UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Natural language generation from abstract meaning representation for brazilian portuguese

Marco Antonio Sobrevilla Cabezudo

Tese de Doutorado do Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional (PPG-CCMC)

ICMC USP
SÃO CARLOS

Marco Antonio Sobrevilla Cabezudo

Natural language generation from abstract meaning representation for brazilian portuguese

Thesis submitted to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP – in accordance with the requirements of the Computer and Mathematical Sciences Graduate Program, for the degree of Doctor in Science. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Prof. Dr. Thiago Alexandre Salgueiro Pardo

USP – São Carlos
April 2023

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

S677n

Sobrevilla Cabezudo, Marco Antonio
Natural language generation from abstract meaning representation for brazilian portuguese / Marco Antonio Sobrevilla Cabezudo; orientador Thiago Alexandre Salgueiro Pardo. -- São Carlos, 2023.
199 p.

Tese (Doutorado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2023.

1. Natural Language Generation. 2. Abstract Meaning Representation. 3. Low-Resource Setting. 4. Brazilian Portuguese. I. Salgueiro Pardo, Thiago Alexandre, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Marco Antonio Sobrevilla Cabezudo

Geração de linguagem natural por meio de representações semânticas abstratas para o português do brasil

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Doutor em Ciências – Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientador: Prof. Dr. Thiago Alexandre Salgueiro Pardo

USP – São Carlos
Abril de 2023

This work is dedicated to my lovely family, mi wife Katherine Vasquez, my daughter Natalia Sobrevilla, my mother Angela Cabezudo, my sister Mayra Sobrevilla, my mamita Victoria Carrizales, and my uncle Cesar Cabezudo.

Besides, I dedicate this work to all future PhD students who feel they have lost their way along their own research. You will arrive at the correct place, be patient and keep researching!

ACKNOWLEDGEMENTS

First of all, I want to thank God for everything he has done for me and in me during this time. I am sure that without his presence, I would not have lived all that I lived in these years.

To my life partner, my wife and my best friend, Katherine Vasquez, who supported me unconditionally all this time. You resisted so many work nights and times in which, despite being present, I was a bit absent. I also would like to thank my beautiful daughter, Natalia, you are the best thing that happened to me, and you have been my impulse to give the best of me in this last period.

To my family who, despite the distance, were always with me encouraging me on this long walk (and asking every day: "Are you going to finish your thesis soon?"). To my mom Angela Cabezudo, my sister Mayra Sobrevilla, my "mamita" Victoria Carrizales, my uncle Cesar Cabezudo, my dad Mauro Sobrevilla and Iván Ferrer. You are the best family I could have had.

To my advisor, prof. Dr. Thiago Alexandre Salgueiro Pardo, for the guidance, the learning acquired, his good cheer and patience. For cheering me up when I felt lost and accepting me again as his student after the master :)

To my friends from the Methodist Church in São Carlos because they have been my family during these years. I want to thank the pastors, Levi Pereira, Gilson Sales and Nuria Lisboa and the members of the church :). In particular, I would also like to thank José Mendoza, Rosalía Taboada, Evandro Lima, Valdemar Abrão, Davi Canton, Israel Cassiano de Oliveira, Claudinei Brito Junior, Marcela Vasco, Anderson Bonilla, Daiany Silva Souza, Verônica Lima, Silveliana Silva, Alex Josué Flórez, Christian Barrantes and Fihama Santos. Your jokes, fun times, soccer days, study and prayer made the doctorate bearable.

To my friends at NILC, for the moments of learning, discussion, jokes and camaraderie. Especially to Marcio Lima Inácio, Renata Ramisch and Rafael Anchiêta (thank you, AMR team. AMRthons were very funny!), Edresson Casanova, Rogério Sousa, Roney Santos (what days in Lisbon!), Laura Quispe, Ana Caroline Brito, Ana Carolina Rodrigues, Henrico Brum, Raphael Rocha, Rafael Martins, Marcos Treviso, Sidney Leal, Nathan Hartmann, Fernando Nóbrega, Leandro Borges, Erick Fonseca, and João Paulo. It has been fun to share (with some of you, again) this time together.

To my friends at USP in São Carlos, because the conversations during lunch and football time were excellent and we had funny moments. In special, to Germain Zanabria, Wilbur Chiuyari, Juan Pablo Mamani, Paul Bustíos, Luis Rosero and Juan Luis Fuentes.

To the professors, Ariani Di Felippo, Helena Caseli, Claudia Barros and Eloize Seno, and to Marcio Lima Inácio, Renata Rasmisch, Rafael Anchiêta, Raphael Rocha, Roney Santos (all of you again :)), João Barbirato, Antônio Neto, and Verônica Lima who they were part of the team that helped me to refine and annotate part of the AMR corpus. Your feedback, suggestions and annotations made this work possible and more fun.

To my peruvian friends around the world, José Cárdenas, Henry Ramos, Roque Lopez, Alessandro Bokan, and Arturo Oncevay, for hearing me during my pain (hehehe), giving suggestions and feedback, and encouraging me to always do better.

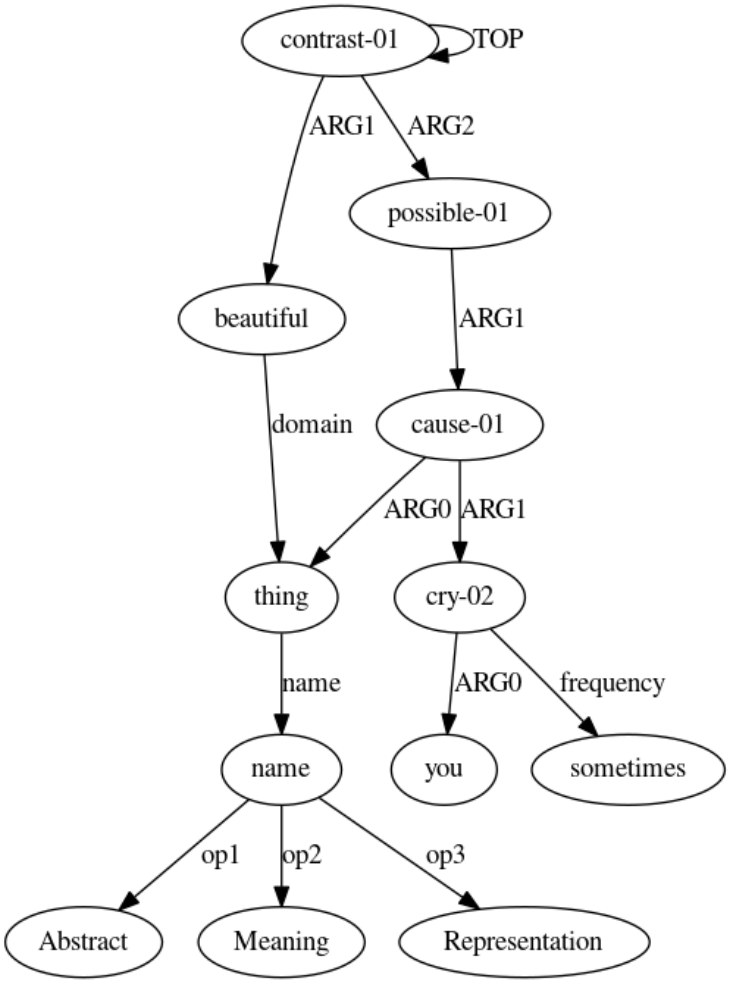
To the team I met at Google Summer of Code, Thiago Ferreira, Diego Moussallem, and Stuart Chen. You were a key piece at a time when I was quite lost and I thank you for everything you taught me.

To my friends at SiDi, for giving me the opportunity to learn a lot about the intersection between academia and industry and for giving me the time to advance with my doctoral work. In particular, I want to thank Fernando Nóbrega, Henrico Brum (again hehehe), Henrique Voni, Iury Americo Melo, Marcelo dos Anjos, and Rodrigo Rodarte.

To my friends at Alana AI, for motivating me and challenging me to always go one step further in what I do.

To the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), for the support. Furthermore, I would like to thank FAPESP (grant 2019/07665-4), IBM Corporation and the Center for Artificial Intelligence (C4AI), which, together with CAPES, gave all the structural and financial support so that this research could be carried out. Besides, I would like to thank to the Center for Mathematical Sciences Applied to Industry (CeMEAI) (FAPESP grant 2013/07375-0) for the computational resources.

Finally, I would like to thank the University of São Paulo and the Institute of Mathematics and Computer Sciences for giving me the opportunity to study as a PhD student and providing all the excellent professors and the amazing environment for carrying out this research.



"the Abstract Meaning Representation is beautiful to be a beautiful thing as you can cause."
(Unknown AMR-to-Text generator)

"So do not fear, for I am with you; do not be dismayed, for I am your God. I will strengthen you and help you; I will uphold you with my righteous right hand."
(Isaiah 41:10)

"Progress is made by trial and failure; the failures are generally a hundred times more numerous than the successes, yet they are usually left unchronicled."
(William Ramsay)

"Just because someone stumbles and loses their way doesn't mean they are lost forever."
(Charles Xavier, "Doctor Strange in the Multiverse of Madness")

"The hardest choices require the strongest wills."
(Thanos, "Avengers: Infinity War")

RESUMO

SOBREVILLA CABEZUDO, M. A. **Geração de linguagem natural por meio de representações semânticas abstratas para o português do brasil**. 2023. 199 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2023.

Abstract Meaning Representation é um formalismo semântico que codifica o significado de uma sentença como um grafo. Essa representação inclui várias informações semânticas, tais como os papéis semânticos, correferência, entidades nomeadas, entre outras. AMR tornou-se um tópico de pesquisa relevante nas áreas de representação semântica, análise semântica e geração de linguagem natural. Seu sucesso se baseia em sua tentativa de abstrair as idiossincrasias sintáticas e seu amplo uso de recursos linguísticos maduros, como o PropBank. A tarefa de geração de texto a partir de AMR (AMR-para-Texto) visa produzir um texto que transmita o significado codificado por um grafo AMR. Para o inglês, isso tem sido amplamente estudado, e várias abordagens como a tradução automática estatística, transdutores grafo/árvore a texto e, recentemente, modelos neurais têm sido explorados. Além disso, o corpus usado contém milhares de instâncias, possibilitando explorar diversos métodos e atingir altos desempenhos. Por outro lado, obter corpora de alta qualidade limita a pesquisa em outras línguas (pois geralmente envolve uma tarefa de anotação difícil e cara), resultando em corpora menores e na incapacidade de replicação de métodos e/ou obtenção de resultados semelhantes aos obtidos no Inglês. Para o Português Brasileiro, existe um corpus AMR contendo frases anotadas do livro "O Pequeno Príncipe" e vários analisadores AMR desenvolvidos. Nesse contexto, esta tese teve como objetivo investigar métodos de geração AMR-para-Texto para o Português Brasileiro, contribuindo para o desenvolvimento dessa linha de pesquisa. Dessa forma, primeiro adaptamos as diretrizes de AMR para o Português Brasileiro, construímos um novo corpus de AMR multigênero e fizemos uma análise de casos difíceis nos gênero de notícias jornalísticas e comentários opinativos. Além disso, adaptamos alguns métodos de geração AMR-para-Texto e os testamos em nosso corpus. Posteriormente, exploramos diversas estratégias para superar o tamanho limitado do corpus. Em particular, exploramos estratégias de língua cruzada usando o corpus AMR em Inglês e estratégias aprimoradas que visavam usar recursos (como modelos pré-treinados) e tarefas (como geração de paráfrases) para melhorar o desempenho dos mesmos. Entre os resultados, avaliamos as potencialidades e limitações de todas as estratégias, com especial enfoque para aquelas úteis para línguas com poucos recursos, sendo que as abordagens de língua cruzada produziram os melhores resultados. As contribuições desta tese também incluem os vários recursos AMR disponibilizados.

Palavras-chave: Geração de Linguagem Natural, Representação Semântica Abstrata, Entorno de Pocos Recursos, Português Brasileiro.

ABSTRACT

SOBREVILLA CABEZUDO, M. A. **Natural language generation from abstract meaning representation for brazilian portuguese**. 2023. 199 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2023.

Abstract Meaning Representation (AMR) is a semantic formalism that encodes the meaning of a sentence as a graph. This representation includes several semantic information, such as semantic roles, coreference and named entities, among others. AMR has become a relevant research topic in meaning representation, semantic parsing, and natural language generation (NLG). Its success is grounded in its attempt to abstract away from syntactic idiosyncrasies and its wide use of mature linguistic resources such as PropBank. The AMR-to-Text generation task aims to produce a text that conveys the meaning encoded by an input AMR graph. For English, this has been widely studied, and several approaches like Statistical Machine Translation, tree and graph to string transducers, and, recently, neural models have been explored. Besides, the corpus used contains thousands of instances, enabling to explore diverse methods and achieve high performance. Conversely, getting high-quality corpora limits the research in other languages (as it usually comprises a difficult and expensive annotation task), resulting in smaller corpora and the inability for state-of-the-art methods to be replicated and/or achieve similar performance to the English ones. For Brazilian Portuguese, there is an AMR corpus containing annotated sentences of the "The Little Prince" book and various AMR parsers developed. In this context, this thesis aimed to investigate diverse AMR-to-Text generation methods, contributing to the development of this research area. In this way, we first adapted the AMR guidelines to Brazilian Portuguese, built a new multi-genre AMR corpus, and made an analysis of hard cases in the news and opinative genres. Moreover, we adapted some AMR-to-Text generation methods and tested them on our corpus. Subsequently, we explored diverse strategies to overcome the limited corpus size. In particular, we explored cross-lingual strategies using the English AMR corpus and advanced strategies that aimed to use resources (such as pre-trained models) and tasks (such as paraphrase generation) to improve the performance. Among the results, we evaluated the strengths and limitations of all strategies, with a special focus on those useful for languages with few resources, being the cross-lingual approaches the ones that produced the best results. The contributions of this thesis also include the various AMR resources made available.

Keywords: Natural Language Generation, Abstract Meaning Representation, Low-Resource Setting, Brazilian Portuguese.

LIST OF FIGURES

LIST OF TABLES

CONTENTS

CHAPTER

1

INTRODUCTION

1.1 Context and Motivation

Since many years ago, researchers have tried to develop technologies for processing human language in order to improve the communication between computers and humans. Figure 1 shows an example of the work of Weizenbaum (1966), in which a chatbot, called ELIZA, tries to keep a dialogue with a user. This has led researchers to explore in depth about how to completely understand texts generated by humans and explore how to interact with them in a natural way.

Figure 1 – Example of the dialogue produced by ELIZA (WEIZENBAUM, 1966)

```
User:   I am unhappy.
ELIZA:  DO YOU THINK COMING HERE WILL HELP YOU NOT TO BE UNHAPPY
User:   I need some help, that much seems certain.
ELIZA:  WHAT WOULD IT MEAN TO YOU IF YOU GOT SOME HELP
User:   Perhaps I could learn to get along with my mother.
ELIZA:  TELL ME MORE ABOUT YOUR FAMILY
User:   My mother takes care of me.
ELIZA:  WHO ELSE IN YOU FAMILY TAKES CARE OF YOU
User:   My father.
ELIZA:  YOUR FATHER
User:   You are like my father in some ways.
```

Source: Jurafsky and Martin (2020).

According to Dale, Eugenio and Scott (1998), Natural Language Processing can be divided in two areas, Natural Language Understanding (NLU), which is in charge of mapping a surface representation (a text or speech) to an underlying representation of the meaning, and Natural Language Generation (NLG) that tries to answer the question of how one maps from some underlying representation of meaning into text or speech. In a more general sense, Natural Language Generation is defined as a subarea of Natural Language Processing and Artificial Intelligence that aims to provide computer systems with the ability to produce understandable texts in natural language from a non-linguistic representation of information (REITER; DALE, 2000).

This subarea has gained relevance in recent years and different applications can be seen

in tasks such as opinion summarization (CONDORI; PARDO, 2017), electronic health record generation (LEE, 2018), dialogue systems (NOVIKOVA; DUŠEK; RIESER, 2017), among others. Furthermore, some companies have made efforts to building natural language generation platforms, for example, ArriaNLG[1], IBM with the Watson project[2] and Alana AI[3].

Concerning the definition provided by Reiter and Dale (2000), it is worth noting that despite the use of non-linguistic representations as inputs (like images or graphs), some authors (VICENTE *et al.*, 2015; GATT; KRAHMER, 2018) consider texts as a possible input and establish a two classes of NLG applications. The first class is called *Data-to-Text* (or also called *Concept-to-Text*) and aims to produce texts from images, numerical data, database tables, semantic representations, among others. Some examples of this class are the generation of descriptions based on data from satellites or sensors and football reports. The last one is called *Text-to-Text* and aims to produce texts by using text written in natural language as input. Some examples of this type are textual simplification, generation of paraphrases and automatic summarization.

Despite this classification, Gatt and Krahmer (2018) note that the the boundaries between the two classes mentioned above are not so clear. For example, automatic summarization is clearly identified as a *Text-to-Text* application if we consider it from the point of view of an extractive approach (since it aims to produce summaries using sentences from the source documents). However, abstractive summarization (which generates sentences not present in any of the source documents) depends more on *Data-to-Text* applications, as they aim to generate new sentences from intermediate representations of the original texts.

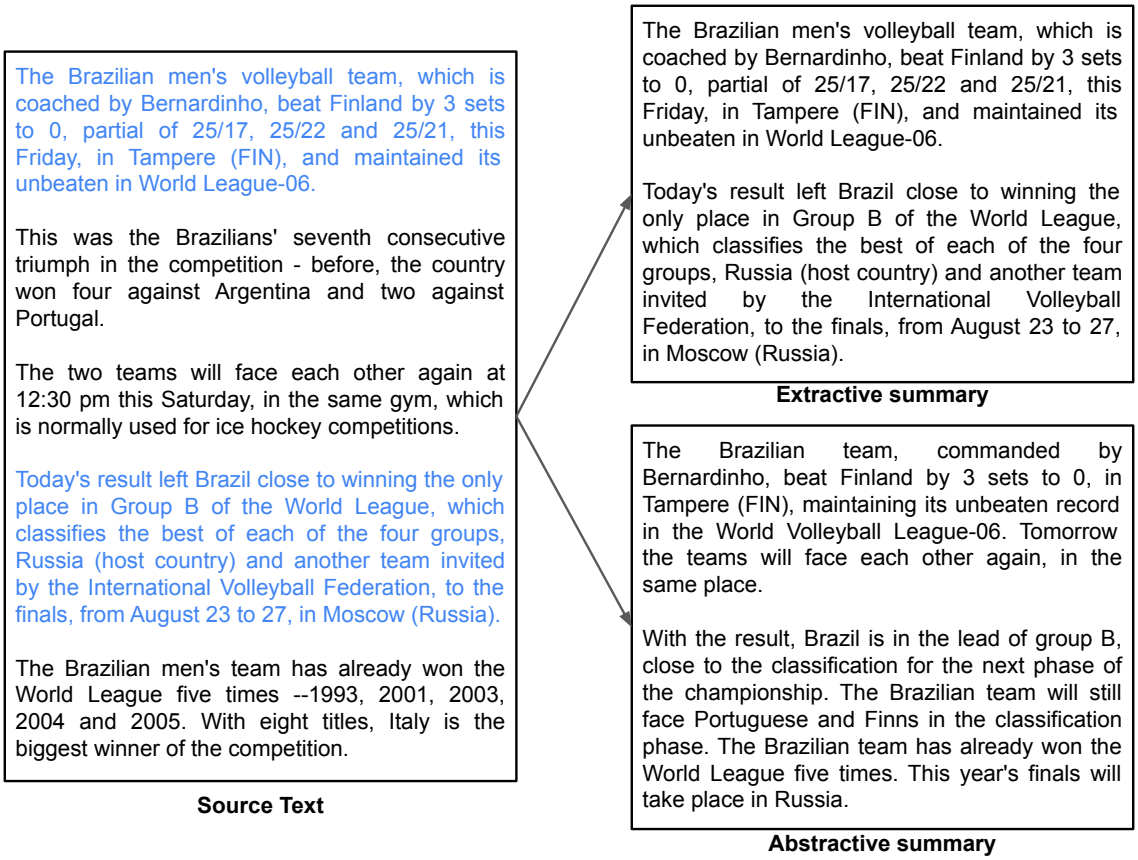
Figure 2 shows a text belonging to the sports domain, and its corresponding extractive summary (located in the upper right of the figure), an abstractive summary (located in the lower right of the figure). In it, one can observe that the extractive summary comprises a number of the most important sentences in the source text (highlighted in blue). This kind of summaries usually presents problems of coherence and cohesion as the selected sentences can express non-related topics, losing the fluency. Additionally, an extractive summary may not reflect some useful information for the user as the summary size limits may be reached depending on the selected sentences.

In the case of the abstractive summary, it can be seen that it is a synthesis of the source text that overcomes the difficulties of coherence and cohesion presented in the extractive summary. However, its generation needs a deeper automatic analysis of the source text to determine what information is more important, how it should be organized, which sentences should be generated as a union (or separation) from others, and how they should be described since these summaries must accomplish linguistic requirements such as grammaticality and meaning similarity. Thus, the generation of natural language becomes extremely important to deal with this kind of

[1] Available at <https://www.arria.com/>. Accessed on January 30, 2021.
[2] Available at <https://www.ibm.com/watson/>. Accessed on January 30, 2021.
[3] Available at <https://alanaai.com/>. Accessed on January 30, 2021.

summarization.

Figure 2 – Example of extractive and abstractive summaries.



Source: Elaborated by the author.

As mentioned in the previous example, *Data-to-Text* approach tends to be important in applications such as abstractive summarization, in which intermediate semantic representations are used to represent the source text(s) and the abstract to be generated and, then, a text is generated from it (MIRANDA-JIMÉNEZ; GELBUKH; SIDOROV, 2014; LIU *et al.*, 2015)[4].

A semantic representation can be defined as a representation that reflects the meaning of the text as it is understood by a language speaker[5]. In general, semantic representations can be classified into shallow and deep representations (ABEND; RAPPOPORT, 2017). Shallow semantic representations covers some aspects of semantics, such as predicates and semantic roles[6] and can be seen in the PropBank project (PALMER; GILDEA; KINGSBURY, 2005) and in the FrameNet project (BAKER; FILLMORE; LOWE, 1998). Deep semantic representations, such as *Universal Networking Language* (UNL) (UCHIDA; ZHU; SENTA, 1996), the one used

[4] It should be noted that the semantic representations addressed must be computationally treatable.

[5] The meaning of a sentence involves events, arguments, adjuncts, predicates, semantic roles, correspondence, temporal and spatial relations, and other information at the semantic or semantic-discursive level.

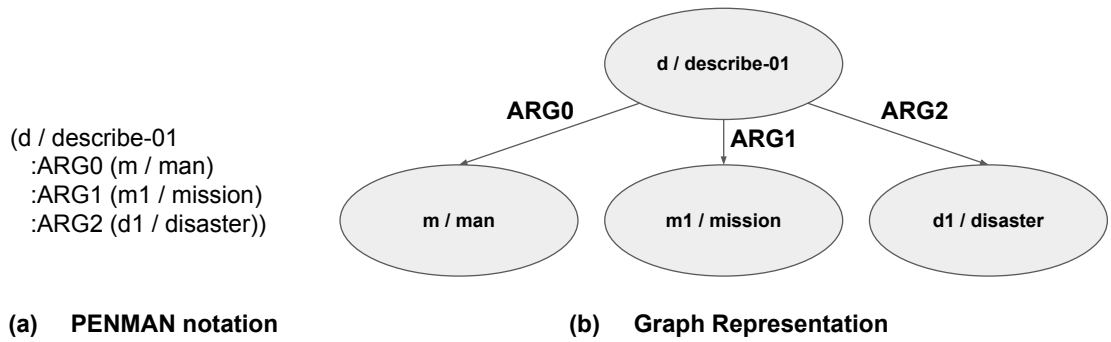
[6] The semantic roles describe the semantic relationships between a predicator and its arguments in a sentence.

in the *Groningen Meaning Bank* (GMB) (BASILE *et al.*, 2012) project and the representation used for *Universal Conceptual Cognitive Annotation* (UCCA) (ABEND; RAPPOPORT, 2013), cover other types of information such as coreference, spatio-temporal relations, word meaning, discourse-level information, among others. Other widely used representations, however, focused on specific tasks, are proposed by Gardent *et al.* (2017) (based on RDF triples) and Dušek, Novikova and Rieser (2018) (based on dialogue systems).

Among all the semantic representations, one of which has gained more relevance is Abstract Meaning Representation (AMR) (BANARESCU *et al.*, 2013). AMR is a semantic formalism that encodes the meaning of a sentence as a directed graph in which concepts are represented by nodes and relations are represented by edges. This representation includes information about semantic roles, named entities, Wikipedia entities, space-time information, and coreference, etc.

AMR has been successful in the research community due to its simpler structure compared to other representations and its wide use of other comprehensive linguistic resources like PropBank (BOS, 2016). Figure 3 shows the AMR graph (sub-figure b) that represents the sentence *The man described the mission as a disaster*[7] and its respective PENMAN notation (MATTHIESSEN; BATEMAN, 1991) (sub-figure a). Also in Figure 3, it is possible to see the concepts associated to the tokens "*described*", "*man*", "*mission*", and "*disaster*" (PropBank frameset *describe-01*, *man*, *mission*, and *disaster*, respectively) and the semantic roles represented by *:ARG0* (Agent), *:ARG1* (Theme) and *:ARG2* (Theme Description).

Figure 3 – AMR example



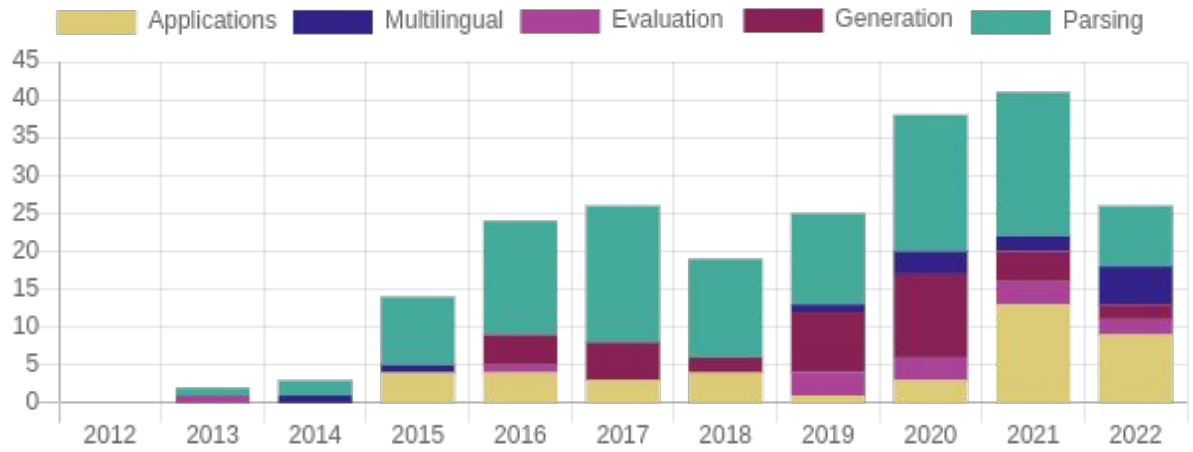
Source: Elaborated by the author.

Some examples of applications in which AMR has been applied are Automatic Text Summarization (HARDY; VLACHOS, 2018; LIAO; LEBANOFF; LIU, 2018; INÁCIO; PARDO, 2021), Dialogue Systems (BONIAL *et al.*, 2020), Event Extraction (RAO *et al.*, 2017), Paraphrase Detection (ISSA *et al.*, 2018; ANCHIÊTA; PARDO, 2020), and Machine Translation (SONG *et al.*, 2019).

[7] Other possible sentences generated by the graph could be *The man's description of the mission: disaster* and *As the man described it, the mission was a disaster* as these are semantically equivalent but syntactically different.

Natural Language Generation is not exempt from this trend. AMR has become one of the most used representations in data-to-text generation work. This can be seen in the proposed shared task at SemEval-2017 (MAY; PRIYADARSHI, 2017). Also, several works have been proposed in the last years. Figure 4 shows the number of works per year according to different AMR research topics and it is possible to see that NLG (wine color sub bar) always presents works, being 2020 the most popular.

Figure 4 – Number of work per year according to different AMR research topics.



Source: Adapted from AMR Biliography (2022).

It is also worth noting that multilingual works have increased in the last years (see blue sub bar in Figure 4). Even though AMR is strongly biased towards English (BANARESCU *et al.*, 2013), several works have tried to build AMR corpora for other languages by adapting the annotation guidelines (MIGUELES-ABRAIRA; AGERRI; ILARRAZA, 2018) or by using AMR as an interlingua and building corpora (semi) automatically (DAMONTE; COHEN, 2018; ANCHIÊTA; PARDO, 2018). Besides, the latter approach has mainly extended the research in cross and multilingual semantic parsing (BLLOSHMI; TRIPODI; NAVIGLI, 2020).

One problem that limits the research in other languages (mainly in the deep learning area) is the difficulty to get high-quality large corpora. On the one hand, adapting guidelines and annotating corpora from scratch allows researchers to tackle specific linguistic phenomena deeply. However, manual annotation is time-consuming and demands a team of experts to perform reliable annotation. There are just two large corpora available, the English corpus (59,255 annotated sentences)[8] and the Chinese corpus (10,325 annotated sentences)[9] and other work reports smaller corpora or the annotation of the book "The Little Prince" (MIGUELES-ABRAIRA; AGERRI; ILARRAZA, 2018; ANCHIÊTA; PARDO, 2018). These corpora result in the inability for state-of-the-art methods to be replicated and/or to achieve similar performance to the larger ones. On the other hand, assuming that AMR is an interlingua can help to accelerate

[8] Available at <https://catalog.ldc.upenn.edu/LDC2020T02>. Accessed 01/03/2021.
[9] Available at <https://catalog.ldc.upenn.edu/LDC2019T07>. Accessed 01/03/2021.

the annotation process as it is only necessary to translate the sentences to a target language and update the alignments to the target language (DAMONTE; COHEN, 2018) (trusting a machine translation system). Nonetheless, this approach can constrain linguistic research in a specific language, preventing the building of corpora for other languages. Besides, some studies have shown that AMR is almost an interlingua (XUE *et al.*, 2014), therefore, there are linguistic phenomena that need adaptations to a target language.

Another alternative to leverage the interlingual potential of AMR is to import the corresponding AMR annotation for each sentence from the source annotated corpus (usually English) and review the annotation to adapt it to the target language (ANCHIÊTA; PARDO, 2018). This seems to be a better alternative than the previous one mentioned. However, it depends on the magnitude of the divergences between languages to make the importation helpful. Some studies have presented an analysis of the differences between AMR representations in different languages, but, as far as we know, there is no study about how these differences can affect the performance in tasks such as semantic parsing or natural language generation.

