Pictures elicit better data. In *Proceedings of the 9th International Natural Language Generation conference*, pages 265–273, Edinburgh, UK. Association for Computational Linguistics.

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.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 866–876, Melbourne, Australia. Association for Computational Linguistics.

Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press, New York, NY, USA.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving Neural Machine Translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas*, pages 223–231. Association for Machine Translation in the Americas.

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.

Nianwen Xue, Ondřej Bojar, Jan Hajič, Martha Palmer, Zdeňka Urešová, and Xiuhong Zhang. 2014. Not an interlingua, but close: Comparison of English AMRs to Chinese and Czech. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, pages 1765–1772, Reykjavik, Iceland. European Languages Resources Association.