Concerning the AMR-to-text generation task in different languages, it is worth noting that most work has focused on English. Earlier methods focused on statistical machine translation (POURDAMGHANI; KNIGHT; HERMJAKOB, 2016), tree and graph to string transducers (FLANIGAN *et al.*, 2016; SONG *et al.*, 2017), transition-based methods (LAMPOURAS; VLACHOS, 2017), neural models such as sequence-to-sequence (FERREIRA *et al.*, 2017; KONSTAS *et al.*, 2017; ZHU *et al.*, 2019a) and graph-to-sequence (BECK; HAFFARI; COHN, 2018; SONG *et al.*, 2018; RIBEIRO; GARDENT; GUREVYCH, 2019; ZHANG *et al.*, 2020), and, recently, pre-trained neural models (MAGER *et al.*, 2020; RIBEIRO *et al.*, 2021b) have become a trend in this field.

In particular, the methods previously mentioned have to deal with some challenges. Methods based on Machine Translation (both statistical and neural) usually rely on alignments between the concepts/relations of the AMR graph and the tokens in the sentence for determining the ordering and selecting the concept/relations that should be part of the flattened version. This way, the quality of the alignments (in the case of alignments generated automatically) and the performance of the methods that use these alignments for generating a flattened version of AMR graphs become crucial on the overall performance of the AMR-to-text generation task[10] .

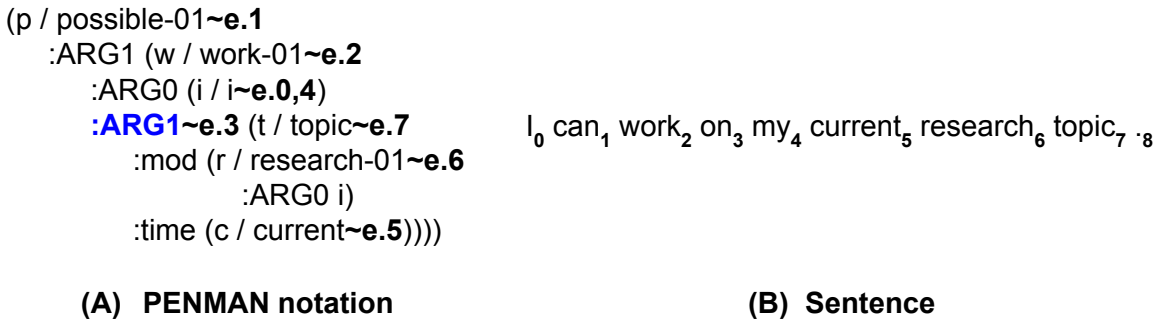
Works like the proposed ones by Pourdamghani, Knight and Hermjakob (2016) and Ferreira *et al.* (2017) use alignments based on tokens for training models that be able to generate an English-like flattened AMR version as, according to the authors, this makes the generation of text easier. On the other hand, Konstas *et al.* (2017) show that the ordering of AMR tokens is not necessary to achieve good performance. Concerning this, Ferreira *et al.* (2017) suggest that data augmentation strategies can reduce the relevance of the ordering. Finally, Mager *et al.* (2020)

[10] A flattened version of an AMR graph is a linearized representation of the AMR graph. An example of this can be seen on Figure 5.

note that PENMAN notation is useful in text generation, highlighting the use of parentheses for introducing structural information.

Figure 5 shows the PENMAN notation of an AMR graph (A) and the corresponding sentence (B) with its respective token-based alignments. Concepts and relations with numbers in bold represent the aligned tokens, and the relation in blue is the only one aligned. Moreover, it shows a flattened version similar to the expected by Ferreira *et al.* (2017) (C) and Mager *et al.* (2020) (D). As it may be seen, the flattened version described in (C) is an English-like version and, also, includes one relation that matches the preposition "*on*", and the flattened version in (D) is the PENMAN notation disregarding the name of the variables in the AMR graph.

Figure 5 – Example of alignments between concepts/relations of AMR graph and words in its corresponding sentence and different versions of a flattened AMR graph.



Source: Elaborated by the author.

In order to overcome the problems that machine translation-based methods face, the Graph-to-text approach emerges as a useful alternative. The Graph-to-text approach models structural information more naturally (without losing information) and produces better results than the previous ones. Furthermore, the performance of methods based on this approach increases more when data augmentation strategies are applied, largely overcoming the performance of the machine translation approach. Conversely, methods based on this approach usually are data-hungry (because of their complexity) and, therefore, their performances can be lower in low-resource settings.

Recently, transfer learning has become widely explored in NLP, and pretrained Transformer-based architectures have outperformed prior State-of-the-Art (SotA) (DEVLIN *et al.*, 2019; RADFORD *et al.*, 2019; LEWIS *et al.*, 2020; RAFFEL *et al.*, 2020). These models are pretrained on large corpora of available unannotated text. Then, they are fine-tuned for specific tasks on smaller amounts of supervised data, relying on the induced language model structure to facilitate generalization. Concerning the AMR-to-text generation task, Mager *et al.* (2020) propose the

use of GPT-2 (RADFORD *et al.*, 2019) to learn the joint distribution of AMR and the text and Ribeiro *et al.* (2021b) study how BART (LEWIS *et al.*, 2020) and T5 (RAFFEL *et al.*, 2020) performs on this task, obtaining improvements in the performance. Additionally, the latter work explores the use of task-adaptative pretraining (TAP), obtaining improvements too. TAP consists of adding more silver data[11] -obtained by a SotA AMR parser-, pretraining the models on silver data, and then continuing the training on the actual data.

One of the biggest problems that most previously mentioned methods have to deal with is data sparsity, which is caused by the ratio of broad vocabulary and a relatively small amount of data (FERREIRA *et al.*, 2017). On the one hand, as defined by Banarescu *et al.* (2013), AMR graphs can be associated with several realisations, i.e., the ratio of tokens/types is low, and this can be problematic for dealing with out-of-vocabulary (oov) or rare words. Besides, this problem gets worse in morphologically rich languages as the diversity of the vocabulary is higher.[12] On the other hand, small corpora harm the performance in some approaches (mainly neural approaches) as they usually need large corpora to return adequate results. For example, English work is evaluated on corpus that comprises approximately 36,000 instances[13], thus, this allows to get acceptable performance, however, this result can dramatically drop in low-resource settings (RIBEIRO *et al.*, 2021b; RIBEIRO; ZHANG; GUREVYCH, 2021).

Concerning the ratio of broad vocabulary, some initial strategies are to use of delexicalisation, which consists of replacing some sparse tokens (usually named entities, dates, or numbers) with dummy tokens (KONSTAS *et al.*, 2017; FERREIRA *et al.*, 2017; BECK; HAFFARI; COHN, 2018) and to apply a copy mechanism (GU *et al.*, 2016; GULCEHRE *et al.*, 2016), which consists of copying some tokens belonging to the input to the output. Both strategies have been applied by Song *et al.* (2018), Ribeiro, Gardent and Gurevych (2019) with success.

Another strategy is to use models based on smaller units (different from words). The main goal of this strategy is to segment words in smaller units for reducing the number of oov and rare words without overly increasing the vocabulary size, since this can produce an explosion in the number of parameters of the model without having enough training examples for proper estimation. In this way, early work applied character-based models in AMR-to-text generation, obtaining good results (KONSTAS *et al.*, 2017), and recently, some work have applied subword-based models (ZHU *et al.*, 2019a; MAGER *et al.*, 2020), which usually use the byte-pair encoding algorithm proposed by Sennrich, Haddow and Birch (2016). This latter demonstrated to be beneficial, keeping an adequate vocabulary size and generating a proper sequence of tokens (in terms of length), different from characters that generate longer ones, harming the training.

Concerning the small corpora, Data Augmentation helps to overcome the data sparsity

[11] Silver data can be defined as data generated by a model automatically.
[12] Many different surface forms can be generated by the same stem/lemma, augmenting the vocabulary.
[13] Current version of the corpus contains 59,255 instances. Available at <https://catalog.ldc.upenn.edu/LDC2020T02>.

as increasing the corpora size reduces the occurrence of out-of-vocabulary and rare words and allows to incorporate more examples related to in-domain instances, which is beneficial for training. This strategy has been usually applied by previous work, getting wide improvements. However, there is a problem in low-resource settings as the quality of the augmented data can be low and introduce noise in the training data, thus, it could be preferred to select higher-quality instances in order to not harm the performance (SOTO *et al.*, 2020). Additionally, some work suggest that the performance can also be improved with a small number of instances if they are in the same domain as the test data, making the data selection an important step in this context (PONCELAS; WAY, 2019).

A different way of augmenting data is to leverage the knowledge from other languages or tasks in high resource settings. As previously mentioned, there is a large AMR corpus for English. An initial approach adopted by Fan and Gardent (2020) was to translate the sentences included in the AMR corpus to several languages (including Portuguese), aiming to explore multilingual AMR-to-text generation. In the case of AMR parsing, several cross-linguistic studies have been performed. Damonte and Cohen (2018) try to project AMR graphs from English sentences to target language sentences through a parallel corpus (not the AMR corpus). On the other hand, Blloshmi, Tripodi and Navigli (2020) try to translate the sentences of the AMR corpus to the target language arguing that the previous approach can generate low-quality AMR graphs.

All the mentioned studies have in common that start (in the case of generation) or end (in the case of parsing) in the English AMR graph, considering it an interlingua. However, AMR-to-text generation for specific languages needs to start from a language-specific graph that handles its linguistic phenomena, thus, an alternative could be also to translate the graph and deal with the differences between languages. This could be even better as this data could serve for generating possibly more information higher-quality silver data. On the other hand, problems in translations and alignments could harm the performance, hence, it would be interesting to evaluate if it is better to preserve the English version of the graphs or not.

1.2 Gaps

As it was shown in Section 1.1, AMR and its applications have been widely explored in English; however, this does not usually happen in other languages. For Portuguese, in particular, there is only a small corpus focused on tales ("The little Prince") (ANCHIÊTA; PARDO, 2018), an aligner (ANCHIÊTA; PARDO, 2020), and some parsers (ANCHIÊTA; PARDO, 2018). Although these are valuable resources and tools, the focus (tales) can constrain research and its application, therefore, it is necessary to extend the corpora to other genres such as journalistic and opinative. On the other hand, several studies have shown the usefulness of AMR in text generation tasks such as Machine Translation (SONG *et al.*, 2019), and Automatic Summarization (HARDY; VLACHOS, 2018); therefore, improvements in NLG from AMR could benefit the

related applications.

Another problem emerges concerning the construction of an AMR corpus for other genre. The current AMR corpus was semi-automatically annotated, leveraging the parallel corpus and importing the annotation from English to Portuguese. Conversely, annotating a corpus in another genre(s) could be expensive and could limit the corpus size, harming the learning and the performance of AMR-to-text generation methods.

1.3 Goals, Hypotheses and Research Questions

The main goal of this work was to explore, develop, adapt and evaluate NLG methods for Brazilian Portuguese from AMR. The hypothesis that guides this goal is that it is possible to develop natural language generation methods for Portuguese from AMR with similar accuracy to the English AMR-to-Text generation task, even in a low-resource setting.

The following specific goals arise from the main goal:

- Creating an AMR corpus for the development and evaluation of NLG methods in Brazilian Portuguese.
- Evaluating the helpfulness of using English-focused AMR corpus for improving the AMR-to-Brazilian Portuguese generation task.
- Comparing pipeline-based methods with end-to-end neural methods for AMR-to-Brazilian Portuguese generation task.
- Evaluating the performance of data augmentation methods in AMR-to-text generation and applying strategies for better selecting the augmented data.

The hypotheses related to the goals proposed in this work are described as follows:

- English AMR corpus, despite the linguistic phenomena differences, improves the AMR-to-text generation task for Brazilian Portuguese.
- Data augmentation improves the performance of Low-resource AMR-to-text generation.
- Pipeline approaches lead to improvements in low-resource AMR-to-text generation.

To achieve our main goal and confirm the hypotheses, some research questions had to be answered:

- How different is English AMR corpus from Portuguese AMR corpus in terms of linguistic phenomena?

- Is it possible to leverage the cross-linguistic potential of the English AMR corpus for increasing the size of the Portuguese AMR corpus and the performance of AMR-to-text generation?
- What is the best strategy for dealing with data sparsity in AMR-to-text generation?
- What is the best way to leverage the knowledge provided by the English AMR corpus?
- How does data augmentation methods behave on AMR-to-text generation in low-resource settings and what is the best way to augment data?

1.4 Thesis Organization

The present thesis is organized into seven chapters, and some of them include published papers and in the process of being published (chapters 3, 4, 5 and 6). This organization is described as follows:

- Chapter 2 presents some definitions of the topics involved in this thesis. In particular, this chapter describes concepts related to Natural Language Generation, Abstract Meaning Representation, Low-resource Natural Language Processing, and some tools and resources included in this work.
- Chapter 3 shows a paper that describes the related work in Natural Language Generation from Semantic representation and some directions and problems of its application for Brazilian Portuguese.
- Chapter 4 describes the manual annotation of a journalistic corpus for Brazilian Portuguese. Besides, a study of some linguistic phenomena during the AMR annotation of journalistic and opinative texts is presented.
- Chapter 5 shows how different cross-lingual strategies that uses English AMR corpus perform on low-resource AMR-to-text generation.
- Chapter 6 presents various strategies to use knowledge provided by resources and tasks for improving the AMR-to-text generation task.
- Finally, Chapter 7 presents the conclusions, contributions and limitations of this work as well as future research directions are addressed.

CHAPTER

2

BASIC CONCEPTS

This chapter presents some definitions about the topics described along this dissertation. In particular, we describe concepts related to Natural Language Generation, Abstract Meaning Representation (our focus), Low-Resource Natural Language Processing and some tools and resources used in this work.

2.1 Natural Language Generation (NLG)

Natural Language Generation (NLG) is a sub-field of Natural Language Processing and Artificial Intelligence that aims to provide computer systems with the ability to produce understandable texts in natural language from a non-linguistic representation of information (REITER; DALE, 2000).

Research and applications in NLG have increased in recent years. Some examples can be seen in the generation of weather forecast reports from graphical weather maps (WANNER *et al.*, 2015), in the generation of opinion summaries (CONDORI; PARDO, 2017) and in the automatic generation of journalistic news (VILCA; CABEZUDO, 2017). In addition to this, some companies have made efforts to build natural language generation platforms such as ArriaNLG [1] and IBM with the Watson project [2].

Although the definition of NLG is focused on producing texts from non-linguistic representations, authors have diversified the input of NLG systems, which can be images, databases, semantic representations, or even texts (GATT; KRAHMER, 2018). This way, in general, NLG systems can be classified into two types according to the input they receive:

- *Data-to-Text* or *Concept-to-Text* (VICENTE *et al.*, 2015), that produces text from data different from text such as numerical data, tables, or semantic representations, and;

[1] Available at <https://www.arria.com/>.
[2] Available at <https://www.ibm.com/watson/>

- *Text-to-Text*, that generates text from texts or sentences. Some examples are automatic abstractive summarization and text simplification.

2.1.1 Natural Language Generation Tasks

NLG systems can be cast as a group of 6 tasks that convey information in natural language (REITER; DALE, 2000; FERREIRA *et al.*, 2019), which are described below:

- Content Determination: this task is responsible for determining what part of the information should be shown to the user. The selection of information may depend on the users or on the communicative intention;
- Discourse Planning or Text Structuring: this task defines how the selected information should be organized, and it is related to the structuring of the discourse;
- Sentence Aggregation: this task is responsible for grouping the information in a sentence to make the text to be generated more readable and avoids redundancies;
- Lexicalization: this task is in charge of deciding which words or expressions in natural language should be used to convey the determined content. Besides, lexicalization is related to the variation of the vocabulary and the context to determine which words to choose;
- Referring Expression Generation (REG): this task decides which expressions should be used for referring entities in the generated text. This task is different from lexicalization as it is a discrimination task, where there is a need to transmit enough information to distinguish one entity from the others; and,
- Linguistic Realization: this task is responsible for generating the text in its final form. This task includes ordering constituents of a sentence, generating correct inflections, and inserting functional words and punctuation marks.

2.1.2 NLG Evaluation

NLG systems can be evaluated extrinsically and intrinsically. The extrinsic evaluation measures the effectiveness of achieving a goal. This effectiveness depends on the context and the purpose of a system—for example, purchase decision after reading machine-generated arguments for and against a product. In general, this evaluation can be carried out through surveys to ask users about the tool's utility. The intrinsic evaluation aims to measure the performance of an NLG system itself, i.e., to evaluate whether the texts generated by these systems are similar to those created by a human. This assessment can be performed using automatic metrics or using human assessment.

Different automatic metrics calculate how similar an NLG system's output is with one or more reference texts. Among them, we can note n-gram overlap metrics, string distance metrics, and, recently, semantic metrics. N-gram overlap metrics like BLEU (PAPINENI *et al.*, 2002), METEOR (LAVIE; AGARWAL, 2007), and chrF++ (POPOVIĆ, 2017), regards the word/character n-gram overlapping between the system's output and the actual output. On the other hand, string distance metrics such as TER (SNOVER *et al.*, 2006), measure the effort to convert the system's output into the actual output.

Unlike the earlier mentioned metrics that rely largely on surface-level matches, semantic metrics regard semantic similarity provided by word embeddings. This way, they can deal better with synonyms and paraphrases. Some of the well-known metrics are BERTScore (ZHANG *et al.*, 2020), which follows an unsupervised strategy to compute the similarity between a reference and an output, and BLEURT (SELLAM; DAS; PARIKH, 2020), which was trained for natural language evaluation purposes, rely on pre-trained language models and have shown improved correlations with human judgments at sentence-level.

Although it is easy to calculate automatic metrics, allowing a quick evaluation, the results they obtain do not reflect aspects that a human evaluates, such as readability and accuracy, among other characteristics. Therefore, another way to evaluate the outputs of natural language generators is through questions to humans about some specific aspects. The most frequently evaluated aspects are listed below (GATT; KRAHMER, 2018):

- Fluency or readability, that measures the linguistic quality of the text; and,
- Accuracy, adequacy, relevance or correctness relative to the input, that measures if the system's output reflects the meaning of the reference.

Human evaluation can be performed using an ordinal scale, where evaluators choose an option from a range of options (for example, using a Likert scale). A problem with this type of evaluation is that it makes it difficult to compare different systems. For example, if a judge chooses the lowest option for a system's output and then evaluates the output from other system that is worse than the previous one, he has no more suitable option to choose.

Another alternative is applying evaluations that use a continuous scale where judges score between 0 and 100 according to their criteria. One of the problems with this type of evaluation is the high variance of the results (GATT; KRAHMER, 2018) since several judges participate in this evaluation. To overcome these difficulties, studies can conduct training sessions where the evaluators can learn together and thus reduce the variance of the evaluations.

Finally, evaluations can be performed using rankings, comparing the outputs of the systems in parallel. This approach was used in the shared task proposed by May and Priyadarshi (2017). In this approach, the judge's task is to rank each of the outputs according to previously defined aspects.

2.2 Abstract Meaning Representation (AMR)

AMR is a semantic representation intended for large-scale annotation of a giant semantic bank (BANARESCU *et al.*, 2013). The authors' main goal is to create a large semantic bank of sentences to treat various semantic phenomena jointly and not in an isolated way, as has been done.

In general, AMR is built based on the following principles:

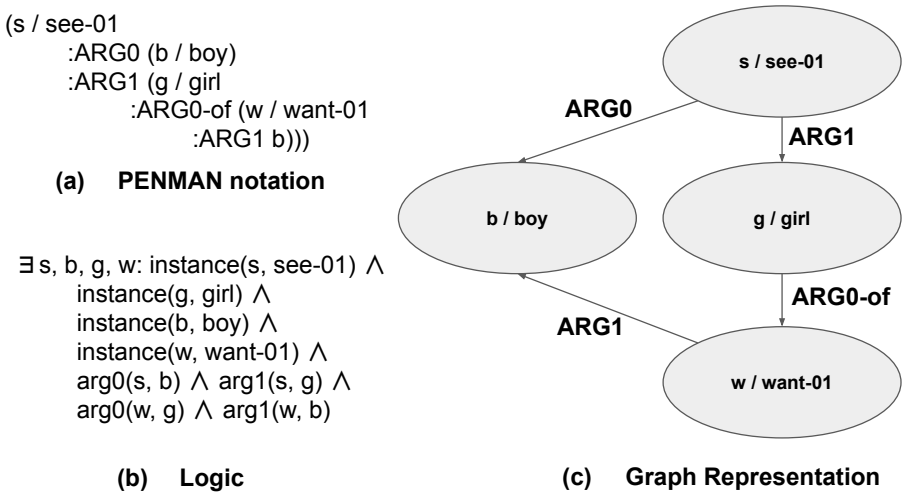
- AMR is a rooted labeled directed graph, easy to read by people and easy to navigate through programs;
- AMR aims to abstract away syntactic idiosyncrasies. This way, sentences "*The boy wants the girl to believe* " (shown in Figure 6) and "*The boy wants to be believed by the girl*" are presented by the same AMR graph although they are syntactically different;
- AMR makes extensive use of PropBank framesets (PALMER; GILDEA; KINGSBURY, 2005) (described in the Subsection 2.4.1)
- AMR is agnostic about how we might want to derive meanings from texts, or vice versa; and
- AMR is strongly biased towards English. It is not an interlingua.

AMR is represented by directed acyclic graphs, in which nodes represent concepts (words in lexicalized form) and edges represent relations between concepts. Besides, concepts are referenced by variables. A sentence can be formatted in AMR by using: (1) PENMAN notation (sub-figure "a" from Figure 6) (MATTHIESSEN; BATEMAN, 1991), (2) first-order logic (sub-figure "b" in Figure 6), and (3) graph (sub-figure "c" from Figure 6).

Figure 6 shows the AMR representation of the sentence *"The boy saw the girl who wanted him"* in PENMAN notation (a), first-order logic (b), and graph (c). The AMR graph contains some concepts such as "(b / boy)", which refers to an instance (called *b*) of the concept *boy* . Another concept is PropBank's frameset of the verb *see-01* (See frameset definition in Section 2.4.1). Relationships are represented by edges and can assume the values of arguments in PropBank's frameset. For example, the concept *boy* is an Arg0 (agent) of the verb *see-01*. Besides, each concept has a unique variable that allows to identify it in the graph. These variables also allow creating references easily. For example, in Figure 6, the concept want-01 also makes reference to the concept *boy* , which fulfills the function of theme (Arg1).

In addition to using the relations coming from PropBank (defined by the ARGs), AMR defines other types of relations and strategies to address some phenomena such as questions,

Figure 6 – AMR representations for sentence "*The boy saw the girl who wanted him*" (a) First-order logic notation (b) PENMAN notation (c) Graph notation. Example extracted from Banarescu *et al.* (2013).



Source: Elaborated by the author.

named entities, and copula verbs, among others. Besides, there are other general semantic relations/concepts, such as quantity, date, named-entity, modality, and negation[3].

The first English AMR corpus comprised the annotation of sentences from the "The Little Prince" tale (BANARESCU *et al.*, 2013). In general, this corpus consists of 1,562 annotated sentences distributed as follows: 145 for the development set, 1,274 for the training set, and 143 for the test set. The annotation of the Little Prince was motivated by the fact that this book is of public and free access, and other semantic projects started by annotating the same book.. This way, it allows different groups to make comparisons between representations of the same text. After the first version of the AMR corpus, other versions for English emerged that will be described below:

- LDC2014T12[4]: First version that contained sentences from several newspapers and discussion forums. This version comprises 13,051 sentences distributed as follows: 10,312 sentences for training, 1,368 for development and 1,371 for testing.
- LDC2015E86: version proposed in the first semantic parsing task (MAY, 2016). This version includes 19,572 sentences distributed as follows: 16,833 sentences for training, 1,368 for development and 1,371 for testing.
- LDC2016E25: version proposed in the shared task of semantic parsing and text generation (MAY; PRIYADARSHI, 2017). It comprises 39,260 sentences distributed as follows: 36,521 sentences for training, 1,368 for development and 1,371 for testing.

[3] Annotation guidelines with the linguistic phenomena and examples of these are available at <https://github.com/amrisi/amr-guidelines/blob/master/amr.md>.

[4] Available at https://catalog.ldc.upenn.edu/LDC2014T12. Accessed on October 1, 2022.

- LDC2017T10[5]: second version available and it is the same as LDC2016E25.
- LDC2020T02[6]: this version contains 59,255 sentences, divided into 55,635 sentences for training, 1,722 for development, and 1,898 for testing.

Concerning AMR corpora for non-English languages, Xue *et al.* (2014) and Urešová, Hajič and Bojar (2014) presented the annotation of 100 Czech and Chinese sentences and analyzed the divergences/similarities between the annotations. Later, Li *et al.* (2016) released the Chinese AMR corpus, which contains AMR-annotated sentences from the "The Little Prince". Then, Li *et al.* (2019) published the most extensive AMR corpus for Chinese, which contains 10,149 sentences belonging to the news texts domain. In addition to these works, efforts have been performed for other languages such as Vietnamese (LINH; NGUYEN, 2019), Turkish (AZIN; ERYIĞIT, 2019), and Spanish (MIGUELES-ABRAIRA; AGERRI; ILARRAZA, 2018), and all worked on the "The Little Prince" and adapted the annotation guidelines to their own languages.

For Brazilian Portuguese, Anchiêta and Pardo (2018) also created the AMR corpus from the Little Prince book for Portuguese. The strategy that the authors used consisted of (1) aligning the sentences between the English and Portuguese versions, (2) mapping the AMR representations of the English corpus to Portuguese, and (3) including the *framesets* of the predicates belonging to the Verbo-Brasil repository (DURAN; ALUÍSIO, 2015)[7], which is similar to the PropBank, and fixing some mistakes in the annotation.

Finally, other works have tried to explore the interlingual potential of AMR by assuming that Enligsh AMR graphs are general language-independent representations and only translating the sentence to other languages. This way, Damonte and Cohen (2018) and Blloshmi, Tripodi and Navigli (2020) explore the cross-lingual AMR parsing task and build datasets for Italian, Spanish, German, and Chinese. In relation to AMR-to-text generation, Fan and Gardent (2020) study multilingual settings, in which the authors aimed to generate text in diverse languages from English AMR graphs.

2.3 Low-Resource Natural Language Processing

Nowadays, there is considerable research in Natural Language Processing (NLP) focused on a few languages with high resources (being English the main one), leaving aside many others with millions of speakers (BENDER, 2019). With the emergence of deep learning, which requires large volumes of data, the performance of diverse NLP applications increased considerably. However, the scarcity of data in low-resource languages made NLP application development a challenging problem as the performance obtained are lower than the former ones.

[5] Available at https://catalog.ldc.upenn.edu/LDC2017T10. Accessed on October 1, 2022.
[6] Available at <https://catalog.ldc.upenn.edu/LDC2020T02>. Accessed on October 1, 2022
[7] More about this repository will be detailed in the Subsection 2.4.1.

Initially, the term "low-resource" was strongly associated with languages; however, this term has been expanded to cover diverse scenarios, such as widely used languages that are not often treated in NLP research, and popular languages in NLP in which there are only small training corpora available for some tasks and uncommon domains.

Concerning the previously mentioned, Hedderich *et al.* (2020) propose to categorize low-resource settings along the following three dimensions:

- the availability of task-specific labels in the target language/domain. Labels are defined/assigned in manual annotation and this is time-intensive and expensive task in several cases;
- the availability of unlabeled language/domain-specific text since current NLP approaches are built on representations trained on unlabeled texts; and,
- the availability of auxiliary data since transfer learning might leverage task-specific labels in a different language or domain.

According to the literature, we can identify three main strategies applied in low-resource settings, which are described as follows:

2.3.1 Additional Labeled Data Generation

In order to alleviate the lack of task-specific labels, diverse approaches have been proposed to augment labeled data via expert insights and automatic methods. Some of the most used strategies are data augmentation, distant supervision or weak supervision, cross-lingual annotation projection, and learning with noise labels.

Data augmentation has its origin in computer vision, where new images are built based on existing ones by applying some operations such as rotations, distortion, scale, and others, without losing their original label. For text, some strategies to generate new instances have been applied. For example, synonyms or related words that share some aspects can be used to replace tokens (FADAEE; BISAZZA; MONZ, 2017), parts of a syntax tree can be modified by applying some operations (DEHOUCK; GÓMEZ-RODRÍGUEZ, 2020), and back-translation (SENNRICH; HADDOW; BIRCH, 2016) or paraphrasing can be applied for generating new sentences.

Distant supervision or weak supervision increases the data size (unlabeled text) by (semi-) automatically assigning labels from an external source or using some heuristics. For example, Corrêa Jr *et al.* (2017) applied distant supervision for increasing data size in a twitter-based sentiment analysis corpus. The authors used emojis like ":)" and ":(" for annotating tweets as positive or negative, respectively. Similarly, Karamanolakis, Hsu and Gravano (2019) create

a simple bag-of-words classifier on a list of seed words and train a deep model for aspect classification.

Another way to augment data is by using cross-lingual projection. It consists of (1) training a task-specific model on a high-resource language, (2) selecting a parallel corpus[8] that includes the high- and the target low-resource language, (3) applying the model on the sentences in the high-resource side, and (4) projecting the annotations to the corresponding sentences in the low-resource side. This strategy have been applied on tasks such as semantic parsing (DAMONTE; COHEN, 2018; BLLOSHMI; TRIPODI; NAVIGLI, 2020), and POS-Tagging (PLANK; AGIĆ, 2018), among others. An alternative to not use parallel corpora is to translate high-resource labeled datasets via machine translation (MONSALVE *et al.*, 2019; FEI; ZHANG; JI, 2020).

All mentioned methods help to get labeled data cheap and rapidly; however, labels usually contain errors that can hurt the performance of the models depending on how noisy the labeled data is. To prevent noise to affect models negatively, authors usually apply noise filtering and noise modeling strategies. Noise filtering consists of removing instances from the training data that have a high probability of being incorrectly labeled, and a classifier is usually trained to make the filtering. On the other hand, in noise modeling, the classifier is not trained on noisy labeled data. However, a noise model is trained for changing from a noisy to a clean label distribution. This can be interpreted as the original classifier being trained on a "cleaned" version of the noisy labels.

2.3.2 Transfer Learning

Unlike the approaches mentioned in the previous sub-section, that aim to increase the task-specific training data, transfer learning decreases the need for labeled target data by transferring representations and models previously learned.

One of the findings that boosted Transfer learning in NLP was the pre-trained word representations. Works like the proposed ones by Mikolov *et al.* (2013) and Bojanowski *et al.* (2017) aimed to generate fixed representations from training on large unlabeled data and resulted in improvements for diverse tasks. In particular, subword-based embeddings (BOJANOWSKI *et al.*, 2017) and byte-pair encoding embeddings (HEINZERLING; STRUBE, 2018) have shown improvements in morphologically rich languages. For example, Zhu *et al.* (2019b) and Regatte, Gangula and Mamidi (2020) showed that these embeddings are beneficial for low-resource sequence labeling and sentiment analysis tasks, respectively, outperforming word-level embeddings.

Recently, models based on the transformer architecture (VASWANI *et al.*, 2017), pre-

[8] Parallel corpora belonging to the OPUS project (TIEDEMANN, 2012) have usually been used in this task. The corpora included in this project are available at <https://opus.nlpl.eu/>.

trained on large data, have emerged, producing improvements in diverse tasks. Some pre-trained models like BERT (DEVLIN *et al.*, 2019) have shown improvements in low-resource languages for which large amounts of unlabeled data are available and task-specific labeled data is limited, such as the named-entity recognition task for Persian (TAHER; HOSEINI; SHAMSFARD, 2019).

It is worth noting that these models have been usually trained on large general data such as news or web-domain texts and it can lead to problems when applied to different and specific domain. To solve this problem, literature suggests to adapt the model to the target domain by fine-tuning (or continuing the pre-training) the model since unlabeled domain-specific data is easier to collect, and then, fine-tune on the task-specific data. Some examples include BioBERT (LEE *et al.*, 2019) that was pre-trained on biomedical PubMED articles, and SciBERT (BELTAGY; LO; COHAN, 2019) for scientific texts. Particularly, Friedrich *et al.* (2020) showed that SciBERT outperforms the original BERT in tasks related to the science domain.

2.3.3 Low-Resource Machine Learning

All the mentioned strategies arise from other areas such as computer vision and machine learning, and in general, NLP strategies have been inspired by these areas in recent decades. An emergent approach is Meta-Learning (FINN; ABBEEL; LEVINE, 2017). It is also known as "learning to learn", and tries to design models that can learn new abilities or adapt to new environments quickly with few training examples. In practice, given a set of auxiliary high-resource tasks and a low-resource target task, meta-learning trains a model to decide how to use the auxiliary tasks to improve the performance on the target task. Some works have proven the usefulness of this approach in tasks like intent detection (BHATHIYA; THAYASIVAM, 2020), machine translation (GU *et al.*, 2018), and natural language generation in task-oriented dialogue systems (MI *et al.*, 2019).

Another approach that is being widely used is adversarial training (AT) (GOODFELLOW *et al.*, 2014). It emerged to solve a transfer learning problem related to the feature mismatching between a pre-training and a specific domain. In this manner, AT helps to prevent models from learning feature representations specific to a domain/language. Some examples of this approach can be seen in the work of Grießhaber, Vu and Maucher (2020), which tried to learn domain-independent representations using adversarial training and the proposed by Kim *et al.* (2017), that aimed to build language-independent representations for cross-lingual transfer.

2.4 Resources and Tools

2.4.1 PropBank Project

The Propositional Bank or PropBank (PALMER; GILDEA; KINGSBURY, 2005) is a project developed for English that is composed of sentences with annotations on semantic roles [9]. In general, PropBank adds information about semantic roles to the Penn Treebank syntactic structures (MARCUS; MARCINKIEWICZ; SANTORINI, 1993). The project aimed to create a large corpus for improving Machine Learning methods in Semantic Role Labeling and allowing to analyze syntactic variations of verbs.

Unlike other resources that use specific names for different semantic roles, which makes it hard to define a general set of semantic roles, PropBank defines semantic roles in a more general way. This way, the arguments of a verb are enumerated, ranging from zero to five (called ArgNs)[10]. Additionally, PropBank defines other roles (most of them adjunct)[11], which are more general, called ArgMs.

In addition to providing a corpus that comprises sentences annotated with semantic roles, PropBank provides a lexicon that contains information about the semantic roles and predicate-argument structures for each entry. Figure 7 presents a full description of the frame file for the verb "die". The components of the frame file are described as follows:

- Roleset: the set of semantic roles that can be used in a frameset;
- Frameset: comprises a roleset plus syntactic frames in which a predicate participates. Represents the direction of entry; and,
- Frame file: is a collection of framesets. The polysemy of the predicate generates the framesets.

For Brazilian Portuguese, there is a PropBank version called PropBank.Br (DURAN; ALUÍSIO, 2012). This project aimed to annotate a Brazilian Portuguese Treebank with semantic roles following the guidelines of the PropBank project. The annotated corpus was the Bosque corpus (belonging to the Floresta Sintá (c) tica) (AFONSO *et al.*, 2002), which is annotated by the parser PALAVRAS (BICK, 2000) and manually revised by linguists. An early version of PropBank.Br contained 1,068 verbs, and 6,142 instances were annotated, representing less than 10% of the size of English PropBank. Hence, the next goal was to increase the corpus. In order to achieve this goal, the lexical resource called Verbo-Brasil (DURAN; MARTINS; ALUÍSIO, 2013) was built, and this served as a basis for annotating a larger corpus.

[9] A semantic role describes the relationship between a predicate (which can be a verb, a name, an adjective, or an adverb) and its arguments.

[10] An argument is a constituent required by a verb.

[11] An adjunct is that which has no mandatory presence in the sentence. Besides, the sentence's meaning is not lost without its presence.

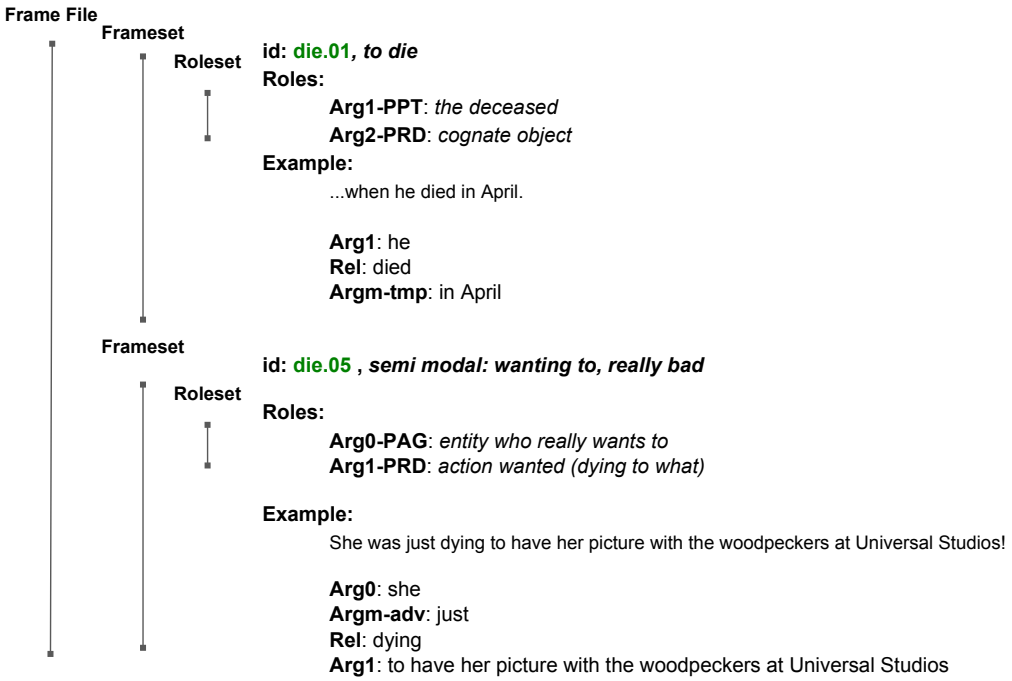


Figure 7 – Frame file for the verb "die". Extracted from http://verbs.colorado.edu/propbank/framesets-english-aliases/die.html

.

The latest version of Verbo-Brasil (DURAN; ALUÍSIO, 2015) contains 2,598 frame files, and 541 of them were manually reviewed. These 541 frame files represent verbs with more than 1000 occurrences in the corpus. Finally, Verbo-Brasil is available through a web interface for searching[12]. Some of the framesets included in it have alignments with English PropBank framesets and with the VerbNet project (KIPPER-SCHULER, 2005).

2.4.2 AMR Tools for Brazilian Portuguese

Among the applications developed for AMR in Brazilian Portuguese, semantic parsers stand out since these can be used for creating additional corpora in data augmentation strategies. Anchiêta and Pardo (2018) proposed a semantic parser based on rules due to the corpus size for Portuguese. The authors used the information extracted from the syntactic parser proposed by Bick (2000) and the information extracted from the semantic role labeling system proposed by Hartmann, Duran and Aluísio (2016) to apply rules on them in order to parse sentences. In addition to this, Anchiêta and Pardo (2022) adapted three additional methods for semantic parsing, two transition-based methods (DAMONTE; COHEN, 2018; WANG; XUE; PRADHAN, 2015) (called AMREager and CAMR, respectively), and a character-based neural models (NOORD; BOS, 2017) (called NeuralAMR). However, the author points out that the rules-based parser obtained the best results. Table 1 shows the performance of each approach in terms of Smatch (CAI; KNIGHT, 2013) and SEMA (Anchieta; Cabezudo; Pardo, 2019). In addition, Seno *et al.* (2022) proposes an AMR parser for Portuguese that follows a cross-lingual approach. The

[12] Available at http://143.107.183.175:21380/verbobrasil/#. Accessed on October 15, 2018.

authors utilized a pre-existing English parser along with multiple bilingual resources in English and Portuguese to transfer the semantic knowledge present in English into equivalent meaning representation in Portuguese.

Table 1 – AMR Parsing Results on Brazilian Portuguese corpus divided in short and long sentences.

	Smatch		SEMA	
	Short	Long	Short	Long
CAMR	0.54	0.44	0.42	0.25
AMREager	0.52	0.41	0.40	0.21
NeuralAMR	0.20	0.15	0.10	0.09
Rule-based	0.66	0.49	0.48	0.28

Source: Anchiêta and Pardo (2022).

Finally, another helpful tool in implementing AMR parsers and natural language generators is the aligner between the reference text and the concepts included in the AMR graph. Alignments are helpful because they allow systems to learn mapping rules between the AMR graph and the reference text and learn how to linearize AMR graphs.

In the literature, there are two types of aligners widely used in AMR. The first one (called JAMR) is based on spans and proposed by Flanigan *et al.* (2014) which automatically aligns word segments with concept fragments from the AMR graph through a search in a set of predefined rules. The second aligner is a token-based one proposed by Pourdamghani *et al.* (2014), which aligns concepts and relationships of the AMR graph with the tokens of the reference text through MGIZA ++ (GAO; VOGEL, 2008), used in automatic translation systems. Additionally, a span-based unsupervised aligner applied in Portuguese is proposed by Anchiêta and Pardo (2020), outperforming previous work in intrinsic and extrinsic (AMR parsing) evaluation. Table 2 shows the performance of each aligner on the Brazilian Portuguese AMR corpus.

Table 2 – Results of AMR aligners on Brazilian Portuguese AMR corpus

Aligner	Precision	Recall	F-Score
JAMR (FLANIGAN *et al.*, 2014)	0.71	0.86	0.78
Unsupervised (POURDAMGHANI *et al.*, 2014)	0.48	0.58	0.53
(ANCHIÊTA; PARDO, 2020)	0.86	0.91	0.89

Source: Anchiêta and Pardo (2020).

2.4.3 Syntax corpora and tools

As commented in the previous chapter, syntax information can be helpful in the AMR-to-text generation task. Cao and Clark (2019) use constituency trees as an intermediate representation in text generation since they claim that constituency trees have the advantage of a well-defined linearization order compared to dependency trees. Besides, constituency trees may be easier to realize, as they effectively correspond to the bracketing of the surface form. On the other hand, most of the work has used dependency trees as intermediate representations by

leveraging the similarities between AMR and dependency trees (LAMPOURAS; VLACHOS, 2017; MILLE *et al.*, 2017). Additionally, the Universal Dependencies (UD) project emerged as a multilingual project that seeks to develop cross-linguistically consistent treebank annotation for many languages, aiming to capture similarities as well as idiosyncracies among typologically different languages (including Portuguese) (NIVRE *et al.*, 2016).

Concerning the syntax corpora for Brazilian Portuguese, there is a corpus with constituency and dependency annotations called the Bosque corpus[13]. Such corpus is a subset of the Floresta Sinta(c)tica treebank (AFONSO *et al.*, 2002). It consists of news running text from Portugal and Brazil, chunked into sentences, syntactically analyzed in tree structures produced by the PALAVRAS parser (BICK, 2000) and fully revised by linguists. In addition to this, the UD project contains additional three corpora. The first is the one proposed by Rademaker *et al.* (2017), who annotated the Bosque corpus under the UD guidelines[14]. The second is one converted from the Google Universal Dependency Treebank (MCDONALD *et al.*, 2013)[15], which contains 12,078 annotated sentences. The last one is one based on Parallel Universal Dependencies Treebanks (ZEMAN *et al.*, 2017) that comprises 1,000 annotated sentences.

There are two well-known constituency parsers for Brazilian Portuguese, the LX-Parser (SILVA; BRANCO; GONÇALVES, 2010)[16], which is a probabilistic, robust constituency parser, and the PALAVRAS parser (BICK, 2000)[17], which is a rule-based parser. In the case of dependency parsing, and in particular UD parsing , the most known are the parser provided by the library Spacy[18] and the parser UDPipe (STRAKA; HAJIČ; STRAKOVÁ, 2016)[19].

Other corpora focused on natural language generation were presented in the shared task of multilingual surface realization (MSR-ST) (MILLE *et al.*, 2018). These corpora are based on Universal Dependencies[20] and have been built for different languages. The MSR-ST presented two tracks; the first one called Shallow track starts from syntactic structures in which word order information has been removed and tokens have been lemmatized, and the last one, called Deep Track, which starts from more abstract structures (similar to AMR) from which, additionally, functional words (in particular, auxiliaries, functional prepositions and conjunctions) and surface-oriented morphological information have been removed. The dataset of the shallow track includes Arabic, Czech , Danish, English, Finnish, French, Italian, Portuguese, Russian and Spanish. The deep track datasets include English, Spanish, and French.

[13] The last version of this corpus is the 8.0 contains 9,368 annotated sentences (FREITAS; ROCHA; BICK, 2008), and it is available at <https://www.linguateca.pt/Floresta/corpus.html#download>.
[14] Available at <https://github.com/UniversalDependencies/UD_Portuguese-Bosque>.
[15] Available at <https://github.com/UniversalDependencies/UD_Portuguese-GSD>.
[16] Available at <http://lxcenter.di.fc.ul.pt/tools/en/conteudo/LXParser.html>.
[17] Available at <https://visl.sdu.dk/visl/pt/parsing/automatic/trees.php>.
[18] Available at <https://spacy.io/models/pt>.
[19] Available at <https://ufal.mff.cuni.cz/udpipe>.
[20] Available at <https://github.com/UniversalDependencies/universaldependencies.github.io>. Accessed October 16, 2020.

2.5 Final Considerations

This chapter presented an overview of all concepts and resources used in this thesis. A synthesis of how all the concepts, resources and tools are applied are described as follows:

- Strategies for Low-Resource Natural Language Generation: this thesis explores two of the three mentioned approaches: (1) the use of additional labeled data generation strategies via data augmentation and cross-lingual approaches using the English AMR corpus[21] as starting point or pivot and (2) the use of transfer learning from pre-training models and syntax knowledge.
- AMR Tools for Brazilian Portuguese: most approaches explored in this thesis use token-based alignments. Therefore, the AMR aligner proposed by Pourdamghani *et al.* (2014) is used for some experiments even though unsupervised AMR aligner produced the worst performance in the work proposed by Anchiêta and Pardo (2020).
- NLG evaluation: the NLG evaluation conducted in this thesis includes automatic metrics such as BLEU, METEOR, chrF++ and BERTScore, and a manual revision of the errors produced by the current models.

[21] The AMR version used in this thesis is available at <https://catalog.ldc.upenn.edu/LDC2017T10>.

CHAPTER

3

LITERATURE REVIEW

This chapter presents a literature review on Natural Language Generation for Brazilian Portuguese. The chapter is divided in two sections. The first section brings one of the papers we publish in this work about the Natural Language Generation area for Brazilian Portuguese and emphasizes the works based on Semantic Representations, mainly the ones focused on Abstract Meaning Representation (CABEZUDO; PARDO, 2019). Finally, the second section complements the previous section with an updated revision of the literature about AMR-to-Text Generation.

3.1 NLG Overview

This section encompasses the paper below.

CABEZUDO, M. A. S.; PARDO, T. Natural language generation: Recently learned lessons, directions for semantic representation-based approaches, and the case of Brazilian Portuguese language. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. Florence, Italy: Association for Computational Linguistics, 2019. p. 81–88. Available at <https://aclanthology.org/P19-2011/>.

Natural Language Generation: Recently Learned Lessons, Directions for Semantic Representation-based Approaches, and the case of Brazilian Portuguese Language

Marco Antonio Sobrevilla Cabezudo and Thiago Alexandre Salgueiro Pardo
Interinstitutional Center for Computational Linguistics (NILC)
Institute of Mathematical and Computer Sciences, University of São Paulo
São Carlos/SP, Brazil
msobrevillac@usp.br, taspardo@icmc.usp.br

Abstract

This paper presents a more recent literature review on Natural Language Generation. In particular, we highlight the efforts for Brazilian Portuguese in order to show the available resources and the existent approaches for this language. We also focus on the approaches for generation from semantic representations (emphasizing the Abstract Meaning Representation formalism) as well as their advantages and limitations, including possible future directions.

1 Introduction

Natural Language Generation (NLG) is a promising area in Natural Language Processing (NLP) community. NLG aims to build computer systems that may produce understandable texts in English or other human languages from some underlying non-linguistic representation of information (Reiter and Dale, 2000). Tools generated by this area are useful for other applications like Automatic Summarization, Question-Answering Systems, and others.

There are several efforts in NLG for English[1]. For example, one may see the works of Krahmer et al. (2003) and Li et al. (2018), which focused on referring expressions generation, and the work of (Gatt and Reiter, 2009), focused on developing a surface realisation tool called SimpleNLG. One may also easily find other works that tried to generate text from semantic representations (Flanigan et al., 2016; Ferreira et al., 2017; Puzikov and Gurevych, 2018b).

For Brazilian Portuguese, there are few works, some of them focused on representations like Universal Networking Language (UNL) (Nunes et al., 2002) or *Resource Description Framework* (RDF) (Moussallem et al., 2018), and other ones that are very specific to the Referring Expression Generation (Pereira and Paraboni, 2008; Lucena et al., 2010) and Surface Realisation tasks (Oliveira and Sripada, 2014; Silva et al., 2013).

More recently, several representations have emerged in the NLP area (Gardent et al., 2017; Novikova et al., 2017; Mille et al., 2018). In particular, Abstract Meaning Representation (AMR) has gained interest from the research community (Banarescu et al., 2013). It 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. This representation includes information about semantic roles, named entities, wiki entities, spatial-temporal information, and co-references, among other information.

AMR has gained attention mainly due to its simplicity to be read by humans and computers, its attempt to abstract away from syntactic idiosyncrasies (focusing only on semantic processing) and its wide use of other comprehensive linguistic resources, such as PropBank (Palmer et al., 2005) (Bos, 2016).

For English, there is a large AMR-annotated corpus that contains 39,260 AMR-annotated sentences[2], which allows deeper studies in NLG and experiments with different approaches (mainly statistical approaches). This may be evidenced in the SemEval-2017 shared-task 9 (May and Priyadarshi, 2017)[3].

For Brazilian Portuguese, Anchiêta and Pardo (2018) built the first corpus using sentences from the "The Little Prince" book. The authors took advantage of the alignment between the English and Brazilian Portuguese versions of the book to import the AMR structures from one language to

[1] Most of the works may be found in the main NLP publication portal at https://www.aclweb.org/anthology/

[2] Available at https://catalog.ldc.upenn.edu/LDC2017T10.

[3] Available at http://alt.qcri.org/semeval2017/task9/.

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 81–88
Florence, Italy, July 28 - August 2, 2019.

another (but also performing the necessary adaptations). They had to use the Verbo-Brasil repository (Duran et al., 2013; Duran and Aluísio, 2015), which is a PropBank-like resource for Portuguese. Nowadays, there is an effort to build a larger AMR-annotated corpus that is similar to the current one available for English.

In this context, this study presents a literature review on Natural Language Generation for Brazilian Portuguese in order to show the resources (in relation to semantic representations) that are available for Portuguese and the existent efforts in the area for this language. We focus on the NLG approaches based on semantic representations and discuss their advantages and limitations. Finally, we suggest some future directions to the area.

2 Literature Review

The literature review was based on the following research questions:

- What was the focus of the existent NLG efforts for Portuguese and which resources were used for this language?
- What challenges exist in the NLG approaches?
- What are the advantages and limitations in the approaches for NLG from semantic representations, specially Abstract Meaning Representation?

Such issues are discussed in what follows.

2.1 Natural Language Generation for Portuguese

In general, we could find few works for Portuguese (considering the existing works for English). These works focus mainly on the referring expression generation (Pereira and Paraboni, 2008; Lucena et al., 2010) and surface realization tasks (Silva et al., 2013; Oliveira and Sripada, 2014), usually restricted to specific domains and applications (like undergraduate test scoring). Nevertheless, there are some recent attempts focused on other tasks and in more general domains (Moussallem et al., 2018; Sobrevilla Cabezudo and Pardo, 2018).

Among the NLG approaches, we may highlight the use of templates (Pereira and Paraboni, 2008; Novais et al., 2010b), rules (Novais and Paraboni, 2013) and language models (LM) (Novais et al., 2010a). In general, these approaches were successful because they were focused on restricted domains. Specifically, template-based methods used basic templates to build sentences. Similarly, some basic rules involving noun and verbal phrases were defined to build sentences. Finally, LM-based methods applied a two-stage strategy to generate sentences. This strategy consisted in generating surface realization alternatives and selecting the best alternative according to the language model.

In the case of LM-based methods, we may point out that classical LMs (based on n-grams) were not suitable because it was necessary to use a large corpus to deal with sparsity of data. Sparsity is a big problem in morphologically marked languages like Portuguese. In order to solve the sparsity of the data, some works used Factored LMs, obtaining better results than the classical LMs (de Novais et al., 2011).

In relation to NLG from semantic representations for Portuguese, we may point out the work of Nunes et al. (2002) (focused on Universal Language Networking), and Moussallem et al. (2018) (focused on ontologies). Another representation was the one proposed by Mille et al. (2018) (based on Universal Dependencies), which is based on syntax instead of semantics.

In relation to NLG tools, we highlight PortNLG (Silva et al., 2013) and SimpleNLG-BP (Oliveira and Sripada, 2014) as surface realisers that were based on SimpleNLG initiative (Gatt and Reiter, 2009)[4]. Finally, other NLG works aimed to build NLP applications, e.g., for structured data visualization and human-computer interaction purposes (Pereira et al., 2012, 2015).

2.2 Natural Language Generation from Semantic Representations

Recently, the number of works on NLG from semantic representations has increased. This increase is reflected in the shared tasks WebNLG (Gardent et al., 2017), E2E Challenge (Novikova et al., 2017), Semeval Task-9 (May and Priyadarshi, 2017) and Surface Realization Shared-Task (Belz et al., 2011; Mille et al., 2018).

In general, there is a trend to apply methods based on neural networks. However, methods

[4]Specifically, SimpleNLG-BP was built using the French version of SimpleNLG due to the similarities between both languages.

based on templates, transformation to intermediate representations and language models have shown interesting results. It is also worthy noticing that most of these methods have been applied to English, except for the methods presented in the shared-task proposed by Mille et al. (2018).

In relation to the shared-tasks mentioned before, we point out that the one proposed by Belz et al. (2011) and Mille et al. (2018) (based on Universal Dependencies) used syntactic representations. Specifically, they presented two tracks, one focused on word reordering and inflection generation (superficial track), and other that focused on generating sentences from a deep syntactic representation that is similar to a semantic representation (deep track). Furthermore, these tasks focused on several languages in the superficial task (including Portuguese) and three languages in the deep track (English, Spanish, and French).

Among the methods used for the superficial track in these shared-tasks, we may highlight the use of rule-based methods and language models in the early years (Belz et al., 2011) and a wide application of neural models in recent years (Mille et al., 2018). In the case of the deep track, it is possible to notice that rule-based methods were applied in the first competition, and methods based on transformation to intermediate representations and based on neural models were applied in the last competition.

The results in these tasks showed that approaches based on transformation to intermediate representations obtained poor results in the automatic evaluation due to the great effort in building transformation rules for their own systems. However, they usually showed better results in human evaluations. This may be explained by the maturity of the original proposed systems. This way, although the coverage of the rules was not good, the results were good from a human point of view.

Differently from the approach mentioned before, methods based on neural models (deep learning) obtained the best results. However, some methods used data augmentation strategies to deal with data sparsity (Elder and Hokamp, 2018; Sobrevilla Cabezudo and Pardo, 2018).

One point to highlight is that the results for Portuguese were poor (compared to similar languages like Spanish). Two reasons to explain this issue are related to the amount of data for Portuguese in this task (less than English or Spanish) and the quality of the existing models for related tasks that were used. Another point to highlight is the division of the general task into two sub-tasks: linearisation and inflection generation. Puzikov and Gurevych (2018a) pointed out that there is a strong relation between the linearisation and the inflection generation, and, thus, both sub-tasks should be treated together.

In contrast to Puzikov and Gurevych (2018a), (Elder and Hokamp, 2018) showed that incorporating syntax and morphological information into neural models did not bring significant contribution in the generation process, but incorporated more difficulty in the task.

Finally, it is important to notice the proposal of Madsack et al. (2018), which trained linearisation models using the dataset for each language independently and in a joint way, using multilingual embeddings. Although the results of this work did not present a lot of variation when used for all languages together, this work suggests that it is possible to train systems with similar languages (for example, Spanish and French) in order to take advantage of the syntax similarities and to overcome the problems of lack of data.

In relation to other used representations (Gardent et al., 2017; Novikova et al., 2017), a large number of works based on deep learning strategies were proposed, obtaining good results. However, the use of pipeline-based methods yielded promising results regarding grammar and fluency criteria in a joint evaluation (for RDF representation), but these methods (which usually use rules) obtained the worst results in the E2E Challenge.

Methods based on Statistical Machine Translation kept a reasonable performance (ranking 2nd in RDF Shared-Task), obtaining good results when evaluating the grammar. The explanation for this result comes from the ability to learn complete phrases. Thus, these methods may generate grammatically correct phrases, but with poor general fluency and dissimilarity to the target output. Finally, methods based on template obtained promising results in restricted domains, like in the E2E Challenge.

2.3 Natural Language Generation from Abstract Meaning Representation

In relation to generation methods from Abstract Meaning Representation, it was possible to highlight approaches based on machine translation

(Pourdamghani et al., 2016; Ferreira et al., 2017), on transformation to intermediate representations (Lampouras and Vlachos, 2017; Mille et al., 2017), on deep learning models (Konstas et al., 2017; Song et al., 2018), and on rule extraction (from graphs and trees) (Song et al., 2016; Flanigan et al., 2016).

Methods based on transformation into intermediate representations focused on transforming AMR graphs into simpler representations (usually dependency trees) and then using an appropriate surface realization system. Authors usually took advantage of the similarity between dependency trees and AMR graphs to map some results. However, some problems in this approach were the need to manually build transformation rules (excepting for Lampouras and Vlachos (2017), who automatically perform this) and the need of alignments between the AMR graph and intermediate representations, which could bring noise into the generation process. Overall, this approach presented poor results (compared to other approaches) in automatic evaluations[5]

Methods based on rule extraction obtained better results than the approach mentioned previously. This approach tries to learn conversion rules from AMR graphs (or trees) to the final text. First methods of this approach tried to transform the AMR graph into a tree before learning rules. As (Song et al., 2017) mentioned, these methods suffer with the loss of information (by not using graphs and being restricted to trees), due to its projective nature. Likewise, (Song et al., 2016) and (Song et al., 2017) could suffer from the same problem (ability to deal with non-projective structures) due to their nature to extract and apply the learned rules. Furthermore, these methods used some manual rules to keep the text fluency. However, these rules did not produce a statistically significant increase in the performance, when compared to learned rules.

Some problems of this approach are related to: (1) the need of alignments between AMR graph and the target sentence, as the aligners could lead to more errors (depending of the performance) in the rule extraction process; (2) the argument realization modeling (Flanigan et al., 2016; Song et al., 2016); and (3) the data sparsity in the rules, as some rules are too specific and there is a need to generalize them.

Methods based on Machine Translation usually outperformed other methods. Specifically, methods based on Statistical Machines Translation (SMT) outperformed methods based on Neural Machine Translation (NMT), which use data augmentation strategies to improve their performance (Konstas et al., 2017). In general, both SMT and NMT-based methods explored some preprocessing strategies like delexicalisation[6], compression[7] and graph linearisation[8] (Ferreira et al., 2017)

In relation to the linearisation, the proposals of Pourdamghani et al. (2016) and Ferreira et al. (2017) depended on alignments to perform linearisation. Both works point out that the way linearisation is carried out affects performance, thus, linearisation is an important preprocessing strategy in NLG. However, Konstas et al. (2017) show that linearisation is not that important in NMT-based methods, as the authors propose a data augmentation strategy, decreasing the effect of the linearisation.

In relation to compression, the dependency of alignments also occurred. Moreover, it is necessary a deep analysis to determine the usefulness of compression. On the one hand, compression contributed positively in the SMT-based methods but, on the other hand, it was harmful in NMT-based methods (Ferreira et al., 2017). It is also important to point out that both compression and linearisation processes were executed in sequence in these works. This could be harmful, as the order of execution could lead to loss of information.

Finally, according to (Ferreira et al., 2017), delexicalisation produces an increase and decrease of performance in NMT-based and SMT-based methods, respectively. An alternative to deal with data sparsity is to use copy mechanisms, which have shown performance increase in NLG methods (Song et al., 2018).

Some limitations of these methods were the alignment dependency (similar to the previous approaches) and the linearisation of long sentences. NMT-based methods could not represent or capture information for long sentences, producing un-

[5]Except for the work of Gruzitis et al. (2017), who incorporated the system proposed by Flanigan et al. (2016) into their pipeline.

[6]Delexicalisation aims to decrease the data sparsity by replacing some common tokens by constants.

[7]Compression tries to keep important concepts and relations in the text generation process.

[8]Linearisation tries to transform the graph into a sequence of tokens.

satisfactory results.

In order to solve these problems, methods based on neural models proposed Graph-to-Sequence architectures to better capture information from AMR graphs. This architecture showed better results than its predecessors, requiring less training data (augmented data) (Beck et al., 2018).

The main difficulty associated to deep learning is the need of large corpora to get better results. Thus, this could be hard to get for languages like Portuguese, as there are no large available corpora as there are for English.

3 Conclusions and Future Directions

This work showed a more recent literature review on NLG, specially those based on semantic representations and for Brazilian Portuguese language. As it may be seen, NLG works for Portuguese were mainly focused on Referring Expression Generation and Surface Realisation. There were a few recent works about NLG from semantic representations like ontologies or Universal Dependencies (although this last one is of syntactic nature), producing poor results.

Some resources for Portuguese were found (additional to AMR-annotated corpus), as corpora for generation from RDF (Moussallem et al., 2018) and from Universal Dependencies (Mille et al., 2018). This opens the possibility to explore the use of other resources for similar tasks in order to improve the AMR-to-Text generation. There are also corpora for languages that are relatively similar to Portuguese. Considering the proposal of Madsack et al. (2018), to learn realisations from languages that share some characteristics with Portuguese (like French or Spanish) is a reasonable alternative.

Among other strategies to deal with lack of data, it is possible to consider Unsupervised Machine Translation and back-translation strategies. The first one tries to learn without parallel corpora (these would be a corpus of AMR graphs and a corpus of sentences). This strategy has proven to be useful in this context (Lample et al., 2018a,b; Freitag and Roy, 2018). In this case, it would be necessary to extend the corpus of AMR annotations, which could represent one of the challenges. The second one aims to generate corpus in a target language (Portuguese) from other languages (as English) in order to increase the corpus size and reduce the data sparseness. In this case, it is necessary to evaluate the influence of the quality of translations and how this affects the performance of the text generator.

Additionally to the above issue, there are currently large corpora for Portuguese (for example, the corpus used by Hartmann et al. (2017)), which may allow to train robust language models.

The main challenges for Portuguese are its morphologically marked nature and its high syntactic variation[9]. These challenges contribute to data sparseness. Thus, two-stage strategies might not be useful, producing an explosion in the search for the best alternative. Moreover, to treat syntactic ordering and inflection generation together could lead to the introduction of more complexity into the models. Therefore, to tackle NLG for Portuguese as two separate tasks seems to be a good alternative, reducing the complexity of the syntactic ordering and treating inflection generation as a sequence labeling problem.

Among the challenges associated to the methods found in the literature, we may highlight two: (1) the alignment dependency, and (2) the need to better understand the semantic representations (in our case, the AMR graphs) to be able to deduce how they may be syntactically and morphologically realized.

Several approaches need alignments to learn rules and ways to linearise and compress data in AMR graphs. This is a problem because there is a need to manually align AMR graphs and target sentences in order to allow the tools to learn to align by themselves and, then, to introduce these tools into some existent NLG pipeline. Thus, limitations in the aligners may lead to errors in the NLG pipeline. This problem could be bigger in NLG for Portuguese as there is limited resources, and some of these do not present alignments. To solve this, it is possible to use approaches those are not constrained by explicit graph-to-text alignments (for example, graph-to-sequence architectures). Furthermore, this could help to join all the available resources for similar tasks (i. e., corpora for other semantic representations), with no need of alignments, in a easy way and train a semantic representation-independent text generation method. However, it is necessary to measure the usefulness of this approach, comparing it with traditional methods.

[9]The interested reader may find an overview of Portuguese characteristics at http://www.meta-net.eu/whitepapers/volumes/portuguese.

Finally, to better understand a semantic representation (and what it means) is very important, as one may better learn the possible syntactic realisations and, therefore, to give a better clue of how sentences may be morphologically constructed. For Portuguese, there is a challenge to deal with different semantic representations. Although the concepts may be shared among different semantic representations, relations are not the same, and the decision on how to code them could generate some problems in the NLG training.

Acknowledgments

The authors are grateful to CAPES and USP Research Office for supporting this work.

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 (LREC 2018)*, Miyazaki, Japan. 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.

Daniel Beck, Gholamreza Haffari, and Trevor Cohn. 2018. Graph-to-sequence learning using gated graph neural networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 273–283. Association for Computational Linguistics.

Anja Belz, Michael White, Dominic Espinosa, Eric Kow, Deirdre Hogan, and Amanda Stent. 2011. The first surface realisation shared task: Overview and evaluation results. In *Proceedings of the 13th European Workshop on Natural Language Generation*, ENLG '11, pages 217–226, Stroudsburg, PA, USA. Association for Computational Linguistics.

Johan Bos. 2016. Expressive power of abstract meaning representations. *Computational Linguistics*, 42(3):527–535.

Magali Sanches Duran and Sandra M. 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, *SEM 2015*, pages 216–221, Denver, Colorado, USA. Association for Computational Linguistics.

Magali Sanches Duran, Jhonata Pereira Martins, and Sandra Maria Aluísio. 2013. Um repositório de verbos para a anotação de papéis semânticos disponível na web (a verb repository for semantic role labeling available in the web) [in portuguese]. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*.

Henry Elder and Chris Hokamp. 2018. Generating high-quality surface realizations using data augmentation and factored sequence models. In *Proceedings of the First Workshop on Multilingual Surface Realisation*, pages 49–53. 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.

Jeffrey Flanigan, Chris Dyer, Noah A. Smith, and Jaime G. Carbonell. 2016. Generation from abstract meaning representation using tree transducers. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 731–739, San Diego California, USA. Association for Computational Linguistics.

Markus Freitag and Scott Roy. 2018. Unsupervised natural language generation with denoising autoencoders. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3922–3929, Brussels, Belgium. Association for Computational Linguistics.

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

Albert Gatt and Ehud Reiter. 2009. Simplenlg: A realisation engine for practical applications. In *Proceedings of the 12th European Workshop on Natural Language Generation*, pages 90–93, Stroudsburg, PA, USA. Association for Computational Linguistics.

Normunds Gruzitis, Didzis Gosko, and Guntis Barzdins. 2017. Rigotrio at semeval-2017 task 9: Combining machine learning and grammar engineering for amr parsing and generation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 924–928. 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. Sociedade Brasileira de Computação.

Ioannis Konstas, Srinivasan Iyer, Mark Yatskar, Yejin Choi, and Luke Zettlemoyer. 2017. Neural amr: Sequence-to-sequence models for parsing and generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 146–157, Vancouver, Canada. Association for Computational Linguistics.

Emiel Krahmer, Sebastiaan Van Erk, and André Verleg. 2003. Graph-based generation of referring expressions. *Computational Linguistics*, 29(1):53–72.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018b. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5039–5049, Brussels, Belgium. Association for Computational Linguistics.

Gerasimos Lampouras and Andreas Vlachos. 2017. Sheffield at semeval-2017 task 9: Transition-based language generation from amr. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 586–591. Association for Computational Linguistics.

Xiao Li, Kees van Deemter, and Chenghua Lin. 2018. Statistical NLG for generating the content and form of referring expressions. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 482–491, Tilburg University, The Netherlands. Association for Computational Linguistics.

Diego Jesus De Lucena, Ivandré Paraboni, and Daniel Bastos Pereira. 2010. From semantic properties to surface text: the generation of domain object descriptions. *Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial*, 14(45):48–58.

Andreas Madsack, Johanna Heininger, Nyamsuren Davaasambuu, Vitaliia Voronik, Michael Käufl, and Robert Weißgraeber. 2018. Ax semantics' submission to the surface realization shared task 2018. In *Proceedings of the First Workshop on Multilingual Surface Realisation*, pages 54–57. Association for Computational Linguistics.

Jonathan May and Jay Priyadarshi. 2017. Semeval-2017 task 9: Abstract meaning representation parsing and generation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 536–545. Association for Computational Linguistics.

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 First Workshop on Multilingual Surface Realisation*, pages 1–12. Association for Computational Linguistics.

Simon Mille, Roberto Carlini, Alicia Burga, and Leo Wanner. 2017. Forge at semeval-2017 task 9: Deep sentence generation based on a sequence of graph transducers. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 920–923. Association for Computational Linguistics.

Diego Moussallem, Thiago Ferreira, Marcos Zampieri, Maria Cláudia Cavalcanti, Geraldo Xexéo, Mariana Neves, and Axel-Cyrille Ngonga Ngomo. 2018. Rdf2pt: Generating brazilian portuguese texts from rdf data. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).

Eder Miranda de Novais, Ivandré Paraboni, and Diogo Takaki Ferreira. 2011. Highly-inflected language generation using factored language models. In *Computational Linguistics and Intelligent Text Processing*, pages 429–438, Berlin, Heidelberg. Springer Berlin Heidelberg.

Eder Miranda De Novais, Thiago Dias Tadeu, and Ivandré Paraboni. 2010a. *Improved Text Generation Using N-gram Statistics*, pages 316–325. Springer Berlin Heidelberg, Berlin, Heidelberg.

Eder Miranda De Novais and Ivandré Paraboni. 2013. Portuguese text generation using factored language models. *Journal of the Brazilian Computer Society*, 19(2):135–146.

Eder Miranda De Novais, Thiago Dias Tadeu, and Ivandré Paraboni. 2010b. Text generation for brazilian portuguese: The surface realization task. In *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas*, YIWCALA '10, pages 125–131, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2017. The e2e dataset: New challenges for end-to-end generation. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 201–206. Association for Computational Linguistics.

Maria Nunes, Graças V Nunes, Ronaldo T Martins, Lucia Rino, and Osvaldo Oliveira. 2002. The decoding system for brazilian portuguese using the universal networking language (unl).

Rodrigo De Oliveira and Somayajulu Sripada. 2014. Adapting simplenlg for brazilian portuguese realisation. In *Proceedings of the Eighth International Natural Language Generation Conference, Including Proceedings of the INLG and SIGDIAL*, pages 93–94, Philadelphia, PA, USA. 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.

Daniel Bastos Pereira and Ivandré Paraboni. 2008. Statistical surface realisation of portuguese referring expressions. In *Proceedings of the 6th International Conference on Advances in Natural Language Processing*, pages 383–392, Gothenburg, Sweden.

JC Pereira, A Teixeira, and JS Pinto. 2015. Towards a Hybrid Nlg System for Data2Text in Portuguese. In *Proceedings of the 10th Iberian Conference on Information Systems and Technologies (CISTI)*, pages 1–6, Aveiro, Portugal. IEEE.

José Casimiro Pereira, António JS Teixeira, and Joaquim Sousa Pinto. 2012. Natural language generation in the context of multimodal interaction in portuguese. *Electrónica e Telecomunicações*, 5(4):400–409.

Nima Pourdamghani, Kevin Knight, and Ulf Hermjakob. 2016. Generating english from abstract meaning representations. In *Proceedings of the Ninth International Natural Language Generation Conference*, pages 21–25, Edinburgh, UK.

Yevgeniy Puzikov and Iryna Gurevych. 2018a. Binlin: A simple method of dependency tree linearization. In *Proceedings of the First Workshop on Multilingual Surface Realisation*, pages 13–28. Association for Computational Linguistics.

Yevgeniy Puzikov and Iryna Gurevych. 2018b. E2e nlg challenge: Neural models vs. templates. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 463–471. Association for Computational Linguistics.

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

Douglas Fernandes Pereira Da Silva, Eder Miranda De Novais, and Ivandré Paraboni. 2013. Um sistema de realização superficial para geração de textos em português. *Revista de Informática Teórica e Aplicada*, 20(3):31–48.

Marco Antonio Sobrevilla Cabezudo and Thiago Pardo. 2018. Nilc-swornemo at the surface realization shared task: Exploring syntax-based word ordering using neural models. In *Proceedings of the First Workshop on Multilingual Surface Realisation*, pages 58–64. Association for Computational Linguistics.

Linfeng Song, Xiaochang Peng, Yue Zhang, Zhiguo Wang, and Daniel Gildea. 2017. Amr-to-text generation with synchronous node replacement grammar. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 7–13. Association for Computational Linguistics.

Linfeng Song, Yue Zhang, Xiaochang Peng, Zhiguo Wang, and Daniel Gildea. 2016. Amr-to-text generation as a traveling salesman problem. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2084–2089, Austin, Texas. Association for Computational Linguistics.

Linfeng Song, Yue Zhang, Zhiguo Wang, and Daniel Gildea. 2018. A graph-to-sequence model for amr-to-text generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 1616–1626. Association for Computational Linguistics.

3.2 AMR-to-Text Generation

Different from earlier work (FLANIGAN *et al.*, 2016; LAMPOURAS; VLACHOS, 2017; MILLE *et al.*, 2017), most recent works are based on end-to-end neural models such as transformers (VASWANI *et al.*, 2017), pre-trained models (RAFFEL *et al.*, 2020; LEWIS *et al.*, 2020) and their variants. The only two works that do not use an end-to-end approach are the ones proposed by Cao and Clark (2019) and Manning (2019).

Cao and Clark (2019) proposed to divide the generation process in two steps: generating a syntax structure from a AMR graph, and generating the surface form from the syntax structure. On the other hand, Manning (2019) proposed a largely rule-based method, that only adds a language model and statistical linearization models to allow for more control over the output.

Other work that tries a different strategy during decoding is the one proposed by Jin and Gildea (2019). The authors propose a transition-based generator on a graph-to-sequence architecture in which the decoder predicts an action in a similar way as AMR parsers work.

About the end-to-end approaches, different strategies have been studied. Some work focused on explore different kind of encoders such as sequences, graphs or trees (DAMONTE; COHEN, 2019), and others explored some strategies to better encode the graphs (RIBEIRO; GARDENT; GUREVYCH, 2019; ZHAO *et al.*, 2020) and reduce memory usage and model complexity (ZHANG *et al.*, 2020).

With the advent of the transformer architecture (VASWANI *et al.*, 2017), works that aimed to introduce structure into the transformers and to better model the relations between indirectly connected concepts in a AMR graph were proposed (ZHU *et al.*, 2019a). Moreover, other works proposed Graph Transformers that produced better results (CAI; LAM, 2020; WANG; WAN; JIN, 2020) and strategies to "connect" distant nodes and having a better overall graph representation (JIN; GILDEA, 2020), or predict/reconstruct the AMR graphs in training/test time (WANG; WAN; YAO, 2020; BAI; SONG; ZHANG, 2020).

Recently, most works are focused on pre-trained models. For example, Mager *et al.* (2020) presented the first AMR-to-Text generation method based on a pre-trained model. The authors used GPT-2 (RADFORD *et al.*, 2019) and fine-tuned it on the AMR-to-Text generation task. Moreover, the authors studied a cycle consistency approach that consisted of generating "n" possible outputs, producing an AMR graph for each output using a off-the-shelf AMR parser, and comparing the generated AMR graph with the actual AMR graph to rerank the possible outputs.

Ribeiro *et al.* (2021b) also studied the helpfulness of pre-trained language models. The authors explored BART (LEWIS *et al.*, 2020) and T5 (RAFFEL *et al.*, 2020) and task-adaptive pre-training strategies for improving the performance of three generation tasks, including AMR-to-Text. On the other hand, Bevilacqua, Blloshmi and Navigli (2021) studied both AMR parsing and AMR-to-Text generation tasks with BART (LEWIS *et al.*, 2020) but modifying

the linearization strategy for the AMR graph. The authors explored three strategies: one based on PENMAN notation, another one based on depth-first traversal, and another one based on breadth-first traversal.

Another interesting work is the one proposed by Ribeiro, Zhang and Gurevych (2021). The authors proposed StructAdapt, an adapter method to encode graph structure into pre-trained language models, obtaining improvements. Also, the authors attempted to use this approach on smaller subsets, showing that StructAdapt performs well even when no large dataset is available, in comparison with traditional pre-trained models fine-tuned on specific tasks.

Hoyle, Marasović and Smith (2021) explored the ability of pre-trained models to encode local graph structures, in particular their invariance to the graph linearization strategy and their ability to reconstruct corrupted inputs, producing improvements. Finally, Bai, Chen and Zhang (2022) explored graph pre-training to improve the structure awareness of pre-trained language models over AMR graphs.

In summary, Table 3 lists the works presented in this section and the performance obtained in each dataset in terms of BLEU. Best works have highlighted in bold.

Table 3 – Results for all the AMR-to-Text generation methods on different AMR corpora in terms of BLEU.

Work	LDC2015E86	LDC2017T10	LDC2020T02
Konstas *et al.* (2017)	22.00	-	-
Song *et al.* (2018)	23.30	-	-
Beck, Haffari and Cohn (2018)	-	23.30	-
Cao and Clark (2019)	-	26.80	-
Damonte and Cohen (2019)	24.40	24.54	-
Manning (2019)	8.70		-
Ribeiro, Gardent and Gurevych (2019)	24.32	27.87	-
Zhu *et al.* (2019a)	29.66	31.82	-
Jin and Gildea (2019)	-	19.51	-
Cai and Lam (2020)	27.40	29.80	-
Mager *et al.* (2020)	-	33.02	-
Zhao *et al.* (2020)	30.58	32.46	-
Wang, Wan and Jin (2020)	25.90	29.30	-
Wang, Wan and Yao (2020)	32.10	33.90	-
Bai, Song and Zhang (2020)	31.48	34.19	-
Zhang *et al.* (2020)	30.80	33.60	-
Jin and Gildea (2020)	-	31.20	-
Ribeiro *et al.* (2021b)	-	45.80	-
Ribeiro, Zhang and Gurevych (2021)	-	**46.60**	**48.00**
Bevilacqua, Blloshmi and Navigli (2021)		45.30	44.90
Hoyle, Marasović and Smith (2021)	-	45.14	-
Bai, Chen and Zhang (2022)	-	**49.80**	**49.20**

Source: Elaborated by the author.

Concerning multilingual AMR-to-Text generation, three works have been found in the literature. The first was proposed by Fan and Gardent (2020) and aimed to generate sentences in 21 languages from English AMR. To do this, the authors used an AMR parser (FLANIGAN *et al.*, 2014) to annotate the English section from Europarl corpus (KOEHN, 2005) and then they use it to train a English AMR-to-XX generation task, in which XX represents a specific language (including Portuguese). In addition, the authors evaluate on the AMR corpus built by Damonte and Cohen (2018), which includes targets for Spanish, German, Italian and Chinese[1] (namely LDC2020T07). The authors used a transformer-based approach and train a model for each language and one for handling all languages in a multilingual setting.

Unlike the work of Fan and Gardent (2020), Ribeiro *et al.* (2021a) study the effect of diverse data augmentation strategies in multilingual AMR-to-Text generation as there is no gold data for non-English languages. Therefore, the authors explore the helpfulness of silver/gold AMRs and sentences in this task. Results on LDC2020T07 show that combining both silver AMRS and sentences leads to improvements.

Finally, Xu *et al.* (2021) leverages the availability of the English AMR corpus and English-to-X parallel datasets to pre-train via multi-task learning. The model is trained on the AMR parsing, AMR-to-Text generation and Machine Translation tasks together and then fine-tuned on the same tasks for different languages. Results on LDC2020T07 show that this approach surpass previous work in AMR parsing and multilingual AMR-to-Text generation.

Overall, the results of the the works previously described are presented on Table 4. It is worth noting that the only work that regarded Portuguese as part of the study is the proposed by Fan and Gardent (2020). However, the results are computed on Europarl corpus instead of the LDC2020T07 corpus.

Table 4 – Multilingual AMR-to-Text Generation BLEU scores on test set. *Results obtained by Fan and Gardent (2020) for Portuguese are obtained on Europarl corpus.

	Spanish	Italian	German	Portuguese*
Fan and Gardent (2020)	21.70	19.80	15.30	21.20
Ribeiro *et al.* (2021a)	30.70	26.40	20.60	-
Xu *et al.* (2021)	31.36	28.42	25.69	-

Source: Elaborated by the author.

3.3 Final Considerations

This chapter described an overview of the works focused on NLG for Brazilian Portuguese, the ones focused on NLG from Semantic Representations, and the ones focused on AMR-to-Text generation. In general, we highlight the following findings:

[1] Available at <https://catalog.ldc.upenn.edu/LDC2020T07>

- Some works tried to use the corresponding AMR graphs at the output side to preserve the meaning of the generated outputs, at training time (WANG; WAN; YAO, 2020; BAI; SONG; ZHANG, 2020) or at inference time (MAGER *et al.*, 2020).

- Pre-trained models such as T5 (RAFFEL *et al.*, 2020) or BART (LEWIS *et al.*, 2020) produced a quite improvement in the area. It is possible to see that the works of Ribeiro *et al.* (2021b) and Bevilacqua, Blloshmi and Navigli (2021) overcame the previous best result (BAI; SONG; ZHANG, 2020) by up to ∼14.00 in terms of BLEU.

- Earlier work attempted to augment data in order to verify the potential of the proposals to leverage this additional data. However, with the advent of pre-trained models, recent work (RIBEIRO *et al.*, 2021b; RIBEIRO; ZHANG; GUREVYCH, 2021) have paid attention to analyse how the models perform in cases where there is not a large dataset available.

- There are few works on Multilingual AMR-to-Text generation and the works focus on generating sentences in a specific language from English AMR, disregarding possible divergences between languages.

- Multilingual works explored adding automatically annotated data and evaluating its helpfulness as well as using related tasks to help the generation task in non-English languages, producing improvements.

CHAPTER

4

CORPUS ANNOTATION

This chapter presents the whole process of definition of AMR guidelines for annotation and the corpora annotation itself. The chapter is divided in two sections. The first section brings a paper about the adaptation of AMR guidelines for annotating Brazilian Portuguese news texts. Finally, the second section complements the previous section with an extension of the corpus, a comparison between the annotation of news texts and opinions and an analysis of hard cases. Both papers aim to answer the research question: *How different is English AMR corpus from Portuguese AMR corpus in terms of linguistic phenomena?*

4.1 Towards a General Abstract Meaning Representation Corpus for Brazilian Portuguese

This section encompasses the paper below.

CABEZUDO, M. A. S.; PARDO, T. Towards a General Abstract Meaning Representation Corpus for Brazilian Portuguese. In: Proceedings of the 13th Linguistic Annotation Workshop. Florence, Italy: Association for Computational Linguistics, 2019. p. 236–244. Available at <https://aclanthology.org/W19-4028/>.

Contributions:

- Adaptation of AMR guidelines for annotating news texts in Brazilian Portuguese.
- First version of a multi-genre AMR corpus for Brazilian Portuguese.

Towards a General Abstract Meaning Representation Corpus for Brazilian Portuguese

Marco Antonio Sobrevilla Cabezudo and Thiago Alexandre Salgueiro Pardo
Interinstitutional Center for Computational Linguistics (NILC)
Institute of Mathematical and Computer Sciences, University of São Paulo
São Carlos/SP, Brazil
msobrevillac@usp.br, taspardo@icmc.usp.br

Abstract

Abstract Meaning Representation (AMR) is a recent and prominent semantic representation with good acceptance and several applications in the Natural Language Processing area. For English, there is a large annotated corpus (with approximately 39K sentences) that supports the research with the representation. However, to the best of our knowledge, there is only one restricted corpus for Portuguese, which contains 1,527 sentences. In this context, this paper presents an effort to build a general purpose AMR-annotated corpus for Brazilian Portuguese by translating and adapting AMR English guidelines. Our results show that such approach is feasible, but there are some challenging phenomena to solve. More than this, efforts are necessary to increase the coverage of the corresponding lexical resource that supports the annotation.

1 Introduction

In recent years, there has been renewed interest in the Natural Language Processing (NLP) community in language understanding and dialogue. Thus, the issue of how the semantic content of language should be represented has reentered into the NLP discussion. In this context, several semantic representations, like Universal Networking Language (UNL) (Uchida et al., 1996), the semantic representation used in the Groningen Meaning Bank (Basile et al., 2012), Universal Conceptual Cognitive Annotation (UCCA) (Abend and Rappoport, 2013), and, more recently, the Abstract Meaning Representation (AMR) (Banarescu et al., 2013), have emerged.

Abstract Meaning Representation 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, wiki entities, spatial-temporal information, and co-references, among other information. AMR may be represented using logic forms (see (a) in Figure 1), PENMAN notation (see (b) in Figure 1), and graphs (see (c) in Figure 1). AMR has gained relevance in the research community due to its easiness to be read by computers and humans (as it could be represented using graphs or first-order logic, which are representations that are more familiar to computers and humans, respectively), its attempt to abstract away from syntactic idiosyncrasies (making the tasks to focus only on semantic processing) and its wide use of other comprehensive linguistic resources, such as PropBank (Bos, 2016).

In relation to its attempt to abstract away from syntactic idiosyncrasies, it may be seen that AMR annotation in Figure 1 could be generated from the sentences "The boy wants the girl to believe him." and "The boy wants to be believed by the girl.", which are semantically similar, but with different syntactic realizations. Regarding the use of linguistic resources, AMR annotation in Figure 1 shows information provided by PropBank, as the framesets "want-01" and "believe-01", and some semantic roles that they require.

The available AMR-annotated corpora for English are large, containing approximately 39,000 sentences. Some efforts have been performed for using AMR as an interlingua and building corpus for Non-English languages, taking advantage of the alignments and the parallel corpora that exist (Xue et al., 2014; Damonte and Cohen, 2018). Other works tried to adapt the AMR guidelines to other languages (Migueles-Abraira et al., 2018), considering its cross-linguistic potential.

It is unnecessary to stress the importance of corpus creation for other languages. Annotated corpora provide qualitative and reusable data for building or improving existing methods and ap-

Proceedings of the 13th Linguistic Annotation Workshop, pages 236–244
Florence, Italy, August 1, 2019.

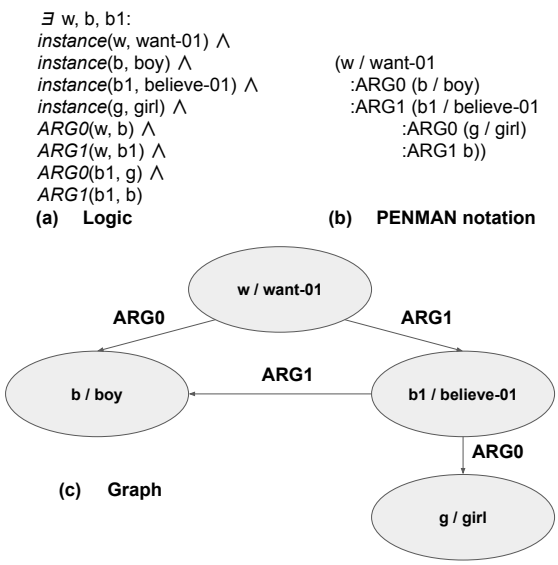


Figure 1: AMR examples

plications, as well as for serving as benchmarks to compare different approaches. In the case of Portuguese language, to the best of our knowledge, there is an unique AMR-annotated corpus, composed by the sentences of the "The Little Prince" book (Anchiêta and Pardo, 2018). The lexical resource they used to annotate some concepts was the Verbo-Brasil (Duran and Aluísio, 2015), which replicates the PropBank experience for Portuguese.

One difficulty related to the above corpus is its unusual writing style (since it is a tale) and its restricted vocabulary, which make the creation or adequacy of general purpose tools a more difficult task. More than this, the corpus is too small, hindering the development or adaptation of methods for tasks that require semantics. In this context, this work intends to show the extension process of the AMR annotation on a general purpose corpus (which covers a wide vocabulary and several domains) using the current AMR guidelines and some adaptations for Portuguese.

This paper is organized as follows. Section 2 briefly introduces some previous work that tried to build AMR corpora for Non-English languages. The corpus in Portuguese is described in Section 3. The annotation methodology and evaluation are described in Section 4 and 5, respectively. The current state of the annotation is reported in Section 6, and, finally, some concluding remarks are presented in Section 7.

2 Related Work

One of the first works that tried to build an AMR-annotated corpus for a Non-English language was proposed by Xue et al. (2014). The main goal of this work was to evaluate the potentiality of AMR to work as an interlingua. In order to achieve this goal, the authors annotated 100 English sentences of the Penn Treebank using AMR and then translated them to Czech and Chinese, which were annotated with AMR as well. Their main finding was that the level of compatibility of AMR between English and Chinese was higher than between English and Czech.

In other research line, Vanderwende et al. (2015) proposed an AMR parser to convert Logic Form representations into AMR for English. The authors also built an AMR-annotated corpus for French, German, Spanish, and Japanese.

Damonte and Cohen (2018) developed an AMR parser for English and used parallel corpora to learn AMR parsers for Italian, Spanish, German, and Chinese. The main results showed that the new parsers overcame structural differences between the languages. The authors also proposed a method to evaluate the parsers that does not need gold standard data in the target languages.

In the case of Spanish, Migueles-Abraira et al. (2018) performed a manual AMR annotation of the book "The Little Prince" using the guidelines of the AMR project. The main goal was to analyze the guidelines and to suggest some adaptions in order to cover the relevant linguistic phenomena in Spanish.

For Portuguese, Anchiêta and Pardo (2018) built the first AMR-annotated corpus taking advantage of the alignments between the book "The Little Prince" for English and Portuguese languages. Thus, the strategy consisted of importing the corresponding AMR annotation for each sentence from the English annotated corpus and revising the annotation to adapt it to Portuguese.

3 The Corpus for Brazilian Portuguese

As mentioned, the AMR-annotated corpus for Brazilian Portuguese was composed by sentences of the "The Little Prince" book (Anchiêta and Pardo, 2018). In order to broaden the annotation to other domains and text genres, our proposal focused on annotating news in several domains.

The news texts were extracted from RSS[1] from *Folha de São Paulo* news agency[2], one of the mainstream agencies in Brazil. The selected news came from different sections/domains: "daily news", "world news", "education", "environment", "sports", "science", "balance and health", "*ilustrada*", "*ilustríssima*", "power", and "technology". Additionally to these sentences, sentences of the PropBank.Br[3] (Duran and Aluísio, 2012) were collected in order to enrich the corpus (PropBank.Br already contains semantic role annotation, which makes the AMR annotation task much easier). It is important to note that PropBank.Br sentences are also from news texts.

The news download interval was from November 25th to November 28th, 2018. Overall, 249 news were collected from different domains, totalizing 7,643 sentences. The news distribution is presented in Table 1.

Section	# News	# Sentences	Avg. tokens by sentence	# Selected sentences
Daily news	48	1,521	22.94	848
World news	43	1,212	24.38	617
Education	13	426	23.72	222
Environment	4	98	25.40	45
Sports	29	875	20.93	531
Science	10	460	23.50	243
Balance and Health	6	159	23.15	88
Ilustrada	27	648	24.10	348
Ilustríssima	7	305	24.41	161
Power	51	1,677	19.93	1,121
Technology	11	262	22.55	149
Total	249	7,643	22.53	4,563

Table 1: News collection statistics

Due to the statistics observed in Table 1 and the difficulty that the task of semantic annotation carries, the scope of the work was focused on annotating only short sentences (but guaranteeing that different domains are covered). In order to define what a short sentence is, the average number of tokens by sentence was calculated and this value was used as threshold. Thus, sentences with a number of tokens below the average (in our case, it was 22.53 tokens) were selected, resulting in 4,563 sentences to be AMR annotated (indicated by the "Selected sentences" column in the table).

In relation to the PropBank.Br sentences (Duran and Aluísio, 2012), the same strategy for selecion was adopted. In total, 3,012 PropBank.Br sentences were added to our corpus.

[1]RSS stands for "Really Simple Syndication".

[2]Available at https://www.folha.uol.com.br/.

[3]PropBank.Br was the basis for the construction of the previously cited Verbo-Brasil.

4 Annotation Methodology

The proposed annotation methodology consisted of two main steps. The first step aimed to independently analyze and think about the sentence structure, while the second step counted with the aid of the AMR Editor tool (Hermjakob, 2013) to produce the AMR annotation in PENMAN format in order to export the annotation.

In relation to the first step, a sequence of actions need to be carried out in order to facilitate the second step. These actions are described as follows:

- To identify the kind of sentence to be analyzed (default, comparative, superlative, coordinate, subordinate, and others). This is useful to determine whether it is necessary to build two or more sub-graphs (in case of coordinate or subordinate sentences) and then to join them using a conjunction (usually coordinate sentences) or a concept of the main sub-graph (in the case of subordinate sentences).

- To identify concepts. Annotators must follow the AMR guidelines[4] in order to define a concept. Thus, they may identify general concepts, concepts from AMR Guidelines or concepts from Verbo-Brasil.

- To identify the main concept from the two previous steps. For example, the main verb could be the main concept in a default sentence.

- To identify the relations among the identified concepts[5].

An example of the execution of the actions is presented in Figure 2. The sentence to be analyzed is "*Ieltsin adotou outras medidas simbólicas para mostrar a perda de poderes do Parlamento.*"("Yeltsin took other symbolic measures to show the loss of Parliament's power."). This is the case of a subordinate sentence. Then, we need to identify the concepts. Thus, some words became general concepts, named-entities or Verbo-Brasil framesets. Then, it was necessary to identify the graph top (in this case, the verb "*adotar*" because

[4]Available at https://github.com/amrisi/amr-guidelines/blob/master/amr.md. Accessed on April 1st, 2019. The adopted version was the 1.2.5.

[5]The relations were extracted from Verbo-Brasil (for core relations) and AMR guidelines (for non-core relations).

it is the main verb of the main sentence "*Ieltsin adotou outras medidas simbólicas*"). Finally, the relations among all concepts were identified.

Similar to the work of Migueles-Abraira et al. (2018), our proposal tried to adapt the AMR guidelines to Brazilian Portuguese, making some modifications on it in order to deal with the specific linguistic phenomena. The general guideline used to annotate a sentence is described as follows:

- To use the framesets of Verbo-Brasil (Duran and Aluísio, 2015) to determine verb senses and the argument structure of verbs.
- To use the 3rd singular person ("*ele*") or the pronoun "that" ("*isso*") in case of NP Ellipsis, clitic or possessive pronouns. Differently from Migueles-Abraira et al. (2018), we propose to use ("*ele*") or "that" ("*isso*") as a default value. We decided to determine this guideline in order to keep some annotation pattern.
- In the case of indeterminate subject, not to use any pronoun.
- In the case of multi-word expression, to identify the one-word synonym of the expression and use it in the annotation, or define a one-word as the join of the words.
- To use the AMR framesets to annotate modal verbs, since Verbo-Brasil does not include that kind of verbs. In order to facilitate the identification of a modal verb, to try to replace by "*poder*" ("can") or "*dever*" ("should") verbs.
- In cases where the difference among two or more senses is subtle, to use the most frequent sense that satisfies the predicted argument structure.
- To use the AMR guidelines and dictionary[6] for the other cases.

The proposed annotation strategy consisted of annotating sentences of shorter size at the beginning and then increasing sentence size up to 22 tokens, according to the annotators' learning. Sentences that had verbs that were not included in the Verbo-Brasil repository were not annotated and the new verbs were put in a list in order to enrich the repository in the future.

Smatch score (Cai and Knight, 2013) was used to calculate the inter-annotator agreement. Unlike the work of Banarescu et al. (2013), which built a gold standard (using the total agreement between the annotators), the way to calculate the inter-annotator agreement consisted in comparing all annotations in an all-against-all configuration, obtaining the average of all inter-annotator agreements. Finally, the annotated versions of the sentences belonging to the agreement sample that were included in the final corpus were chosen by an adjudicator (since that more than one possible annotation exists).

5 Evaluation

In relation to the overview of the annotation process, it is important to know that the annotation team was originally composed of 14 annotators[7] that belong to the areas of Computer Science and Linguistics (all of them focused on Natural Language Processing). These annotators participated in two training sessions. In the first session, the task and the resources to be used were presented. The participants were trained by annotating sentences of PropBank.Br (Duran and Aluísio, 2012) in order to perceive the difficulty of the task. The second session aimed to answer questions about the annotation, show the inter-annotator agreement during the training stage, some common mistakes, and launch the annotation process.

5.1 Inter-annotator Agreement

The results of the inter-annotator agreement are presented in Table 2. During the training stage, the agreement was measured once in each week (with 4-5 sentences to annotate per week). Currently, the annotators are building AMR annotations for more sentences until they reach 100 sentences (as in the original AMR project) in order to have an adequate sample to measure the agreement.

In general, the Smatch was 0.72, with the minimum being 0.70 and the maximum 0.77. These results are similar to the obtained by the work of Banarescu et al. (2013) (between 0.70 and 0.80), although the number of sentences assessed in English was 100 (in our case, there were 34 sentences) and the number of annotators was 4 (we

[6] Available at https://www.isi.edu/~ulf/amr/lib/amr-dict.html. Accessed on April 1st, 2019.

[7] During the annotation process, some of the annotators gave up.

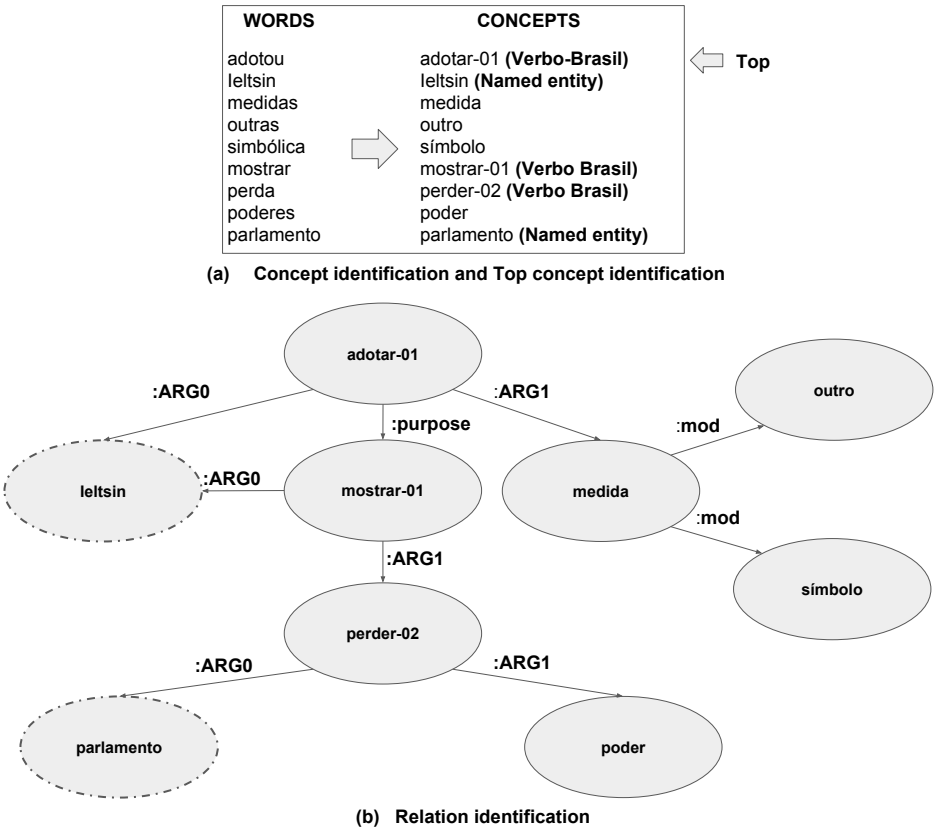


Figure 2: Example of the annotation steps

had from 5 to 7).

Week	# Annotators	# Sentences	Smatch
1	5	5	0.77
2	7	5	0.72
3	5	4	0.73
-	-	20	0.70
Total		34	0.72

Table 2: Annotation agreement

5.2 Disagreement Analysis

It is important to highlight some reasons that led to the occurring disagreements. One of the reasons was the difficulty identifying some kinds of verbs, as modal, copula, light and auxiliary verbs. Additionally, due to the use of English framesets for modal verbs, there were cases where the frameset to be used was difficult to be determined. For example, the sentence "*A quem podemos nos aliar?*" ("Who can we ally with?") was encoded as follows:

(r / **recommend-01**
:ARG1 (a / aliar-01
:ARG0 (n / nós)
:ARG1 (a2 / amr-unknown)))

(p5 / **possible-01**
:ARG1 (a8 / aliar-01
:ARG1 (n3 / nós)
:ARG2 (a9 / amr-unknown)))

As one may see, the modal verb "*poder*" was encoded as "recommend-01" and "possible-01", depending on the interpretation of the annotator. This problem occurred because a modal verb in Portuguese may be translated in different ways to English according to the context.

Another difficulty was the identification of verbs whose modality could not be easy to identify. For example, the verb "*conseguir*" (usually translated to "get") in the sentence "*Ele contou que conseguiu adquirir 20 entradas porque ofereceu Cr$ 5.000 ao bilheteiro.*" ("He said he was able to get 20 tickets because he offered Cr$ 5.000 to the ticket clerk.") was annotated using a Verbo-Brasil frameset (without modal verb) by some annotators and using the AMR frameset (for modal verb) by others. To solve this difficulty, the guidelines (adapted for Portuguese) suggested that they should try to substitute verbs for some modal verbs

as "*dever*" or "*poder*". In the previous sentence, the verb "*conseguir*" could be replaced by the verb "*poder*". This way, "*conseguir*" might be identified as a modal verb.

As for the modal verbs, the annotation of auxiliary verbs also presented some difficulties. Some annotators used the Verbo-Brasil framesets and others omitted that verb annotation, being this last one the correct way to annotate. For example, this happens for the verb "*ficar*" in the sentence "*Eles ficaram aguardando o resultado da negociação.*" ("They were waiting for the outcome of the negotiation."), where the verb fulfills an auxiliary function, and, therefore, it should not be considered in the final AMR representation.

Another difficulty was related to the identification of the verb sense in the Verbo-Brasil repository. This identification was problematic in some cases. For example, the verb "*admitir*" in the sentence "*Ele não treinava como devia, o que não admito*" ("He did not train as he should, what I do not admit") was associated to the concept "*admitir-01*" (whose meaning is related to confess or acknowledge as truth) and to the concept *admitir-02* (whose meaning is related to agree, allow, or tolerate). In this case, i.e., when the verb sense is difficult to identify, the suggestion was to select the most frequent sense (usually the first in the sense list) that covers all the arguments in the sentence.

In a similar way, sometimes the identification of the argument labels and the relations between concepts presented challenges to the annotators. For example, the word "*porque*" in the sentence "*Ele contou que conseguiu adquirir 20 entradas porque ofereceu Cr$ 5.000 ao bilheteiro.*" was associated to the relation "*cause*". However, some annotators omitted this relation.

In relation to the reference annotation, we may highlight that the annotators had disagreements in some cases, mainly when they had to choose where the reference should be inserted. For example, in the sentence "*A empresa considera os equipamentos ultrapassados e quer adquirir modelos modernos.*" ("The company considers the equipment to be outdated and wants to acquire modern models.") represented in the two following ways), the concept "*empresa*" ("company") was used as reference for "querer-01" and "adquirir-01" by some annotators and as reference only for "*querer-01*" by others.

```
(e / and
   :op1 (c / considerar-01
      :ARG0 (e2 / empresa)
      :ARG1 (e3 / equipamento)
      :ARG2 (u / ultrapassado))
   :op2 (q / querer-01
      :ARG0 e2
      :ARG1 (a2 / adquirir-01
         :ARG0 e2
         :ARG1 (m / modelo
            :mod (m2 / moderno)))))
```

```
(e / and
   :op1 (c6 / considerar-01
      :ARG0 (e / empresa)
      :ARG1 (e12 / equipamento
      :ARG2 (u2 / ultrapassado)))
   :op2 (q / querer-01
      :ARG0 e
      :ARG1 (a12 / adquirir-01
         :ARG1 (m / modelo
            :mod (m2 / moderno)))))
```

In relation to part of speech tags, we remark that there were problems in the annotation of some adjectives and nouns. In the case of adjectives, there were some difficulties to nominalize some adjectives (pertainym adjectives). For example, the adjective "*tributária*" ("tributary") in the expression "*carga tributária*" ("Tax burden") refers to a type of "*carga*" ("charge"), therefore, the concept "*tributo*" ("tribute") should be used instead of "*tributária*". In the case of nouns, there were difficulties to convert some nouns into verbs and to deal with some nouns like executors of some action. For example, the word "*competivididade*" ("competitiveness") was encoded using the concept "*competivididade*" (wrong way) and using the concept "*competir-01*" (correct way). Another example is the word "*bilheteiro*" ("ticket clerk"), which was encoded using the concept "*bilheteiro*" by some annotators. However, the correct encoding was to interpret "*bilheteiro*" as "*pessoa que vende bilhetes*" ("person that sells tickets") and, thus, encoding it as follows:

```
(p / pessoa
   :ARG0-of (v / vender-01
      :ARG1 (b / bilhete)
```

Finally, another difficulty was associated to the

use of temporal expressions. For example, the expression "*até agora*" ("until now") was encoded in several ways by the annotators. In this case, this expression was treated as fixed, using the concept "até-agora".

5.3 Common Mistakes

Some of the frequent errors made in the annotation process include the following:

- No lemmatization: there were several cases where some annotators did not use the lemmas to represent the concepts. In this way, this decreased inter-annotator agreement and could harm the annotation quality. For example, the concept "*equipamento*" ("equipment") should be used instead of "*equipamentos*" ("equipments"), and the concept "*ele*" ("he") instead of "*eles*" ("they").

- Specific characters for Portuguese: the AMR Editor tool was developed for annotating English sentences. Thus, this tool does not work well when a sentence to be annotated includes words with characters used in Portuguese like "*â*" or "*ç*". To solve this problem, it was suggested that annotators omit these characters when using the editor (replacing by one general character like "*a*" and "*c*") and then restore the correct characters as a post-editing step. However, these errors occurred, impairing the agreement.

- Variable errors or format errors: some annotators opted not to use the AMR Editor tool to build the AMR graphs, resulting in mistakes related to the number of parenthesis of the PENMAN notation and the variable declaration repetition. For example, the concept "*correr*" ("run") was represented by the variable "c" and the concept "*coelho*" ("rabbit") was also represented by the same variable, producing an error in the graph representation.

5.4 Annotation Challenges

During the annotation process (after the training stage), several challenges emerged. In what follows, some of these challenges are briefly discussed.

- Expressions or short sentences. Although the length of the sentences (or expressions) were tiny (3-5 words), expressions like "*nada demais?*", "*De quem é a culpa?*", "*Não, em hipótese alguma.*" were difficult to annotate. In some cases, it happened due to lack of context. In other cases, to identify which concepts should be included in the representation and how these concepts should be related was a hard task. This representation problem may be reflected in the inter-annotator agreement decay down to 0.70 (in comparison with the previous agreement).

- Multi-word expressions (MWE). Expressions like "*toda hora*", "*todo mundo*", or "*estar na moda*" in the sentence "*Academias especializadas estão na moda.*" were examples of multi-word expressions that annotators could not represent as a 1-word synonym (as the guideline indicates). In these cases, annotators join the words (for example, "*toda-hora*" is described as AMR dictionary suggests) or tried to separate the concepts in the graph. Another problem was the MWE identification. Expressions like "*na moda*" could be difficult to identify as a MWE and bring some challenges into the annotation.

- Particularities of Portuguese. Some expressions are specific for Portuguese or similar languages. For example, we may see a double negation in the sentence "***Não** temos **nenhuma** intelectualidade pronta.*", which does not naturally occur in English. Thus, annotators omitted one of the negations to preserve the meaning of the sentence.

- Indeterminate subjects. In some cases, the subject was indeterminate and the annotators did not annotate the reference. For example, in the sentence "*bebe-se*", the particle "*se*" did not show who is the subject, so, it was not marked in the representation.

6 Current State of the Annotation

Currently, the corpus is composed by 299 AMR-annotated sentences (considering the inter-annotator agreement sample), which include 907 concepts and 711 relations (excluding "instance", "name", and "op" relations). It is important to notice that there are 26 verbs (or verb senses) that did not appear in the Verbo-Brasil and it is necessary

to analyze them in order to increase the coverage of the repository in the future.

Table 3 and Table 4 show the statistics about the concepts and the top 10 most frequent relations annotated in the corpus. For comparison purposes, Table 4 also shows the top 10 most frequent relations annotated in the AMR-annotated corpus based on "The Little Prince" book for Brazilian Portuguese.

One point to remark in relation to Table 4 is that both corpora keep the same proportion in the first relations (the top 5); then, both show slightly different distributions. In the case of "The Little Prince", relations like "degree" and "poss" are more frequent. One reason to explain this is that tales use intensifiers like "more" or "less" and possessives like "mine" or "his" in their vocabulary. On the other hand, news texts, and the sentences and expressions contained in it, describe facts and usually use numbers to report quantities ("quant" relation). More than this, some expressions collected until now (due to their short size) describe imperatives like "*arranje!*" ("get it"). Thus, the imperative mode is frequent in the corpus. It is expected that, when the news corpus grows, these relation will change a bit.

Concepts	Frequency
General concepts	504
Verbo-Brasil concepts	235
Named entities	66
Modal verbs	20
Amr-unknown	33
Other entities and special frames	49

Table 3: Statistics of concepts in the corpus

Current corpus			"The Little Prince" corpus		
Relation	Freq.	%	Relation	Freq.	%
ARG1	173	24.33	ARG1	1,734	25.88
ARG0	140	19.69	ARG0	1,520	22.69
polarity	70	9.85	mod	678	10.12
mod	69	9.70	ARG2	454	6.78
ARG2	53	7.45	polarity	295	4.40
domain	35	4.92	time	246	3.67
quant	25	3.52	domain	211	3.15
time	23	3.23	degree	194	2.90
manner	20	2.81	manner	187	2.79
mode	17	2.39	poss	162	2.42

Table 4: Ten most frequent relations in the news corpus and in the "The Little Prince" corpus

7 Concluding Remarks

This paper showed the process of the AMR annotation on a general purpose corpus using the current AMR guidelines and some adaptations for Portuguese. In general, most of the guidelines could be translated to Portuguese. However, there were some cases that needed improvements, as the use of modal verbs and multi-word expressions. On the other hand, the adopted PropBank-like lexical resource (Verbo-Brasil) needs to increase its coverage.

As future work, besides extending Verbo-Brasil, we plan to try back-translation strategies to accelerate the annotation process.

More details about the corpus and the related ongoing work may be found at the OPINANDO project webpage[8].

Acknowledgments

The authors are grateful to CAPES and USP Research Office for supporting this work and to the several corpus annotators that have collaborated with this research.

References

Omri Abend and Ari Rappoport. 2013. Ucca: A semantics-based grammatical annotation scheme. In *Proccedings of the 10th International Conference on Computational Semantics*, pages 1–12, Potsdam, Germany. Association for Computer Linguistics.

Rafael Anchiêta and Thiago Pardo. 2018. Towards AMR-BR: A SemBank for Brazilian Portuguese Language. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*, pages 974–979, Miyazaki, Japan. 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.

Valerio Basile, Johan Bos, Kilian Evang, and Noortje Venhuizen. 2012. Developing a large semantically annotated corpus. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 3196–3200, Istanbul, Turkey. European Language Resource Association (ELRA).

[8]Available at https://sites.google.com/icmc.usp.br/opinando/

Johan Bos. 2016. Expressive power of abstract meaning representations. *Computational Linguistics*, 42(3):527–535.

Shu Cai and Kevin Knight. 2013. Smatch: an evaluation metric for semantic feature structures. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 748–752, Sofia, Bulgaria. Association for Computational Linguistics.

Marco Damonte and Shay B. Cohen. 2018. Cross-lingual abstract meaning representation parsing. In *Proceedings of the 16th 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. 2012. Propbank-br: a brazilian treebank annotated with semantic role labels. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 1862–1867, Istanbul, Turkey. European Language Resources Association (ELRA).

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.

Ulf Hermjakob. 2013. Amr editor: A tool to build abstract meaning representations.

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

Hiroshi Uchida, M Zhu, and T Della Senta. 1996. UNL: Universal networking language–an electronic language for communication, understanding, and collaboration. *Tokyo: UNU/IAS/UNL Center*.

Lucy Vanderwende, Arul Menezes, and Chris Quirk. 2015. An amr parser for english, french, german, spanish and japanese and a new amr-annotated corpus. In *Proceedings of the 13th Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 26–30, Denver, Colorado. Association for Computational Linguistics.

Nianwen Xue, Ondrej Bojar, Jan Hajic, Martha Palmer, Zdenka Uresova, 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 Language Resources Association (ELRA).

4.2 The AMR-PT Corpus: Manual Annotation of Hard Cases of Sentences from Journalistic and Opinative Texts

This section encompasses the paper below.

INÁCIO, M. L.; CABEZUDO, M. A. S.; RAMISCH, R.; DI FELIPPO, A.; PARDO, T. A. S. The AMR-PT corpus and the semantic annotation of challenging sentences from journalistic and opinion texts. In SciELO Preprints. https://doi.org/10.1590/1678-460x202255159. 2022. Available at <https://preprints.scielo.org/index.php/scielo/preprint/view/4652>.

Contributions:

- Refinement of AMR guidelines for annotating news texts and opinions in Brazilian Portuguese.
- Extension of the AMR corpus for Brazilian Portuguese and annotation of opinative sentences.
- Analysis of hard cases in both domains (news texts and opinative).

The AMR-PT corpus and the semantic annotation of challenging sentences from journalistic and opinion texts
O corpus AMR-PT e a anotação semântica de sentenças desafiadoras de textos jornalísticos e opinativos

Marcio Lima Inácio[1]
Marco Antonio Sobrevilla Cabezudo[2]
Renata Ramisch[3]
Ariani Di Felippo[4]
Thiago Alexandre Salgueiro Pardo[5]

Abstract

One of the most popular semantic representation languages in Natural Language Processing (NLP) is Abstract Meaning Representation (AMR). This formalism encodes the meaning of single sentences in directed rooted graphs. For English, there is a large annotated corpus that provides qualitative and reusable data for building or improving existing NLP methods and applications. For building AMR corpora for non-English languages, including Brazilian Portuguese, automatic and manual strategies have been conducted. The automatic annotation methods are essentially based on the cross-linguistic alignment of parallel corpora and the inheritance of the AMR annotation. The manual strategies focus on adapting the AMR English guidelines to a target language. Both annotation strategies have to deal with some phenomena that are challenging. This paper explores in detail some characteristics of Portuguese for which the AMR model had to be adapted and introduces two annotated corpora: AMRNews, a corpus of 870 annotated sentences from journalistic texts, and OpiSums-PT-AMR, comprising 404 opinionated sentences in AMR.

Keywords: Corpus Annotation; Knowledge Representation; Semantics

[1] Universidade de São Paulo. São Carlos – Brasil / Universidade de Coimbra. Coimbra – Portugal. https://orcid.org/0000-0002-0875-4574. E-mail: mlinacio@dei.uc.pt

[2] Universidade de São Paulo. São Carlos – Brasil. https://orcid.org/0000-0001-7625-9914. E-mail: msobrevillac@usp.br

[3] Redação Nota 1000. https://orcid.org/0000-0003-3372-6150. E-mail: renata.ramisch@redacaonota1000.com.br

[4] Universidade Federal de São Carlos. São Carlos – Brasil. https://orcid.org/0000-0002-4566-9352. E-mail: arianidf@gmail.com

[5] Universidade de São Paulo. São Carlos – Brasil. https://orcid.org/0000-0003-2111-1319. E-mail: taspardo@icmc.usp.br

Resumo

Abstract Meaning Representation (AMR) é uma linguagem de representação semântica bastante popular em Processamento de Línguas Naturais (PLN). Ela codifica o significado das sentenças em grafos orientados (enraizados). Para o inglês, há um grande *corpus* com anotação AMR que subsidia métodos e aplicações de PLN. Para a anotação de *corpora* em línguas que não sejam o inglês, incluindo o português brasileiro, tem-se aplicado estratégias automáticas ou manuais. As automáticas se baseiam essencialmente no alinhamento entre *corpora* paralelos e na herança da anotação AMR, enquanto as estratégias manuais focalizam na adaptação das diretrizes originais de anotação AMR (para o inglês) em função da língua-alvo. Ambas as estratégias, automática ou manual, precisam lidar com certos fenômenos linguísticos desafiadores. Neste trabalho, exploram-se características do português para as quais o modelo AMR foi adaptado e apresentam-se dois *corpora* anotados: AMRNews, *corpus* composto por 870 sentenças anotadas, provenientes de textos jornalísticos, e o *corpus* OpiSums-PT-AMR, contendo 404 sentenças opinativas em AMR.

Palavras-chave: Anotação de *corpus*; Representação de conhecimento; Semântica.

1. Introduction

Natural Language Processing (NLP) is a research field that aims at developing computational systems that are able to perform tasks involving interpretation and/or generation of natural languages such as automatic translation and summarization, sentiment analysis, text simplification, and speech recognition and synthesis, among several other tasks (Jurafsky & Martin, 2008).

NLP has significantly advanced in the last decade due to the good results obtained with artificial neural networks, in particular, deep learning (Goodfellow et al., 2016) and distributional word embedding models as *word2vec* (Mikolov et al., 2013) and BERT (Devlin et al., 2019). Despite such recent advances, Natural Language Understanding (NLU) or Natural Language Interpretation (NLI) has remained as a trending challenging topic in the NLP community. Defined as the subtopic of NLP that deals with machine reading comprehension, NLU is considered an AI-hard or AI-complete problem (Yampolskiy, 2013).

Given the considerable commercial interest in NLU because of its application in large-scale content analysis, recent works focused on different semantic representation languages have emerged. Some examples are the semantic representation used in the *Groningen Meaning Bank* (Basile et al., 2012), *Universal Conceptual Cognitive Annotation* (UCCA) (Abend & Rappoport, 2013), *Universal Decompositional Semantics* (White et al., 2016), and the model used in the *Parallel Meaning Bank* (Abzianidze et al., 2017).

In the NLU scenario, *Abstract Meaning Representation* (AMR) is a very popular and prominent semantic model, which arose to answer the need to build a semantic bank that includes different semantic phenomena. It aims at encoding the meaning of a sentence with a (relatively) simple representation in the form of a directed rooted graph (Banarescu et al., 2013). This representation includes semantic roles, named entities, spatial-temporal information and polarity, among other semantic information levels.

The AMR-annotated corpus for English is large, with approximately 39,000 sentences. The Chinese AMR corpus is also of respectable size, containing 10,149 sentences[1]. Differently from such situations, there are small annotated corpora for other languages, likely due to the high

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

complexity that building this kind of corpora represents. It is unnecessary to highlight the relevance of building corpora for other languages. Annotated corpora provide qualitative and reusable data for building or improving existing methods and applications and serving as benchmarks to compare different approaches.

Some efforts have been conducted to build AMR corpora for non-English languages. Some tried to use AMR as an interlingua and automatically mapped the alignments between parallel corpora (Anchiêta & Pardo, 2018a; Damonte & Cohen, 2018; Xue et al., 2014). In general, these works exploit an AMR parser for English and parallel corpora to learn AMR parsers for other languages (such as Italian, Spanish, German, and Chinese). Other works tried to adapt the AMR guidelines to annotate corpora in other languages[2] (Sobrevilla Cabezudo & Pardo, 2019; Migueles-Abraira et al., 2018), leveraging its cross-linguistic potential.

It is a known fact that automatic alignments can accelerate the annotation process but can also result in some limitations dealing with syntactic phenomena that account for several cross-lingual differences. For example, Damonte and Cohen (2018) mention that the automatic alignments generate AMR corpora with several mistakes, mostly involving concept identification. In another work, Anchiêta and Pardo (2018a) show that hidden subject and complex predicates are some linguistic phenomena not taken into account in the creation of an AMR corpus for Brazilian Portuguese (BP) via automatic alignment.

Manual AMR annotation can be an interesting direction for corpora building despite increasing annotation time. Some works focused on this approach are proposed by Sobrevilla Cabezudo & Pardo (2019) and Migueles-Abraira et al. (2018). However, performing manually AMR annotation for other languages, such as BP, is not a trivial task since the semantic representation model proposed in AMR is biased towards English, as stated by its original developers (Banarescu et al., 2013).

To build the first version of the AMRNews corpus in BP, Sobrevilla Cabezudo & Pardo (2019) manually annotated 299 sentences belonging to several news domains[3], adapting some of the current AMR guidelines. In order to increase the size of the corpus, we recently annotated 571

[2] Available at https://github.com/amrisi/amr-guidelines/blob/master/amr.md and detailed at https://amr.isi.edu/doc/amr-dict.html.

[3] Differently from the previous work focused on the "The Little Prince" book (Anchiêta & Pardo, 2018a).

more sentences using the same strategy, resulting in the version 2.0 of the corpus with 870 annotated sentences.

We also focused on the annotation of opinions, creating the OpiSums-PT-AMR corpus. Concerning a different domain from AMRNews, it enables a more semantic-focused comparative analysis between texts from both domains. Furthermore, this initiative provides data to be used in future research within the Sentiment Analysis area, as semantic knowledge can be an important feature to be taken into account in this type of processing, as argued by Cambria et al. (2015). To this extent, we used as basis the OpiSums-PT corpus (López Condori et al., 2015), comprising 1,502 sentences from comments about 17 different products, among which 404 have been annotated in AMR within the scope of this paper.

In this work, we explore the BP challenging linguistic phenomena for which the AMR model had to be adapted and present and detail the two annotated corpora: the AMRNews and the OpiSums-PT-AMR corpora. We also take advantage of the different domains of each AMR corpus and present a comparative analysis between opinions and news, highlighting important differences on the occurrence of semantic phenomena between each type of text.

This paper is organized as follows. Section 2 introduces the AMR fundamentals. In Section 3, we describe some works related to AMR corpus building. Afterwards, in Section 4, we present both the AMRNews and Opisums-PT-AMR corpora and report their annotation methodology and evaluation. We also perform a statistical description of each corpus together with some comparative analysis between this data. In Section 5, we explore some phenomena of our corpora in Portuguese and the correspondent adaptations of the English guidelines. Finally, some final remarks are presented in Section 6.

2. Abstract Meaning Representation

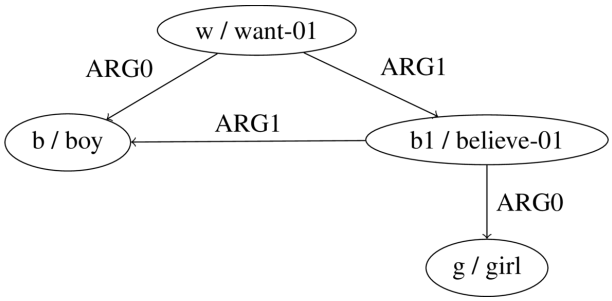
AMR is a semantic representation language designed to represent or encode the logical meaning of a sentence, abstracting away from elements of the surface syntactic structure, such as morphosyntactic information and word ordering (Banarescu et al., 2013). In a propositional-style logic, AMR is able to capture who is doing what to whom in a sentence. In such formalism, words that do not significantly contribute to the meaning of a sentence are left out of the annotation.

The AMR annotation is more frequently represented as a single-rooted directed acyclic graph with labeled nodes (concepts) and edges (relations) among them (see Figure 1). Nodes represent the main events and entities that occur in a sentence, and edges represent semantic relationships among nodes. AMR concepts are either (i) words in their lexicalized forms (e.g., boy), (ii) predicate-argument structure as defined by the PropBank resource (Palmer et al., 2005) (e.g., want-01), or (iii) special keywords such as "date-entity", "government-organization", and others. In the example of Figure 1, the concepts are `want-01`, `believe-01`, `boy` and `girl`, and the relations are `:ARG0` and `:ARG1`, represented by labeled directed edges in the graph. The symbols `w`, `b`, `b1` and `g` are variables and may be re-used in the annotation, corresponding to reentrances (multiple incoming edges) in the graph.

Overall, AMR has become popular in the NLP research community due to its attempt to abstract away from syntactic idiosyncrasies and its wide use of other comprehensive linguistic resources, such as PropBank[4] (Palmer et al., 2005), supposedly being relatively simpler than other semantic languages.

Concerning its attempt to abstract away from syntactic idiosyncrasies, it may be seen that the AMR annotation examples in Figure 1 could be generated from the sentences "The boy wants the girl to believe him" and "The boy wants to be believed by the girl", which are semantically similar, but with different syntactic realizations.

Figure 1 — AMR graph for the sentence "The boy wants the girl to believe him".



[4] The PropBank project created a corpus of text annotated with information about basic semantic propositions. More information at https://propbank.github.io/.

In relation to the use of linguistic resources, Figure 2 shows a predicate-argument structure (or *frameset*) provided by PropBank, which is essentially a verb linked to a list of possible arguments and their semantic roles. In this case, the frameset `want.01` represents the "desire to possess or do (something)" sense. It has two arguments, `Arg0` and `Arg1`, with the semantic roles `wanter` and `thing wanted`.

Figure 2 — Example of PropBank frameset.

```
Frameset want.01 "possession
desiring"
Arg0: wanter
Arg1: thing wanted
Ex: [Arg0 I] want [Arg1 a flight from
Ontario to Chicago].
```

Furthermore, AMR also offers approximately 100 additional relations, which are used to annotate different types of information, such as quantities (for example, `:quant`, `:unit`, `:scale`), dates (`:day`, `:month`, `:year`, `:weekday`), and others (`:mod`, `:manner`, `:location`, `:name`, `:polarity`). AMR may also be represented in first-order logic (see (a) in Figure 3) or in the PENMAN notation (Matthiessen & Bateman, 1991) (see (b) in Figure 3), for easier human reading and writing.

Figure 3 — Other AMR notations.

```
∃ w, b, b1, g:
instance(w, want-01) ∧
instance(b, boy) ∧
instance(b1, believe-01) ∧
instance(g, girl) ∧
ARG0(w, b) ∧
ARG1(w, b1) ∧
ARG0(b1, g) ∧
ARG1(b1, b)
```

(a) Logic

```
(w / want-01
    :ARG0 (b / boy)
    :ARG1 (b1 / believe-01
        :ARG0 (g / girl)
        :ARG1 b))
```

(b) PENMAN

3. Related Work

Although AMR was not initially planned to be an interlingual semantic representation (Banarescu et al., 2013), some efforts in this line have been made to build non-English corpora. Nowadays, there are aligned and parallel AMR corpora available in Czech, Chinese, Spanish, and BP (Anchiêta & Pardo, 2018a; Damonte & Cohen, 2018; Xue et al., 2014), built mainly in a semiautomatic way.

Xue et al. (2014) is probably the first work that addressed the construction of AMR annotated corpora for non-English languages. Aiming to evaluate AMR's potential to work as an interlingua, the authors annotated 100 English sentences from the Penn TreeBank (Marcus et al., 1993) with AMR. Such sentences were translated to Czech and Chinese, and annotated with AMR as well. As a result, Xue et al. (2014) observed that the level of compatibility of AMR between English and Chinese is higher than between English and Czech.

Since annotating AMR manually is time consuming and demands a team of experts to perform reliable annotation, some efforts were made to develop AMR parsing and converting tools from other semantic representations. Vanderwende et al. (2015) proposed an AMR parser to convert logic form representations into AMR for English. As a result, AMR-annotated corpora for French, German, Spanish, and Japanese have been released. Damonte and Cohen (2018) also developed an AMR parser for English. In their work, they used parallel corpora to learn AMR parsers for Italian, Spanish, German, and Chinese, and discovered that the tools were able to overcome structural differences between the languages. Another result of this work is the method proposed by the authors to evaluate the parsers, which exempt the need of gold standard data for the target languages.

Despite their usefulness, automatic alignment and conversion strategies do not necessarily reflect the complexity of some linguistic phenomena in non-English languages, so manual annotation or revision is necessary. Thus, other annotating teams tried to adapt the AMR guidelines to their languages (Sobrevilla Cabezudo & Pardo, 2019; Linh & Nguyen, 2019; Migueles-Abraira et al., 2018).

Migueles-Abraira et al. (2018) performed a manual AMR annotation of the book "The Little Prince", refining the original AMR guidelines and comparing Spanish and English in terms of similarity of the occurring phenomena. As a result of this work, the authors identified some relevant specific phenomena that proved to be challenging during the annotation process, such as ellipsis, third person possessives and clitic pronouns.

Concerning Portuguese, two AMR-annotated corpora were created. Both corpora used Verbo-Brasil[5] (Duran & Aluísio, 2015) as a lexical resource to annotate the framesets, that is based on the same representation scheme of the PropBank lexical repository. The first one was automatically built, leveraging the alignments between sentences of the "The Little Prince" book in English and Portuguese (Anchiêta & Pardo, 2018a). Specifically, such corpus is the result of an aligner based on pre-trained word embeddings and Word Mover's Distance function (Kusner et al., 2015) to match word tokens in the sentences and nodes in the corresponding AMR graphs. The Little Prince corpus has a rather unusual genre (tales), and is composed of sentences with restricted vocabulary, mainly related to the story. Furthermore, the number of sentences is small: only 1,527 annotated sentences. The other corpus is the AMRNews corpus, whose first version is described by Sobrevilla Cabezudo & Pardo (2019). The next section presents the current version of the corpus and reports its annotation methodology and evaluation, and also introduces a novel initiative on annotating opinions into the AMR formalism. Both corpora compound the AMR-PT corpus initiative.

4. The AMR-PT Initiative

General Description

The AMR-PT initiative comprises two corpora in different domains. One focuses on news texts, named AMRNews-PT, and the other one on opinionated texts, named OpiSums-PT-AMR. AMRNews is a news corpus with manually annotated sentences following the English AMR guidelines with some language-specific adaptations (Sobrevilla Cabezudo & Pardo, 2019). The journalistic texts were extracted from the Folha de São Paulo news agency[6]. The selected data came from different domains such as "Daily news", "World news", "Education", "Environment", "Sports", "Science", "Balance and Health", "Ilustrada",

[5] Available at http://143.107.183.175:21380/verbobrasil/.
[6] Available at https://www.folha.uol.com.br/.

“Ilustríssima”, “Power”, “Tourism”, “Food”, and “Technology”. To enrich the corpus, news sentences were also extracted from PropBank.Br[7] (Duran & Aluísio, 2012), since it already contains semantic role information, which makes the AMR annotation much easier. The document's download period was November 25th-28th, 2018. Currently, this corpus contains 870 manually annotated sentences and the size for each sentence is up to 23 tokens[8].

In its turn, the OpiSums-PT-AMR corpus comprises 404 manually annotated sentences from the OpiSums-PT (López Condori et al., 2015) corpus, which was created based on comments about 13 books — originally from the ReLi corpus (Freitas et al., 2014) — alongside opinions concerning four electronic products obtained from the Buscapé[9] e-commerce website. Each product has 10 comments with multiple sentences each. Every document also does not exceed 300 tokens.

Annotation Procedure

Both annotations of journalistic and opinionated sentences followed the same process (Sobrevilla Cabezudo & Pardo, 2019). This means that it was guided by the original AMR guidelines[10] including the adaptations performed by the authors, and the lexical repository used was Verbo-Brasil (Duran & Aluísio, 2015).

The initial annotation focused on journalistic sentences (Sobrevilla Cabezudo & Pardo, 2019) and the team was originally composed of 14 annotators that belong to the areas of Computer Science and Linguistics, and with large experience in NLP. These annotators took part in two training sessions. In the first session, the task and the resources to be used were presented. The participants were trained by annotating sentences of PropBank.Br (Duran & Aluísio, 2012) for perceiving the difficulty of the task. The second session aimed at answering questions about the annotation, showing the inter-annotator agreement in the training stage, some common mistakes, and launching the annotation process. This process resulted in 299 annotated sentences.

---

[7] PropBank.Br was the basis for the construction of the previously cited Verbo-Brasil repository.

[8] Due to the difficulty that the task of semantic annotation carries, the scope of this corpus was focused on annotating only short sentences (but guaranteeing that different domains are covered). In order to define what a short sentence is, the average number of tokens by sentence was calculated and this value was used as a threshold. Thus, sentences with a number of tokens below the average (in our case, it was 23 tokens) were selected.

[9] Available at https://www.buscape.com.br.

[10] Available at https://github.com/amrisi/amr-guidelines/blob/master/amr.md. The adopted version was the 1.2.6.

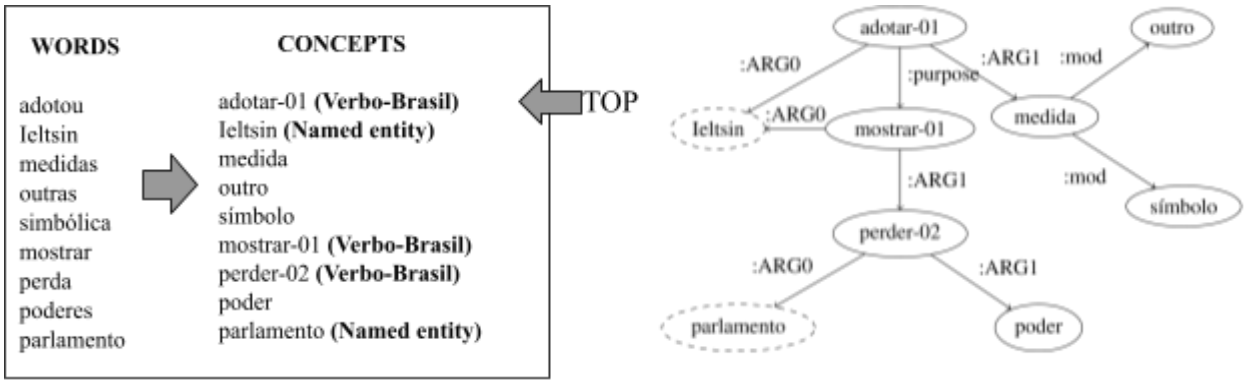
In general, the annotation procedure consisted of two general steps. The first step aimed at analyzing the sentence structure, while the second step counted with the aid of the AMR Editor tool (Hermjakob, 2013), which produces an AMR PENMAN format to be exported into textual files of easy processing and consulting. Furthermore, the fundamentals of the manual annotation process in the first step were the following:

a) Identification of sentence type (i.e., default, comparative, superlative, coordinate, subordinate, and others), which determines whether it is necessary to build two or more sub-graphs (in case of coordinate or subordinate sentences) and then to join them using a conjunction (usually coordinate sentences) or a concept of the main sub-graph (in the case of subordinate sentences).
b) Concept identification, which was based on the AMR guidelines. Specifically, the annotators identify general concepts, either from the AMR guidelines or from Verbo-Brasil.
c) Identification of the main concept, which is done based on the two previous steps. To illustrate, the main verb could be the main concept in a default sentence.
d) Identification of relations among the identified concepts[11].

This sequence of actions (a-d) can be illustrated with Figure 4. To annotate the sentence "*Ieltsin adotou outras medidas simbólicas para mostrar a perda de poderes do Parlamento*" ("Yeltsin took other symbolic measures to show the loss of Parliament's power."), the annotators firstly identify that it includes a subordinate clause, which means that its correspondent AMR graph should have sub-graphs. Then, they identify the concepts. In Figure 4 (a), we see that some words became general concepts (i.e., `medida`, `outro`, `símbolo` and `poder`), named-entities (`Ieltsin` and `Parlamento`) or Verbo-Brasil framesets (`adotar-01`, `mostrar-01` and `perder-02`).

[11] The relations were extracted from Verbo-Brasil (for core relations) and AMR guidelines (for non-core relations).

Figure 4 — Example of the manual annotation procedure (Sobrevilla Cabezudo & Pardo, 2019).



(a) Concept identification and Top concept identification

(b) Relation identification

Following the concepts identification, it was necessary to identify the graph root: in this case, the verb `adotar-01`, because it is the main verb of the main clause "*Ieltsin adotou outras medidas simbólicas*" ("Yeltsin took other symbolic measures"). Finally, the relations among all concepts were identified (e.g., `:ARG0` between `Ieltsin` and `adotar-01`), which encodes that `Ieltsin` has the `adopter` semantic role according to the corresponding frameset.

After the initial annotation and with some learned lessons, we focused on annotating both journalistic and opinionated sentences. This process was performed by three human experts with previous background on AMR and its guidelines and consists in (1) annotating a set of sentences and (2) discussing the hard cases and other interesting aspects of the annotation in an iterative way. So far, the annotation process resulted in 404 opinionated and 870 journalistic sentences (resulting from the previously annotated news corpus of Sobrevilla Cabezudo & Pardo (2019) and the newly 571 annotated sentences).

It is worth noting that, before the annotation of opinionated sentences, all texts were normalized using the Enelvo[12] tool (Bertaglia & Nunes, 2016), given that guidelines state that misspellings should be normalized during annotation. If the annotators think that there is a normalization error, they could check the original text and mention this in a note.

[12]Available at https://github.com/tfcbertaglia/enelvo.

Finally, and in a similar way to Sobrevilla Cabezudo & Pardo (2019), each sentence with a verb that was not present in the Verbo-Brasil repository was not annotated and the given verb was added in a list to enable further development of the resource (in future work).

Evaluation

To compute the inter-annotator agreement, a random subset of all sentences to be annotated was shared among all the annotators and they could not discuss this subset. The agreement measure used was Smatch[13] (Cai & Knight, 2013). Unlike the work of Banarescu et al. (2013), which built a gold standard (using the total agreement between the annotators), we calculated the inter-annotator agreement by comparing all annotations in an all-against-all configuration, obtaining the average of all inter-annotator agreements[14]. A set of 50 sentences was used for calculating the agreement within the journalistic corpus. Meanwhile, for the OpiSums-PT-AMR, 70 sentences of different lengths were initially selected to compose the agreement set, however, due to the complexity of the annotation process, only 17 were actually annotated by all three experts and, therefore, were considered to calculate the agreement for this specific corpus.

The overall agreement for the journalistic part of the corpora achieved a Smatch value of 0.73, which is a good value, considering that the inter-annotator agreement in the original AMR project ranged between 0.70 and 0.80. Besides, 34 (from the 50) sentences were annotated by 5-7 annotators in the initial procedure and the last 16 sentences were annotated by 3 annotators.

The annotation of sentences from the opinions domain resulted in an average agreement of 0.90, which can be considered high, when compared with other works on the matter. This can be due to the fact that the 17 sentences used are shorter and, therefore, easier to achieve some consensus.

---

[13]Available at https://github.com/snowblink14/smatch. It is interesting to notice that, differently from annotation efforts for other linguistic phenomena, the Smatch metric is the dominant metric for AMR annotation (instead of Kappa or other metrics (Banerjee et al., 1999)), following the original work on AMR (Banarescu et al., 2013). It evaluates the triples formed by the relations and the associated nodes in an AMR structure. Moreover, Smatch does an additional task of mapping the variables in the AMR representation in a way to maximize the results.

[14] Finally, the annotated versions of the sentences belonging to the agreement sample that were included in the final corpus were chosen by an adjudicator (as more than one possible annotation exists).

From the obtained annotation, we can make a comparative analysis between the two domains within this AMR-BP initiative, pointing out how the texts differ in terms of semantic phenomena and how they are captured by the AMR representation.

News vs Opinionated texts

In total, the AMRNews corpus includes 4,192 concepts (excluding `name`) and 3,758 relations (excluding `:instance`), whilst OpiSums-PT-AMR comprises 3,064 concepts and 3,159 relations. As a first comparative analysis, we can observe the distribution of the different types of phenomena captured by the concepts within the AMR graphs. These statistics can be seen in Table 1.

Table 1 — Statistics of concepts in both the AMRNews and OpiSums-PT-AMR corpora.

Concepts	**Frequency**	
	AMRNews	**OpiSums-PT-AMR**
General concepts	1,977	1,770
Verbo-Brasil concepts	866	641
Named entities	311	125
Modal verbs	45	25
Amr-unknown[15]	80	7
Other entities and special frames	104	169
Constants[16]	660	215
Negative polarity	135	79

In a more detailed analysis, we present in Table 2 the 15 most frequent relations in each corpus. It is possible to see that the 3 most frequent relations are the same (and in the same order) in the two corpora. One point to remark in relation to the table is that, in the news texts, the sentences and expressions contained in them describe facts and usually use numbers to report quantities (through the `:quant` relation). More than this, some expressions collected until now describe

[15] AMR uses the concept "amr-unknown" to indicate wh-questions.
[16] Constants include numbers, strings and symbols that are not traditional concepts and, therefore, are not given variable names.

imperatives like "*Arranje*!" ("Get it!"). Thus, the imperative mode (`:mode` relation) is frequent in the corpus. It is expected that, when the news corpus grows, these relations will change a bit.

Table 2 — Fifteen most frequent relations in both the AMRNews and OpiSums-PT-AMR corpora.

OpiSums-PT-AMR			**AMRNews**		
Relation	**Frequency**	**Freq. (%)**	**Relation**	**Frequency**	**Freq. (%)**
ARG1	652	20.64%	ARG1	715	19.03%
op	624	19.75%	op	706	18.79%
ARG0	485	15.35%	ARG0	512	13.62%
mod	314	9.94%	name	311	8.28%
ARG2	208	6.58%	mod	268	7.13%
name	125	3.96%	ARG2	196	5.22%
domain	96	3.04%	polarity	169	4.50%
polarity	80	2.53%	domain	143	3.81%
time	67	2.12%	quant	105	2.79%
topic	56	1.77%	time	98	2.61%
poss	56	1.77%	location	75	2.00%
snt	55	1.74%	manner	49	1.30%
quant	44	1.39%	poss	48	1.28%
degree	39	1.23%	topic	45	1.20%
ARG3	38	1.20%	mode	36	0.96%

We can also note, from both Table 1 and Table 2 (through the `:name` relation), that news texts contain a higher proportion of named entities. However this phenomenon is still common in opinions, as `:name` is the 6th most common relation in OpiSums-PT-AMR. It is also worth pointing out that the `:degree` relation, used mainly with amplifiers and downtoners, are more common in opinionated texts, especially when taking into account its associated concept

(have-degree-91), as can be seen in Table 3, which includes the ten most frequent framesets for both corpora.

Table 3 — Ten most frequent framesets in both the AMRNews and OpiSums-PT-AMR corpora.

OpiSums-PT-AMR			AMRNews		
Frameset	**Frequency**	**Freq. (%)**	**Frameset**	**Frequency**	**Freq. (%)**
cause-01	44	5.27	ter-01	42	4.14
ler-01	42	5.03	contrast-01	29	2.86
ter-01	35	4.19	possible-01	27	2.66
gostar-01	33	3.95	dizer-01	24	2.36
contrast-01	27	3.23	fazer-01	23	2.27
escrever-01	25	2.99	haver-01	17	1.67
have-rel-role-91	21	2.51	querer-01	17	1.67
have-degree-91	21	2.51	acontecer-01	15	1.48
possible-01	17	2.04	saber-01	13	1.28
fazer-01	15	1.80	cause-01	13	1.28

Analyzing the results in Table 3[17], it is also important to mention that the higher frequencies of some concepts — such as ler-01 (to read), escrever-01 (to write) and have-rel-role-91 (used to indicate personal relationship between people) — are due to the type of products about which the opinions are written, mainly books.

A noteworthy observation to be made is that opinions have framesets used within contexts with some degree of sentiment associated, e.g., gostar-01 (to like) and have-degree-91. Meanwhile, news texts have more descriptive concepts, such as ter-01 (to have), dizer-01 (to say), fazer-01 (to do/to make) and acontecer-01 (to happen), among others.

[17] Some framesets in the table come directly from the English PropBank and not from Verbo-Brasil due to the original guidelines developed by Sobrevilla Cabezudo & Pardo (2019), in which modal verbs (possible-01) and some conjunctions (contrast-01, cause-01) are annotated in such way to keep consistency with the original AMR guidelines (Banarescu et al., 2013). Some other framesets are AMR-exclusive, for instance, have-rel-role-01 and have-degree-91, and were kept in English.

One of the limitations of annotating AMR in BP is related to Verbo-Brasil, since the lexical units that are not represented in this resource could not be annotated. We found 161 verbs that were not present in Verbo-Brasil, as *opinar*, *duvidar*, *ficar* (in the sense of "dating") and *devorar* (in the sense of "outperforming"). Overall, almost 14% of the analyzed sentences had verbs not found in Verbo-Brasil. These sentences were discarded and, therefore, not included in our corpora. Thus, this work also shows directions for developing and improving the linguistic resources used for building the annotated corpora, as well as adapting original methods and guidelines, which remain for future work.

5. Linguistic Phenomena and Adapted Guidelines

The experience of annotating news and opinionated corpora in BP with AMR allowed the identification of some challenging phenomena that could not (totally or partially) be represented with AMR. Thus, we conducted a linguistic analysis of them that could offer possible solutions to other language annotation teams that face similar issues. Although we are not able to propose definitive solutions to all the problems, we believe that they are possible satisfactory strategies. The hard cases discussed here are diminutives, null subject, pronoun ambiguity, and multiword expressions. We do not aim, however, to present an extensive or exhaustive analysis for each example and issue in the corpus.

Diminutives

From the 404 sentences of OpiSums-PT-AMR, there are five sentences with one diminutive case each (1-5). Such diminutives are basically formed by replacing the unstressed final vowel -o or -a of a word with the affix *-inho* or *-inha* according to its gender. There are other rules of diminutive formation[18] in BP, but there are no occurrences of them in the corpus.

1. Aquele filme meio [*chatinho*]$_{adj}$ e clichê que está passando na televisão. ("That rather boring and cliché film that is on television")

[18] Diminutive in BP can also be formed as follows: (i) with nouns and adjectives ending in *-s* or *-z*, the affix *-inho/-inha* is also added to the stem word (e.g., *japonês* ("Japanese man") > *japonesinho* ("little Japanese guy"), and *voz* ("voice") (fem.) > *vozinha* ("little voice"), and (ii) with all other nouns, the affix -zinho/-zinha is added to the word (e.g., *papel* ("paper") (masc.) > *papelzinho* ("scrap of paper"), and *mão* ("hand") (fem.) > *mãozinha* ("little hand").

2. Livro bem [*chatinho*]$_{adj}$ (“[A] pretty boring book”)

3. Muito [*engraçadinho*]$_{adj}$! ;) (“Very funny! ;)”)

4. Lindo, [*fininho*]$_{adj}$ e discreto. (“[It´s] Beautiful, very thin and discrete”)

5. Acaba se atrapalhando com a sua “[*anjinha*]$_{noun}$”. (“He/She ends up messing with his/her little angel’”)

In the examples (1) and (2), the diminutive form *chatinho* is used to temper an unpleasant quality. In (3) and (4), however, the meaning is quite different from (1) and (2); they have the meaning of “nice and…” or having a quality to exactly the desirable degree (i.e., *engraçadinho* > “good and funny”, and *fininho* > “good and thin”). As illustrated by sentence (5), diminutive forms very often connote cuteness, affection or pleasantness (more examples are: “*Que tal uma cervejinha gelada?*” / “What about a nice and cold beer?” or “*Adoro pezinho de bebê*” (“I love babies’ little feet”)[19].

According to Alves (2006), diminutive forms can be classified in terms of their function in semantic diminutives and pragmatic diminutives. The first group expresses “reduced size /quantity /intensity” meanings, which are based on inherent properties or features of the objects. The second one expresses more subjective meanings, and refers to how the speaker perceives objects and their properties, which are guided by social and cultural factors. Thus, we first classified the cases in semantic diminutives (4) and pragmatic diminutives (1, 2, 3 and 5) for understanding the different meanings of such words before the AMR annotation[20]. This task was strongly based on world knowledge, since the sentences are out of context, as it is established by the AMR guidelines.

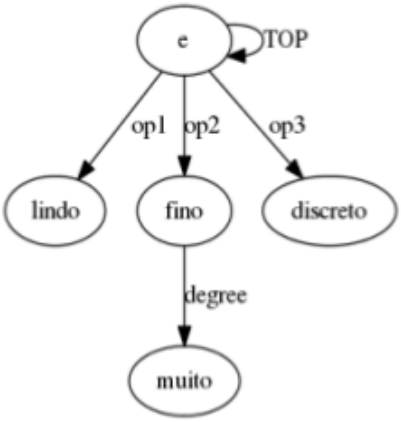
We then turn to the AMR guidelines for diminutive annotation. While the semantic diminutive *fininho* in (4), for example, is easily represented in AMR with the `:degree` relation (as in Figure 5), which links two concepts, i.e., `fino` (“small”)) and `muito` (“very”)), the pragmatic diminutive is much more difficult to represent, since it corresponds to non-literal meanings. In other words, the concepts represented by pragmatic diminutives do not literally mean a `:degree` relation, so using the same annotation in both constructions seems inappropriate. Consequently, we used two different annotation schemes for diminutives: while the semantic

[19] It’s worth noting that the same can happen with the augmentative. In our corpora, however, there was no occurrence of augmentative forms.

[20] The different meanings of diminutive are not an idiosyncrasy of Portuguese, however, we are not aware of specific AMR guidelines in the literature to annotate these cases.

ones are represented as usual with the `:degree` relation (Figure 5), the pragmatic diminutives are not lemmatized, and the concept remains as a diminutive, as in Figure 6.

Figure 5 — Annotation of sentence 4 with a case of semantic diminutive.

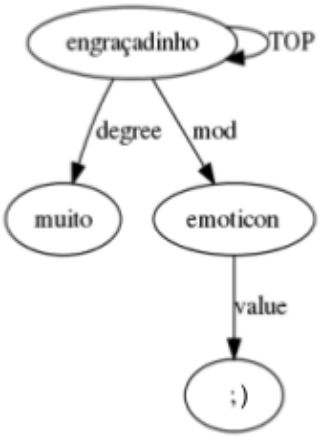


(a) AMR Graph

```
(e0 / e
    :op1 (l0 / lindo)
    :op2 (f0 / fino
      :degree (m0 / muito))
    :op3 (d0 / discreto))
```

(b) PENMAN notation

Figure 6 — Annotation of sentence 3 with a case of pragmatic diminutive.



(a) AMR Graph

```
(e0 / engraçadinho
    :degree (m0 / muito)
    :mod (e1 / emoticon
           :value ";)"))
```

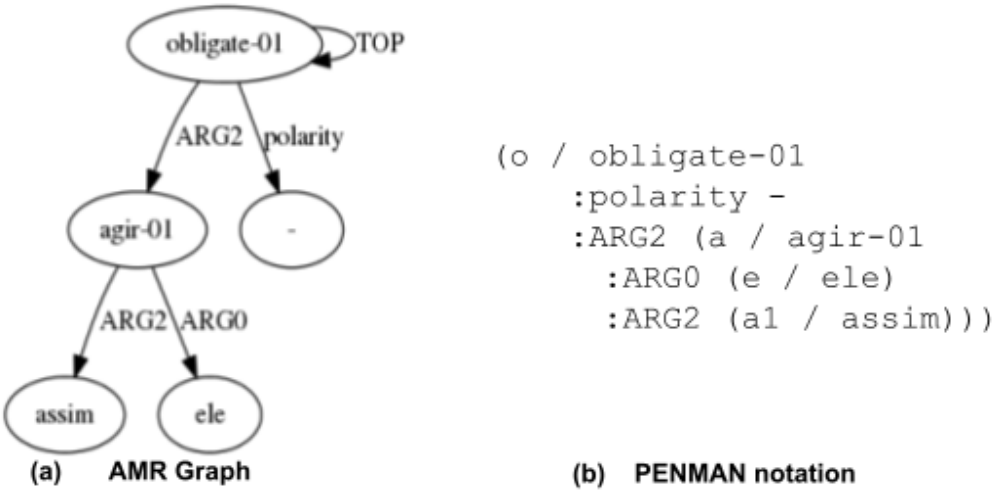
(b) PENMAN notation

Null subject and pronoun ambiguity

In BP, as in other romance languages (e.g., Spanish), but different from English, the subject does not have to be necessarily expressed in the sentence. In the example shown in Figure 7, the subject is not present in the sentence ("*Não precisaria agir assim.*") ("[He/She/You] wouldn't have to act like this"), but is probably clear in the sentence source-text. However, the verb ("*precisaria*") indicates that the person referred to is a third person in singular, since the

verb has this conjugation. In this situation, the annotation team decided to annotate the ARG0 role explicitly, even if it is not in the sentence. The reason for such a definition is that it permits the explicit identification of the ARG0 role. Thus, it would be possible to recover the agreement information. However, this decision led to another problem: the third-person pronoun ambiguity. In BP, a verb conjugated in the third person can refer to the second person in singular ("you") or to the third person in singular ("he" or "she"), so it had also to be decided if the pronoun annotated would be the second or the third person pronoun, and, in the later case, if it is masculine or feminine. Thus, the decision was to annotate it as the third person masculine (he). The same decision was kept for the ambiguity of possessive pronouns, when *seu*/*sua* can refer both to yours and his/her.

Figure 7 — Null subject annotation.



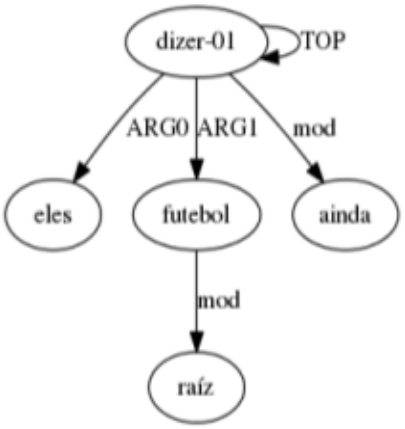
This decision was based on the argument that, as the initial annotation focused on a journalistic corpus, it was expected to be more frequent that the null subject refers to a person about whom something is being reported. Besides, the decision for the masculine is based on the original lemmatization rules for concepts in Portuguese, that orient to lemmatize the modifiers in their masculine singular form.

Another problem arises when the verb is in plural form and the subject is indeterminable, as in "*Dirão até que é futebol raiz*" ("[They/Someone] will also say it is the old soccer") (Figure 8). In this example, the annotators were oriented to explicitly annotate the ARG0 as "they", but mention the fact that the sentence contains an indeterminable subject. In this case, the standard

orientation of always using singular was changed in favor of the possibility to represent this phenomenon.

Lastly, there are some cases where it is not possible to identify if the pronoun is a personal one or a demonstrative one, as in "*Pode até criar problema emocional e retração social*" ("[She/He/It/You] can even create emotional problem and social withdrawal") (Figure 9). The subject could be a person, a fact or the entire previous sentence that is being taken up, and it is impossible to recover this information without context (note that AMR considers only the sentence level for the annotation). In this case, the annotators were oriented to not explicitly annotate any pronoun.

Figure 8 — Indeterminable subject explicitly annotated.

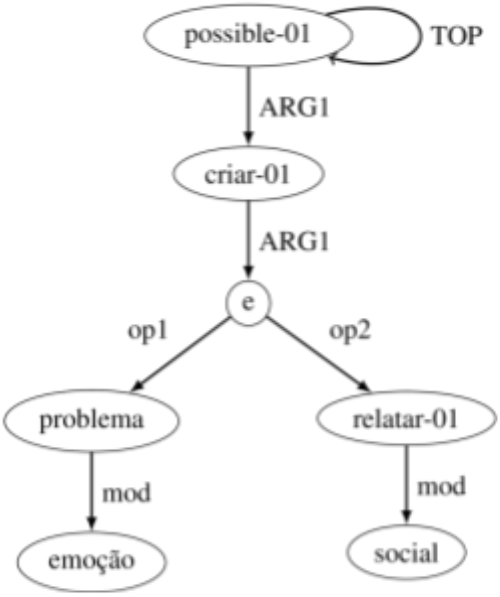


(a) **AMR Graph**

```
(d / dizer-01
    :ARG0 (e / eles)
    :ARG1 (f / futebol
      :mod (r / raíz))
    :mod (a / ainda))
```

(b) **PENMAN notation**

Figure 9 — Indeterminable subject not explicitly annotated.



(a) Grafo AMR

```
(p / possible-01
    :ARG1 (c / criar
        :ARG1 (e / e
            :op1 (p2 / problema
                :mod (e2 / emoção))
            :op2 (r / retratar-01
                :mod (s / social)))))
```

(b) PENMAN notation

Multiword expressions

A specially challenging phenomenon was the annotation of multiword expressions (MWE). MWEs are (continuous or discontinuous) sequences of words with some degree of orthographic, morphological, syntactic or semantic idiosyncrasy with respect to what is considered general grammar rules of a language (Baldwin & Kim, 2010). Another important property of MWEs is the semantic non-compositionality, i.e., it is impossible to deduce the meaning of the whole unit based only on the meaning of its parts (Constant et al., 2017).

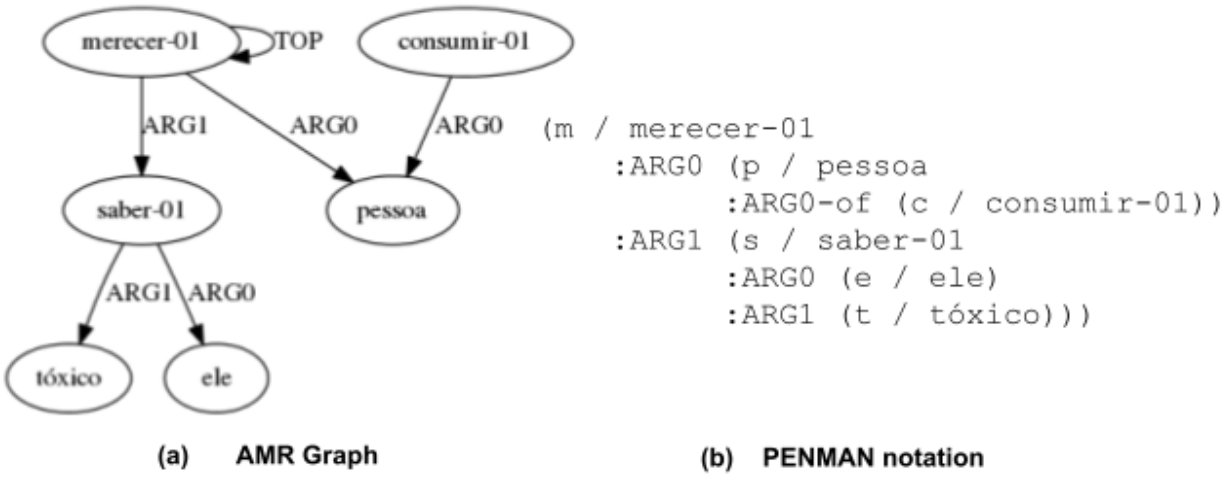
The original AMR guidelines define that MWEs should be represented as a unique concept that is synonym or equivalent to the MWE. One example that occurs frequently are the light-verb constructions (LVCs), such as “The girl made adjustments [adjust] to the machine”. LVCs are composed by a verb that does not add much semantics to the expression (the light verb) (Wittenberg et al., 2014), followed by a predicative noun that represents a state or an event. While many cases have indeed a unique verbal form that can replace the MWE (“make adjustments” has the equivalent full verb “adjust”), in some cases this is not possible.

In Figure 10, there is no full verb that could directly substitute the MWE “*ter direito*”, so in cases like that the team decided to find synonyms (in the example, the full verb “*merecer*”

means “to deserve”). In other examples, such as in the idiom “*pagar mico*” (literally, “to pay the monkey”), that means “to completely embarrass yourself”, the best synonym in BP is also a MWE: “*passar vergonha*” (“to get embarrassed”).

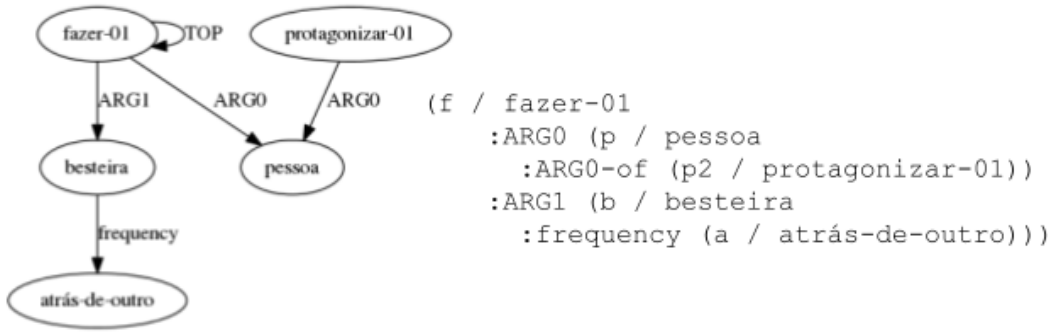
The solution in these cases was finding a concept in the Verbo-Brasil repository that could represent this structure with core arguments, resulting in annotating light-verbs as full verbs and ignoring the fact that the element predicating the sentence is actually the predicative noun. This case demanded a lot of discussion every time the team faced a new MWE, because some synonyms do not express the same meaning as the composed construction. This issue arises also because, different from PropBank, Verbo-Brasil has very few MWEs and no predicative nouns as framesets. Increasing the number and the diversity of the repository would probably solve most of these problems, but this would cost time and another team of experts to improve the lexical resource before continuing the AMR annotations.

Figure 10 — MWE annotated by synonym.



Furthermore, some MWEs are composed of non-lexical words and have adverbial meanings, as “*atrás de*” in the sentence “*A protagonista faz uma besteira atrás da outra*” (“The protagonist makes a mess after the other”) (cf. Figure 11).

Figure 11 — MWE annotated with hyphens.



(a) AMR Graph **(b) PENMAN notation**

The prepositional compound *atrás de* typically has the locative meaning of "behind", but in the MWE "*um* [noun] *atrás de outro*" (e.g., "*uma besteira atrás da outra*") it acquires the sense of a temporal sequence of things (as represented by "after", in English). In this case, the team annotated the expression as it occurs, using hyphens, as implied by some examples presented at the original AMR dictionary guidelines for the English language (e.g., "He can recite the poem by heart." is annotated with a concept `by-heart`, to indicate the manner in which the poem is recited). This leads to many concepts that have to be represented in this way, which is not a good long-term solution, since it makes room for annotators using hyphens whenever a compound arises.

The best way to deal with this phenomenon continues to be an open question, and it would be useful to analyze the frequency of MWEs in the corpus for proposing other (better) solutions for annotating them. One option would be to improve Verbo-Brasil and adding not only multiword framesets, but also the predicative nouns and their argument structure, so they could be used for the annotation as it has been made for English. This challenge highlights the importance of robust lexical resources that are not always available for under-represented languages. This could be one of the most important constraints in using AMR as an interlingua. Another possible option (that was not taken by now in the BP team) is using other lexical repositories that are specific for MWEs, such as the Parseme corpus for verbal MWEs, that is available for Portuguese, as for many other languages (Ramisch et al., 2018). This could be used at least as a consulting repository to identify when an expression is a real MWE (since this identification is not trivial), and for future improvement of Verbo-Brasil.

6. Final Remarks

This work presented and detailed two AMR annotated corpora for Brazilian Portuguese — the AMRNews and the OpiSums-PT-AMR corpora — and carried out a comparative analysis between opinions and news, highlighting important differences on the occurrence of semantic phenomena between each type of text. The released version of the AMR Corpus for Brazilian Portuguese is available at the web portal of the POeTiSA project[21]. Although the amount of AMR annotated data for Portuguese is still small (due to the hard task that AMR annotation represents), it has already subsidized NLP initiatives for the Portuguese language, as semantic parsing (Anchiêta & Pardo, 2018b, 2022), text generation (Sobrevilla Cabezudo & Pardo, 2022), and opinion summarization (Inácio & Pardo, 2021).

We also explored the language-specific challenges that appeared during the AMR annotation process and some strategies to deal with these. As it could be seen, some of them may be better handled (diminutives). However, there are other phenomena which are hard to deal with and a deeper study has to be conducted. On the other hand, projects aiming to build unified multilingual sembanks for NLP tasks have to follow a minimum pattern by annotating similar phenomena to allow comparing them in terms of frequency and structure among different languages. In this way, annotation adaptations should be restricted to specific phenomena (and as general as possible to capture similar phenomena in similar languages), so the core idea of the AMR scheme rests true for as many languages as possible.

There are also phenomena that may lead to further research of the AMR semantic representation, such as metaphorical language, which is situated in a boundary interface between semantics and pragmatics, according to Legroski (2009). This type of phenomenon also has a degree of relation to multiword expressions, which, as we discuss throughout this paper, present a challenge for annotation.

Gender is also an interesting path for research. Migueles-Abraira (2017) includes grammatical gender annotation for their Spanish version of AMR, under the argument that it has influence on the understanding of a sentence. For example, the word “caixa” may represent two different concepts in Portuguese: box or clerk, depending on its gender (feminine or masculine, respectively). This decision, however, needs to be further discussed taking into account the

[21] https://sites.google.com/icmc.usp.br/poetisa

lemmatization process of words into concepts (since, in Portuguese, the lemmas are commonly represented by the words' masculine forms) and which other morphological aspects (e.g., number) should be included in a semantic representation as the AMR.

Acknowledgments

The authors of this work would like to thank the Center for Artificial Intelligence (C4AI-USP) and the support from the São Paulo Research Foundation (FAPESP grant #2019/07665-4) and from the IBM Corporation. This research also had the support of Coordination for the Improvement of Higher Education Personnel (CAPES) and the OPINANDO project[22] (PRP #668).

Conflict of interests (multiple authors)

(X) The authors declare they have no conflict of interest.

Credit Author Statement

We, Marcio Lima Inácio, Marco Antonio Sobrevilla Cabezudo, Renata Ramisch, Ariani Di Felippo and Thiago Alexandre Salgueiro Pardo, hereby declare that we do not have any potential conflict of interest in this study. The first two authors have carried out the collection of the texts to be annotated and, alongside Renata Ramisch, performed the annotation of the copus in AMR. Marcio Inácio and Marco Cabezudo have also been responsible for the contrastive data analysis. All authors contributed to the qualitative discussion about the phenomena observed in the data and the writing of this work. Ariani Di Felippo and Thiago Pardo were responsible for the project supervision. All authors approve the final version of the manuscript and are responsible for all aspects, including the guarantee of its veracity and integrity.

[22] https://sites.google.com/icmc.usp.br/opinando/

Bibliographical References

Abend, O., & Rappoport, A. (2013). UCCA: A semantics-based grammatical annotation scheme. *Proceedings of the 10th International Conference on Computational Semantics – Long Papers*, 1–12.

Abzianidze, L., Bjerva, J., Evang, K., Haagsma, H., van Noord, R., Ludmann, P., Nguyen, D.-D., & Bos, J. (2017). The Parallel Meaning Bank: Towards a multilingual corpus of translations annotated with compositional meaning representations. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 242–247.

Alves, E. (2006). O diminutivo no português do Brasil: Funcionalidade e tipologia. *Estudos Linguísticos*, *35*, 694–701.

Anchiêta, R. T., & Pardo, T. A. S. (2018a). Towards AMR-BR: A SemBank for Brazilian Portuguese Language. *Proceedings of the eleventh international conference on language resources and evaluation*, 974–979.

Anchiêta, R. T., & Pardo, T. A. S. (2018b). A Rule-Based AMR Parser for Portuguese. *Proceedings of the 16th Ibero-American Conference on Artificial Intelligence*, 341–353. https://doi.org/10.1007/978-3-030-03928-8_28.

Anchiêta, R. T., & Pardo, T. A. S. (2022). Abstract Meaning Representation Parsing for the Brazilian Portuguese Language. *Proceedings of the International Conference on Computational Processing of Portuguese*, 429–434. https://doi.org/10.11606/T.55.2020.tde-29072020-120805.

Baldwin, T., & Kim, S. N. (2010). Multiword expressions. In N. Indurkhya & F. J. Damerau (Eds.), *Handbook of Natural Language Processing* (2nd ed.), 267–292. CRC Press.

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., & Schneider, N. (2013). Abstract meaning representation for

sembanking. *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, 178–186.

Banerjee, M., Capozzoli, M., McSweeney, L., & Sinha, D. (1999). Beyond kappa: A review of interrater agreement measures. *The Canadian Journal of Statistics*, *27*(1), 3–23.

Basile, V., Bos, J., Evang, K., & Venhuizen, N. (2012). Developing a large semantically annotated corpus. *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, 3196–3200.

Bertaglia, T. F. C., & Nunes, M. das G. V. (2016). Exploring word embeddings for unsupervised textual user-generated content normalization. *Proceedings of the 2nd Workshop on Noisy User-generated Text*, 112–120.

Cai, S., & Knight, K. (2013). Smatch: An evaluation metric for semantic feature structures. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, 748–752.

Cambria, E., Poria, S., Bisio, F., Bajpai, R., & Chaturvedi, I. (2015). The CLSA model: A novel framework for concept-level sentiment analysis. *Proceedings of the Computational linguistics and intelligent text processing conference*, 3–22. https://doi.org/10.1007/978-3-319-18117-2_1.

López Condori, R. E., Pardo, T. A. S., Avanço, L. V., Filho, P., Bokan, A., Cardoso, P., Dias, M., Nóbrega, F., Sobrevilla Cabezudo, M. A., Souza, J., Zacarias, A., Seno, E., & Di Felippo, A. (2015). A qualitative analysis of a corpus of opinion summaries based on aspects. *Proceedings of the 9th Linguistic Annotation Workshop*, 62–71. http://dx.doi.org/10.3115/v1/W15-1607.

Constant, M., Eryiğit, G., Monti, J., van der Plas, L., Ramisch, C., Rosner, M., & Todirascu, A. (2017). Multiword Expression Processing: A Survey. *Computational Linguistics*, *43*(4), 837–892. https://doi.org/10.1162/COLI_a_00302.

Damonte, M., & Cohen, S. B. (2018). Cross-lingual Abstract Meaning Representation parsing. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1146–1155. https://doi.org/10.18653/v1/N18-1104.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *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)*, 4171–4186.

Duran, M. S., & Aluísio, S. M. (2015). Automatic generation of a lexical resource to support semantic role labeling in portuguese. *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, 216–221.

Duran, M. S., & Aluísio, S. M. (2012). Propbank-Br: A Brazilian Treebank annotated with semantic role labels. *Proceedings of the Eighth International Conference on Language Resources and Evaluation* , 1862–1867.

Freitas, C., Motta, E., Milidiú, R. L., & César, J. (2014). Sparkling vampire... LOL! Annotating opinions in a book review corpus. *New Language Technologies and Linguistic Research*, 128–146.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning* (1st ed.). MIT Press.

Hermjakob, U. (2013). *AMR editor: A tool to build abstract meaning representations*. Available at https://amr.isi.edu/editor.html

Inácio, M. L., & Pardo, T. A. S. (2021). Semantic-Based Opinion Summarization. *Proceedings of Recent Advances in Natural Language Processing*, 624–633. https://doi.org/10.11606/D.55.2021.tde-13092021-141741.

Jurafsky, D., & Martin, J. H. (2008). *Speech and language processing. An introduction to Natural Language Processing, Computational Linguistics and Speech Recognition* (2nd ed.). Prentice Hall.

Kusner, M. J., Sun, Y., Kolkin, N. I., & Weinberger, K. Q. (2015). From word embeddings to document distances. *Proceedings of the 32nd International Conference on International Conference on Machine Learning*, 957–966.

Legroski, M. (2009). Definindo Metáfora. *Revista Polidisciplinar Eletrônica da Faculdade Guairacá*, *1*(2), 15–31.

Linh, H., & Nguyen, H. (2019). A Case Study on Meaning Representation for Vietnamese. *Proceedings of the First International Workshop on Designing Meaning Representations*, 148–153. https://doi.org/10.18653/v1/W19-3317.

Migueles-Abraira, N. (2017). *A Study Towards Spanish Abstract Meaning Representation* [Master thesis]. Universidad del País Vasco.

Migueles-Abraira, N., Agerri, R., & Diaz de Ilarraza, A. (2018). Annotating Abstract Meaning Representations for Spanish. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, 3074–3078.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *Proceedings of the International Conference on Learning Representations*, 1–12.

Palmer, M., Gildea, D., & Kingsbury, P. (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, *31*(1), 71–106. https://doi.org/10.1162/0891201053630264.

Ramisch, C., Ramisch, R., Zilio, L., Villavicencio, A., & Cordeiro, S. (2018). A Corpus Study of Verbal Multiword Expressions in Brazilian Portuguese. *Proceedings of the International Conference on Computational Processing of the Portuguese Language*, 24–34. https://doi.org/10.1007/978-3-319-99722-3_3.

Sobrevilla Cabezudo, M. A., & Pardo, T. A. S. (2019). Towards a general Abstract Meaning Representation corpus for Brazilian Portuguese. *Proceedings of the 13th Linguistic Annotation Workshop*, 236–244. https://doi.org/10.18653/v1/W19-4028.

Sobrevilla Cabezudo, M. A., & Pardo, T. A. S. (2022). Low-resource AMR-to-Text Generation: A Study on Brazilian Portuguese. *Procesamiento del Lenguaje Natural*, *68*, 85–97.

Yampolskiy, R.V. (2013). Turing Test as a Defining Feature of AI-Completeness. In X. S. Yang (Ed.), *Artificial Intelligence, Evolutionary Computation and Metaheuristics*, pp. 3-17. Springer.

White, A. S., Reisinger, D., Sakaguchi, K., Vieira, T., Zhang, S., Rudinger, R., Rawlins, K., & Van Durme, B. (2016). Universal decompositional semantics on Universal Dependencies. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1713–1723.

Wittenberg, E., Jackendoff, R., Kuperberg, G., Paczynski, M., Snedeker, J., Wiese, H., & Wittenberg, E. (2014). The processing and representation of light verb constructions. In A. Bachrach, I. Roy, & L. Stockall (Eds.), *Structuring the argument: Multidisciplinary research on verb argument structure*, 10, 61–80. John Benjamins Publishing Company.

Xue, N., Bojar, O., Hajič, J., Palmer, M., Urešová, Z., & Zhang, X. (2014). Not an interlingua, but close: Comparison of English AMRs to Chinese and Czech. *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, 1765–1772.

CHAPTER

5

CROSS-LINGUAL APPROACHES

This chapter presents works that aim to show how to leverage the knowledge provided by the English AMR corpus to improve the AMR-to-Text generation task for Brazilian Portuguese and answer the following research questions:

- *is it possible to leverage the cross-linguistic potential of the English AMR corpus for increasing the size of the Portuguese AMR corpus and the performance of AMR-to-text generation?*
- *what is the best strategy for dealing with data sparsity in AMR-to-text generation?*
- *what is the best way to leverage the knowledge provided by the English AMR corpus?*

The chapter is divided in three sections. The first section brings a paper that aims to verify the potential use of a translated English AMR corpus for the text generation task in Brazilian Portuguese. The second section presents a study on diverse approaches and criteria for dealing with low-resource AMR-to-Text generation task. Finally, the last section brings a paper that focus on evaluating cross-lingual approaches for improving the text generation task.

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

This section comprehends the paper below.

CABEZUDO, M. A. S.; MILLE, S.; PARDO, T. 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). Hong

Kong, China: Association for Computational Linguistics, 2019. p. 94–103. Available at <https://aclanthology.org/D19-6313/>.

Contributions:

- Translation of the English AMR corpus to help the text generation in Brazilian Portuguese.
- Exploring criteria to better select instances from an English AMR corpus in order to improve the performance on the AMR-to-Text generation task for Brazilian Portuguese.

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.
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).

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

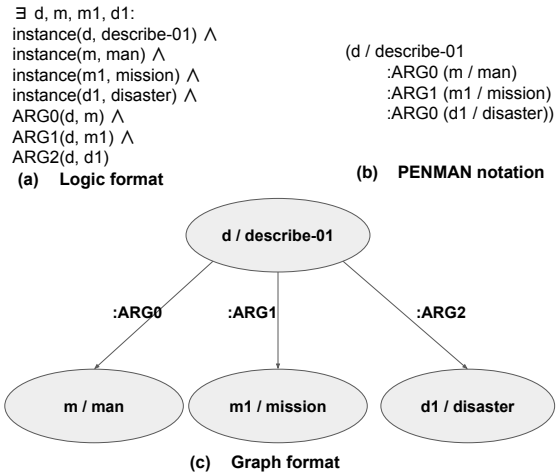


Figure 1: AMR example for the sentence "The man described the mission as a disaster"

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.

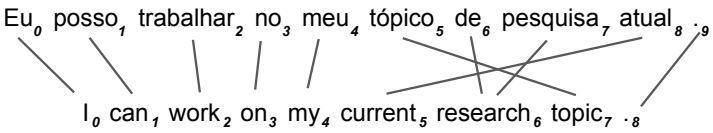
Besides, the AMR graph pre-processing was composed of a delexicalization and a pre-ordering step without compression (described as +Delex-Compress+Preorder in the original paper) as this configuration got one of the best results.

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.



(a) Alignments between English and Brazilian Portuguese sentences

```
(p / possible-01~e.1                    (p / possible-01~e.1
    :ARG1 (w / work-01~e.2                  :ARG1 (w / trabalhar-01~e.2
        :ARG0 (i / i~e.0,4)                     :ARG0 (i / eu-eu~e.4,0)
        :ARG1~e.3 (t / topic~e.7                :ARG1~e.3 (t / tópico~e.5
            :mod (r / research-01~e.6               :mod (r / research-01~e.7,6
                :ARG0 i)                                :ARG0 i)
            :time (c / current~e.5))))              :time (c / atual~e.8))))
```

(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

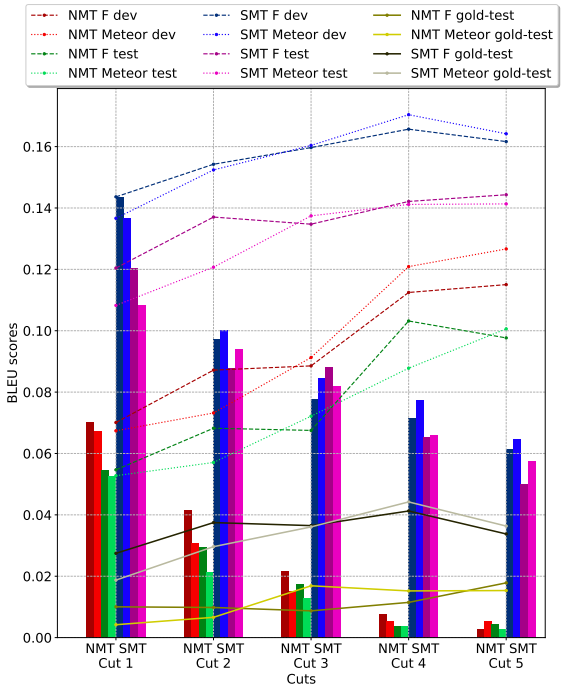


Figure 3: BLEU scores

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.

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.

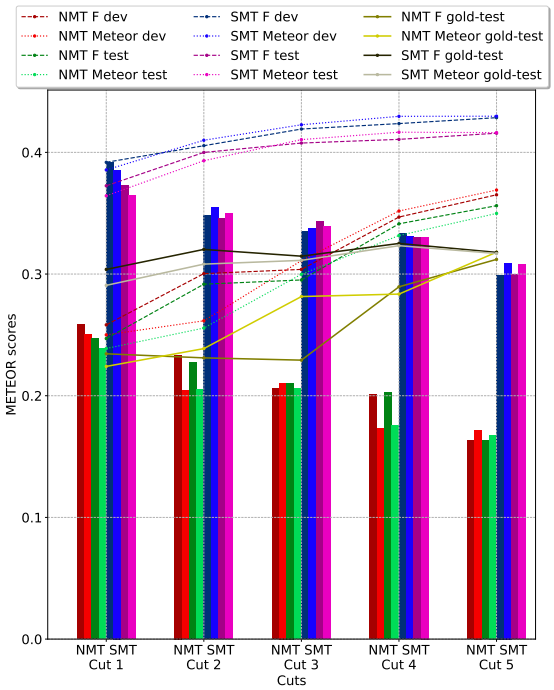


Figure 4: METEOR scores

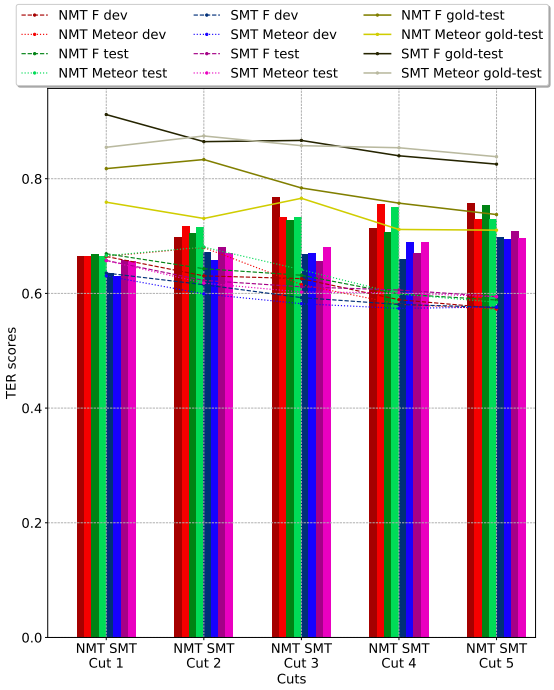


Figure 5: TER scores

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.

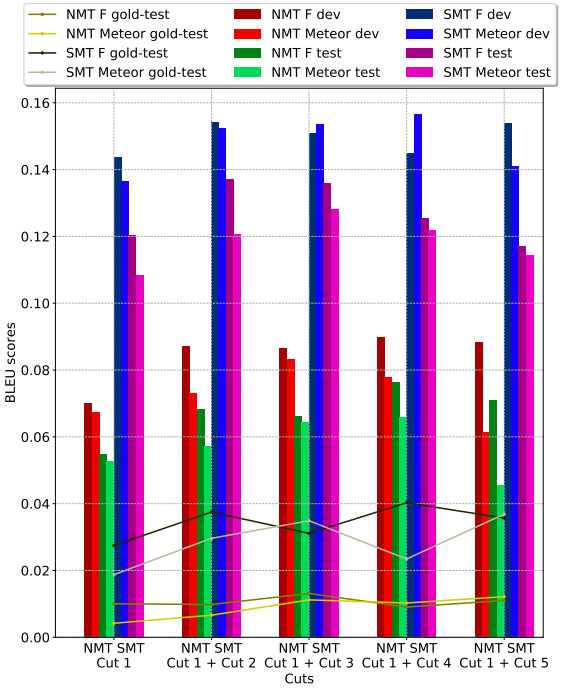


Figure 6: BLEU scores for the cut 1 plus the other cuts

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

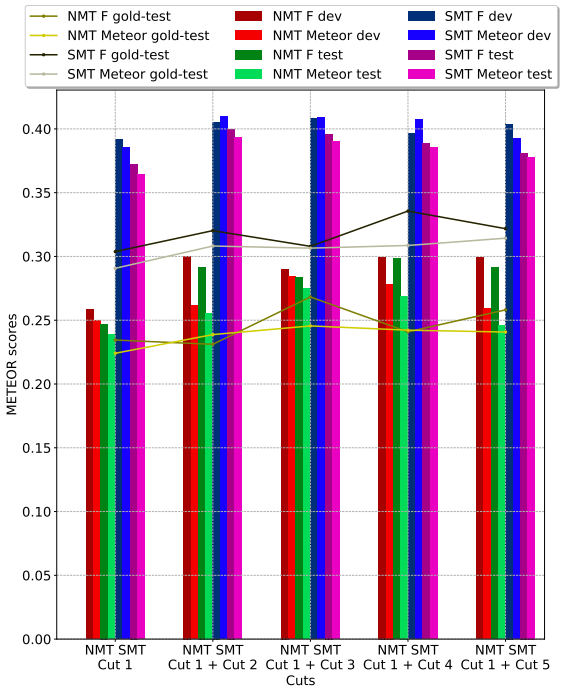


Figure 7: METEOR scores for the cut 1 plus the other cuts

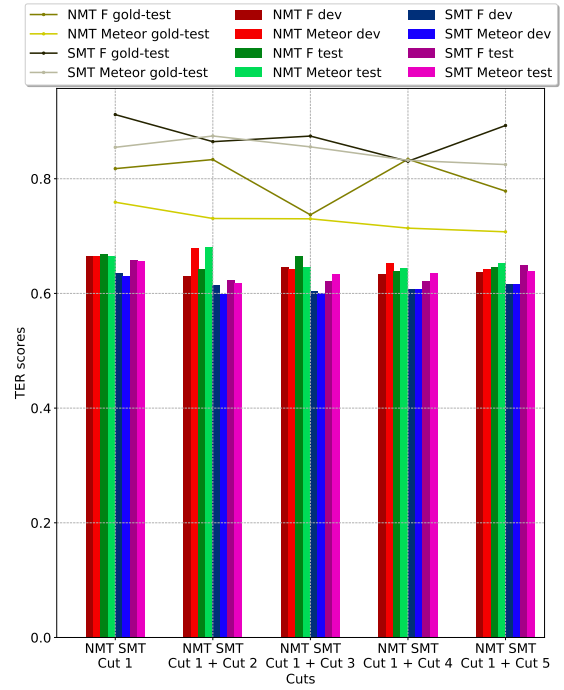


Figure 8: TER scores for the cut 1 plus the other cuts

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.

5.2 Low-resource AMR-to-Text Generation: A Study on Brazilian Portuguese

This section comprises the paper below.

CABEZUDO, M. A. S.; PARDO, T. Low-resource AMR-to-Text Generation: A Study on Brazilian Portuguese. Procesamiento del Lenguaje Natural, Vol 68. p. 85-97. Sociedad Española para el Procesamiento del Lenguaje Natural. Available at <http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6408>.

Contributions:

- Study of three approaches for tackling the low-resource AMR-to-Text generation task: Statistical Machine Translation, Neural Machine Translation, and Graph-to-Sequence-based models.
- Comparative study of diverse criteria such as the granularity of the input representations and linearization strategies for each approach.

Procesamiento del Lenguaje Natural, Revista nº 68, marzo de 2022, pp. 85-97 recibido 19-12-2021 revisado 30-01-2022 aceptado 05-02-2022

Low-resource AMR-to-Text Generation: A Study on Brazilian Portuguese

Generación de Texto a partir de AMR en Contexto de Bajos Recursos: Un Estudio para el Portugués Brasileño

Marco Antonio Sobrevilla Cabezudo, Thiago Alexandre Salgueiro Pardo
Interinstitutional Center for Computational Linguistics (NILC)
Institute of Mathematical and Computer Sciences, University of São Paulo
msobrevillac@usp.br, taspardo@icmc.usp.br

Abstract: This work presents a study of how varied strategies for tackling low-resource AMR-to-text generation for three approaches are helpful in Brazilian Portuguese. Specifically, we explore the helpfulness of additional *translated* corpus, different granularity levels in input representation, and three preprocessing steps. Results show that *translation* is useful. However, it must be used in each approach differently. In addition, finer-grained representations as characters and subwords improve the performance and reduce the bias on the development set, and preprocessing steps are helpful in different contexts, being delexicalisation and preordering the most important ones.
Keywords: AMR-to-Text Generation, Low-resource setting, Brazilian Portuguese.

Resumen: Este trabajo presenta un estudio de cómo diversas estrategias para abordar la generación de textos a partir de AMR en contextos de bajos recursos para tres enfoques son útiles en portugués brasileño. Específicamente, exploramos la utilidad de un corpus traducido, diferentes niveles de granularidad en la representación de entradas y tres técnicas de preprocesamiento. Los resultados muestran que el corpus traducido es útil. Sin embargo, debe usarse en cada enfoque de manera diferente. Además, las representaciones más detalladas, como las basadas en caracteres y subpalabras, mejoran el rendimiento y reducen el sesgo en el conjunto de validación, y los pasos de preprocesamiento son útiles en diferentes contextos, siendo la deslexicalización y el preordenamiento los más importantes.
Palabras clave: Generación de Texto a partir de AMR, Contexto de Bajos Recursos, Portugués Brasileño.

1 Introduction

Abstract Meaning Representation (AMR) is a semantic formalism that encodes the meaning of a sentence as a rooted, acyclic, labeled, and directed graph (Banarescu et al., 2013). This representation includes several semantic information, like semantic roles and named entities, among others.

AMR has become a relevant research topic in meaning representation, semantic parsing, and natural language generation (NLG). Its success is grounded on its attempt to abstract away from syntactic idiosyncrasies, and surface forms, its wide use of mature linguistic resources such as PropBank (Palmer, Gildea, and Kingsbury, 2005), and its usefulness on tasks like text summarisation (Liao, Lebanoff, and Liu, 2018), event detection (Li et al., 2015a) and machine translation (Song et al., 2019).

The goal of the AMR-to-Text generation task is to produce a text that represents the meaning encoded by an input AMR graph. For English, there are several works and approaches for this, as techniques of Statistical Machine Translation (Pourdamghani, Knight, and Hermjakob, 2016), tree and graph to string transducers (Flanigan et al., 2016) and, recently, neural models following sequence-to-sequence (Castro Ferreira et al., 2017; Konstas et al., 2017) and graph-to-sequence architectures (Beck, Haffari, and Cohn, 2018) or pretrained models (Mager et al., 2020; Ribeiro et al., 2020). For other languages, there are some multilingual work (Fan and Gardent, 2020) that tries to generate sentences in several languages. However, they use the AMR for English as in-

ISSN 1135-5948 DOI 10.26342/2022-68-6

put and do not capture some particular linguistic phenomena. In a different line, Sobrevilla Cabezudo, Mille, and Pardo (2019) try to generate Brazilian Portuguese (BP) sentences from the corresponding AMR for BP; nonetheless, the corpus is small (only 299 instances).

One problem that limits the research in other languages is the difficulty to get high-quality corpora (due to the difficult and expensive annotation task that it represents), resulting in smaller corpora and the inability for state-of-the-art methods to be replicated and/or achieve similar performance to the English ones.

It is well-known that the lack of data deteriorates the performance produced by neural models, which usually are data-hungry. To tackle this problem, some authors make use of data augmentation techniques, cross-lingual projection, and other strategies for increasing the corpus size (Hedderich et al., 2021). In the case of AMR-to-text generation, Sobrevilla Cabezudo, Mille, and Pardo (2019) proposed to translate both AMR and English sentences to their corresponding BP ones and then used the translated corpus as training/development set and a gold BP subset as test.

One problem associated with scarce corpus is data sparsity. Particularly, sparsity usually happens at input level in Natural Language Processing tasks. Word representation presents problems with unseen and rare words, resulting in low performance. Many works have proposed employing different granularities in input representation to solve this problem. The most commonly used are subwords (specifically Byte-pair encoding) (Sennrich, Haddow, and Birch, 2016) and characters, resulting in better results. In AMR-to-text generation, some work (Konstas et al., 2017; Mager et al., 2020) used finer-grained representations producing improvements; however, its benefits have not been studied in depth in low-resource settings.

This work explores three different strategies on three approaches for tackling low-resource AMR-to-text generation in Brazilian Portuguese. Specifically, we focus on machine translation and graph-to-sequence-based approaches and study the helpfulness of adding a *translated* corpus, using finer-grained representations and applying diverse preprocessing strategies.

It is worth noting that, even though the current state-of-the-art model for this task uses pretrained models (Mager et al., 2020; Ribeiro et al., 2020) and there are pretrained models for Brazilian Portuguese (Carmo et al., 2020), our goal is to show how to use simpler models and what kind of information could be helpful in low-resource settings or for other languages in which there are no pretrained models.

In general, our main contributions are:

- An analysis of the helpfulness of an additional translated corpus in different settings;
- An exploratory study about the effects of diverse granularity levels in input representation for low-resource AMR-to-text generation; and,
- A deep analysis of three commonly used preprocessing strategies in AMR-to-text generation: delexicalisation, compression, and linearisation.

We start by briefly reviewing AMR fundamentals (Section 2) and presenting the main related work (Section 3). Section 4 reports the techniques and methods that we investigate, while the achieved results are discussed in Section 5. Section 6 concludes this paper.

2 *Abstract Meaning Representation*

As previously mentioned, AMR aims to encode the meaning of a sentence in a directed, labeled, acyclic, and rooted graph (Banarescu et al., 2013). Furthermore, this representation may comprehend semantic information related to semantic roles, named entities, spatial-temporal information and coreferences, among others.

Figure 1 presents an example of an AMR graph for the sentence "The boy destroyed the room". It is worth noting that, as AMR abstracts away the syntactic information, multiple possible sentences can correspond to this graph. This way, another possible sentence that represents the graph could be "the destruction of the room by the boy".

The current AMR-annotated corpus for English contains 59,255 instances[1]. For Non-English languages, there are some efforts to

[1] `https://catalog.ldc.upenn.edu/LDC2020T02`

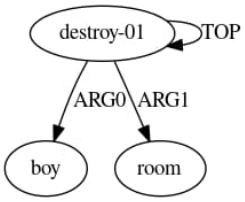


Figure 1: AMR example for the sentence "The boy destroyed the room.".

build corpora leveraging the alignments and existing parallel corpora by using AMR as an interlingua (Xue et al., 2014; Anchiêta and Pardo, 2018). Additionally, other works adapt the AMR guidelines to their languages (Sobrevilla Cabezudo and Pardo, 2019). However, most corpora are far from presenting a size similar to the English one.

For Brazilian Portuguese, as far as we know, there are two AMR corpora, one focused on annotating the sentences of "The Little Prince" book (Anchiêta and Pardo, 2018), and another one that contains manually annotated news text sentences (Sobrevilla Cabezudo and Pardo, 2019). Similarly to Banarescu et al. (2013), some concepts of both corpora were annotated using Verbo-Brasil (Duran and Aluísio, 2015), a lexical resource analogous to PropBank (Palmer, Gildea, and Kingsbury, 2005). Concerning the size of these corpora, the "Little Prince" corpus contains 1,527 annotated sentences (instances), and the second corpus comprises 299 instances, being both small and making it hard to replicate results obtained by state-of-the-art methods.

3 Related Work

In the last years, several AMR-to-Text generation methods for English have been proposed. Initially, methods inspired on Statistical Machine Translation (SMT) techniques (Pourdamghani, Knight, and Hermjakob, 2016) and tree-to-string or graph-to-string transducers (Flanigan et al., 2016) were proposed. Recently, neural models as sequence-to-sequence (Neural Machine Translation or NMT) (Castro Ferreira et al., 2017; Konstas et al., 2017) and, mainly, graph-to-sequence (Beck, Haffari, and Cohn, 2018) and pretrained-based ones (Mager et al., 2020), have emerged, outperforming the previous approaches.

To the extent of our knowledge, the only work focused on AMR-to-Text generation for a Non-English language is proposed by Sobrevilla Cabezudo, Mille, and Pardo (2019). The authors explore the automatic construction of an AMR corpus for Brazilian Portuguese (BP) from its English version and evaluate SMT and NMT approaches on a BP test set composed of 299 instances. Other non-English work (Fan and Gardent, 2020) have tried to generate sentences in diverse languages from English AMR graphs. Although the results are promising, this work does not deal with some specific linguistic phenomena as the previous one does.

In what follows, we detail the dataset that we use in this work and the methods that we investigate.

4 AMR-to-Text Generation

4.1 Data

The methods that we investigate are trained on two corpora and their combinations. The first one is an updated version of the AMR corpus for Brazilian Portuguese (Sobrevilla Cabezudo and Pardo, 2019), which represents our target (*gold*) dataset. This version is a manually annotated corpus comprising 870 instances divided into 402, 224, and 244 instances for training, development, and test, respectively.

The second one is a portion of an automatically generated AMR corpus for Portuguese and represents our augmented (*translated*) dataset. This corpus is generated by translating both AMR graphs and sentences from the English AMR corpus[2] to Portuguese and inheriting the alignments between node/edges and surface tokens[3] (Sobrevilla Cabezudo, Mille, and Pardo, 2019).

In general, this corpus comprises 18,219 and 1,027 instances in the training and development set, respectively, that correspond to the higher-quality translations according to BLEU (Papineni et al., 2002) and METEOR (Lavie and Agarwal, 2007) scores.[4] It is worth noting that, differently from the work of Sobrevilla Cabezudo, Mille, and Pardo (2019), that translates only aligned concepts

[2]In this work, we use the LDC2016E25 corpus to perform the experiments.

[3]Surface tokens are those included in the reference sentence.

[4]The actual portion of the dataset contains 20,000 and 1,271 instances for training and development, respectively. However, some instances were filtered out because they presented some format errors.

in the AMR graphs, all concepts in the AMR graphs are translated.

4.2 Machine Translation-based Techniques

AMR-to-text generation receives an AMR graph as an input and generates a text in natural language; however, Machine Translation models are trained on linear input/output pairs. This way, we need to generate a flattened version of the AMR graph as input. Some flattened versions that have been used in the literature are the ones generated by the PENMAN notation (Matthiessen and Bateman, 1991) and the depth-first search (DFS) algorithm. However, other preprocessing steps can generate a flattened AMR version. Figure 2 shows an example of a flattened AMR version for the sentence *A crise na Venezuela foi um assunto que permeou as reuniões.* ("The crisis in Venezuela was an issue that permeated the meetings.").

In order to evaluate how the use of various flattened AMR versions affect the performance in AMR-to-text generation, we explore the strategies that include the preprocessing steps used by Castro Ferreira et al. (2017). In particular, the preprocessing steps are:

- Delexicalisation: that anonymises some entities of the graph;
- Compression: that determines which nodes and relations should be in the flattened graph; and,
- Linearisation: that determines how the nodes and relations should be put into the flattened graph.

We study two machine translation approaches, a statistical phrase-based one (Koehn, Och, and Marcu, 2003) as a strong baseline and one based on neural models (Bahdanau, Cho, and Bengio, 2015) in a similar way to Castro Ferreira et al. (2017).

4.2.1 Statistical Machine Translation (SMT)

The training parameters in SMT are the same of Castro Ferreira et al. (2017) and a 5-gram language model trained on the Brazilian Portuguese corpus provided by Hartmann et al. (2017) by using KenLM (Heafield, 2011). Furthermore, we use Moses (Koehn et al., 2007) to train the statistical machine translation models.

4.2.2 Neural Machine Translation (NMT)

The architecture and the parameters used in NMT are described as follows: the encoder and the decoder are a 1-layer RNN, and a 2-layers RNN with LSTM, each with a 512D hidden unit, respectively. Besides, the RNN decoder also uses bilinear attention (Luong, Pham, and Manning, 2015). Furthermore, the vocabulary is shared, and we apply weight tying between the source, target, and output layers. Additionally, source and target word embeddings are 512D each, and both are trained jointly with the model.

Among other parameters, the maximum sequence length in the decoder is 80, and we apply dropout with a probability of 0.25 in source embeddings. Moreover, models are trained using the Adam optimizer with a learning rate of 0.0003, a learning rate reduce factor of 0.5, and the learning rate decays if perplexity does not improve after 3 checkpoints/epochs. Besides, we use mini-batches of size 16. Finally, we apply early stopping for model selection based on perplexity scores. Training is halted if a model does not improve on the development set for more than 8 checkpoints/epochs. Sockeye[5] (Hieber et al., 2017) provides all other parameters.

4.3 Graph-to-Sequence (G2S)

Unlike previous approaches, which depend on preprocessing steps and can lose information, the Graph-to-Sequence approach tries to capture the whole graph information more effectively. This work also follows the Graph-to-Sequence approach proposed by Beck, Haffari, and Cohn (2018), that models AMR graphs using a Gated Graph Neural Network (GGNN) (Li et al., 2015b).

In general, model input is defined by the nodes (concepts and relations) and positional embeddings of a graph. To consider AMR relations as nodes, the authors transform the original AMR graph into its respective Levi graph[6] (Levi, 1942). Finally, the output is a version of the original sentence.

We use the same architecture and parameters as Beck, Haffari, and Cohn (2018). Thus, the number of layers in the GGNN encoder

[5] `https://github.com/beckdaniel/sockeye/`

[6] A Levi graph is a modification of a labeled graph so that relations are converted into nodes generating an unlabeled graph.

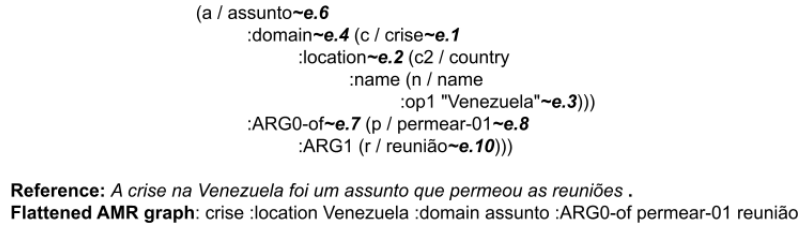


Figure 2: Sentence *A crise na Venezuela foi um assunto que permeou as reuniões.* ("The crisis in Venezuela was an issue that permeated the meetings."), its corresponding AMR graph and a flattened version that includes only aligned nodes/edges. Alignments in AMR graph are in bold.

is 8. All dimensionalities are fixed at 512D except for the GGNN encoder, which uses 576D. The decoder uses a 2-layer LSTM and the Bilinear attention proposed by (Luong, Pham, and Manning, 2015). The remained parameters are the same as the NMT approach.

4.4 Preprocessing Strategies

The preprocessing strategies that we test in this work include:

- Delexicalisation: we delexicalise constants like named-entities or numbers, replacing the original information for tags such as __name1__ and __quant1__ for NMT (Castro Ferreira et al., 2017) and person_1 and quantity_1 for G2S (Beck, Haffari, and Cohn, 2018). A list of tag-values is kept, aiming to rebuild the output sentence after generation;
- Compression: it is performed using a Conditional Random Field (CRF) and executed sequentially over a flattened representation obtained by depth-first search through the AMR graph, and its name and the parent name represent each element. We use the CRFSuite toolkit[7] (Okazaki, 2007) to train our model;
- Linearisation: we apply two strategies. The first consists of performing a depth-first search through the AMR graph, printing the elements (nodes and edges) according to the visiting order. The other strategy is based on the 2-step maximum entropy classifier developed by Lerner and Petrov (2013) and adapted by Castro Ferreira et al. (2017) (we called it preordering). Given an AMR graph represented by a tree, this consists of ordering a head and its corresponding subtrees, i.e., defining which subtrees should be at left/right of the head, and then ordering the subtrees in each built group (left and right side of the head).

All models are tested on inputs/outputs that include or not the preprocessing steps. However, we only explore compression and linearization (preordering) for SMT and delexicalisation for G2S. In addition, when compression is not considered, we include all elements from an AMR graph (nodes and edges).

4.5 Representation Levels

We explore three different representation levels for both input (AMR graph) and output (sentence): words, subwords, and characters. It is expected that finer-grained representations, such as subwords and characters, produce better results, handling in a better way rare words or even possible mismatches between the *translated* and the *gold* corpora.

Subwords are generated by using the Bertimbau's vocabulary provided by Souza, Nogueira, and Lotufo (2020)[8] that uses the sentencepiece tool[9] and the BPE algorithm (Sennrich, Haddow, and Birch, 2016). In the case of the flattened AMR graph, we do not decompose the relations. This way, relations such as "*:ARG0*" or "*:mod*" are kept intact, differently from concepts, such as "*ferida*", that are changed to "*fer ##ida*" in the case of subwords and "*f e r i d a*" in the case of characters.

It is worth mentioning that, in the case of G2S, each subword/character is represented

[7] https://www.chokkan.org/software/crfsuite/

[8] https://github.com/neuralmind-ai/portuguese-bert

[9] https://github.com/google/sentencepiece

by a node, and all subwords/characters that compose a concept are linked sequentially in two directions. For example, we create an edge from subword "fer" to "##ida" and vice-versa.

We present and analyze the achieved results in what follows.

5 Results and Analysis

Tables 1, 2, 3, and 4 show the overall results for SMT, NMT and G2S approaches in terms of BLEU (Papineni et al., 2002), METEOR (Lavie and Agarwal, 2007), and chrF++ (Popović, 2017) evaluation metrics [10]. The tables contain the results when the *translated* corpus (T), the *gold* corpus (G), a join of the training *translated* and *gold* corpora (T + G), and a join of the training/development *translated* and *gold* corpora (T + G Train/dev) are used. In addition, the results of using some preprocessing steps and representation levels are shown. Preprocessing steps are identified as +D (delexicalisation), +C (compression), and +P (preordering) and the opposite when these are not included in the preprocessing.

In general, the best result[11] for SMT happens when we train the model on T + G and use compression and preordering. Likewise, the best result for NMT occurs when the training is performed on T + G, using delexicalisation and preordering, and char-level representations. At last, G2S performs better when the model is trained on T + G train/dev, and lexicalisation and bpe-level presentation are applied.

Results on *gold* corpus show that SMT is by far the best approach to be used in the case of low-resource settings. It is expected as neural models usually need lots of data to achieve good performance, and SMT uses a pre-built language model that guides the decoding, differently from NMT and G2S in which the language model is built during training. In particular, using compressing (+C) and preordering (+P) produces the best results, being preordering the most critical preprocessing step, similarly to the results obtained by Castro Ferreira et al. (2017).

Concerning neural models, NMT produces the best performance; however, this is far from the SMT one yet. Char-level representation and Delexicalisation (+D) are the best strategies when BLEU is evaluated. However, lexicalisation (-D) is better when the metric is chrF++. Moreover, preordering (+P) seems useful when char-level representation is used. Finally, G2S presents the worst performance, being char-level representation and delexicalisation (+D) the best strategies.

In the following subsections, we will study how the performance changes in different contexts and try to answer three questions: (1) how helpful is the *translated* corpus? (2) what are the most useful preprocessing steps? (3) how fine-grained should be the representations to achieve better performance?

5.1 How helpful is the *translated* corpus?

To determine the helpfulness of the *translated* corpus, we study the performance when models are trained on T and T + G.

In general, the *translated* corpus is helpful as all models trained on it present better results than models trained on only *gold* corpus, however, there exists a mismatch between *translated* and *gold* corpora, as values for all measures in development set are quite higher than the obtained in test set (see results on *translated* corpus - T). This behavior can be generated by domain mismatch, in which the vocabulary is different even though both corpora are on news, or by structure mismatch between AMR graphs, since *translated* AMR graphs are English-biased and can introduce noise during training (as its size is bigger than the *gold* corpus).

Regarding the change in the performance when *gold* corpus is added to the *translated* one (T + G), SMT gets leveraging the data increase better. On the other hand, NMT performance presents a slight improvement when *gold* corpus is added. Finally, the G2S performance slightly drops in all cases and can suggest that there is a structural mismatch between the *translated* and gold AMR graphs, as this approach considers structural information, different from SMT or NMT, which use a flattened version with some nodes/edges included in it.

In order to evaluate how to deal with the possible mismatch, we add the *translated* development set (1,027 instances) to the *gold* one as well. Table 4 shows the result for each

[10] We execute 4 runs for each experiment and show the mean and standard deviation for NMT and G2S.

[11] Best results are highlighted in bold in Tables.

		DEV			TEST		
		BLEU	METEOR	chrF++	BLEU	METEOR	chrF++
Gold	+C+P	11.58	0.31	0.48	10.00	0.30	0.48
	+C-P	11.36	0.29	0.47	7.95	0.26	0.46
	-C+P	6.06	0.24	0.43	6.05	0.24	0.43
	-C-P	7.31	0.24	0.44	4.89	0.22	0.43
Translated	+C+P	27.18	0.45	0.57	9.98	0.29	0.47
	+C-P	26.10	0.44	0.56	10.50	0.28	0.46
	-C+P	23.73	0.42	0.55	10.47	0.30	0.48
	-C-P	24.02	0.42	0.55	7.83	0.26	0.46
Translated + Gold	+C+P	**18.67**	**0.38**	**0.52**	**14.83**	**0.33**	**0.49**
	+C-P	17.75	0.37	0.51	11.96	0.32	0.47
	-C+P	17.38	0.37	0.51	13.91	0.32	0.49
	-C-P	14.86	0.35	0.50	11.96	0.32	0.48

Table 1: Overall SMT results.

setting and approach. Unlike the previous setting (T+G), both SMT and NMT present a small improvement in all metrics. However, G2S presents bigger improvements, suggesting that adding *translated* instances can make models more robust to possible structural divergences, leading to performance improvements.

5.2 What are the most useful preprocessing strategies?

5.2.1 Statistical Machine Translation

Pre-ordering (+P) seems to lead to improvements, however, this improvement is notorious when translated + *gold* corpora are used in the training set. Another point to highlight is the importance of compression (+C). Initial experiments (T and T + G) show that compression leads to slight improvements. However, no compression (-C) produces the best results when the classifier is trained on T + G train/dev.

5.2.2 Neural Machine Translation

Delexicalisation (+D) seems to be a good strategy for word and char-level representations, but it is not relevant for bpe-level. Moreover, compression (+C) generally harms the performance or produces mixed results, being better when lexicalisation (-D) is applied in char-level representation. Finally, pre-ordering (+P) seems to produce small improvements in all settings.

5.2.3 Graph-to-Sequence

About Graph-to-Sequence approach, Delexicalisation (+D) improves the performance when word and char-level presentations are used. However, the contrary happens when bpe-level representation is used. A possible explanation is that delexicalisation reduces data sparseness when word-level representation is applied together and allows to deal with large graphs in the case of char-level representation. However, in the case of bpe, delexicalisation seems to introduce noise and makes the model more prone to generate hallucinations.

5.3 How fine-grained should be the representations to achieve better performance?

Concerning the representation levels, characters and bpe produce the best and second-best performance for NMT. The main gain in both representations is in terms of METEOR and chRF++, which is expected as these representations are finer-grained and the evaluation measures take stems and characters into account.

Different from NMT, bpe produces the best performance for G2S. However, and as it was previously mentioned, this performance happens when delexicalisation is applied. This way, we hypothesise two possible problems: (1) word-level representations suffer more from mismatch problems as experiments on T and T + G show low performance, and (2) char-level representations can generate larger AMR graphs for which semantics can be challenging to be captured by G2S.

Another point to highlight is that finer-grained representations usually help reducing the bias to the development set, mainly when char-level representations are used. Consequently, mismatch problems are mitigated. This can be seen in the difference between development and test performance for experiments on T and T + G train/dev. For example, Figure 3 shows the difference mentioned for NMT. Experiments on T + G present a BLEU overall difference of 10.45, 9.9, and 5,67 between development and test for word, bpe, and char-level representations. Similarly, differences for METEOR and chrF++ are 0.11, 0.11, and 0.03, and 0.11, 0.09, and 0.00, respectively.

			DEV			TEST		
			BLEU	METEOR	chrF++	BLEU	METEOR	chrF++
G	word	+D+C+P	0.00±0.00	0.06±0.00	0.05±0.00	2.66±0.14	0.10±0.00	0.13±0.01
		+D+C-P	0.87±0.87	0.10±0.01	0.12±0.00	2.48±0.37	0.11±0.00	0.14±0.01
		+D-C+P	0.00±0.00	0.10±0.01	0.11±0.00	2.61±0.23	0.11±0.00	0.13±0.00
		+D-C-P	0.37±0.63	0.10±0.01	0.11±0.01	2.39±0.18	0.10±0.00	0.13±0.01
		-D+C+P	0.00±0.00	0.03±0.02	0.02±0.02	0.00±0.00	0.03±0.02	0.02±0.02
		-D+C-P	0.00±0.00	0.03±0.02	0.02±0.02	0.00±0.00	0.03±0.02	0.02±0.02
		-D-C+P	0.00±0.00	0.02±0.00	0.01±0.00	0.00±0.00	0.02±0.00	0.01±0.00
		-D-C-P	0.00±0.00	0.02±0.00	0.01±0.00	0.00±0.00	0.02±0.00	0.01±0.00
	bpe	+D+C+P	0.00±0.00	0.02±0.00	0.01±0.00	0.00±0.00	0.01±0.00	0.01±0.00
		+D+C-P	0.34±0.58	0.05±0.04	0.07±0.05	0.88±0.90	0.06±0.04	0.08±0.06
		+D-C+P	0.33±0.56	0.07±0.03	0.09±0.04	1.33±0.81	0.08±0.04	0.10±0.05
		+D-C-P	0.33±0.57	0.03±0.03	0.04±0.05	0.39±0.67	0.03±0.03	0.04±0.05
		-D+C+P	0.00±0.00	0.03±0.02	0.02±0.02	0.00±0.00	0.03±0.02	0.02±0.02
		-D+C-P	0.00±0.00	0.03±0.02	0.02±0.02	0.00±0.00	0.03±0.02	0.02±0.02
		-D-C+P	0.00±0.00	0.02±0.00	0.01±0.00	0.00±0.00	0.02±0.00	0.01±0.00
		-D-C-P	0.00±0.00	0.02±0.00	0.01±0.00	0.00±0.00	0.02±0.00	0.01±0.00
	char	+D+C+P	0.59±0.67	0.11±0.03	0.22±0.06	3.12±0.37	0.15±0.02	0.26±0.05
		+D+C-P	1.61±1.00	0.11±0.01	0.19±0.01	2.80±0.27	0.11±0.00	0.19±0.00
		+D-C+P	2.28±0.36	0.12±0.01	0.22±0.04	3.12±0.10	0.13±0.01	0.22±0.03
		+D-C-P	1.63±0.09	0.10±0.00	0.18±0.00	2.88±0.35	0.11±0.00	0.19±0.00
		-D+C+P	1.35±0.82	0.14±0.05	0.27±0.09	1.77±1.14	0.14±0.05	0.28±0.09
		-D+C-P	0.00±0.00	0.11±0.00	0.26±0.00	0.48±0.82	0.13±0.01	0.27±0.01
		-D-C+P	1.45±0.87	0.16±0.01	0.31±0.01	0.70±1.22	0.16±0.01	0.31±0.01
		-D-C-P	0.72±0.74	0.09±0.04	0.20±0.07	0.00±0.00	0.09±0.04	0.19±0.07
—	word	+D+C+P	11.02±1.37	0.26±0.02	0.32±0.01	4.16±0.65	0.20±0.01	0.29±0.01
		+D+C-P	4.66±0.19	0.18±0.01	0.24±0.01	2.46±0.29	0.13±0.00	0.19±0.00
		+D-C+P	20.53±0.56	0.38±0.00	0.46±0.00	5.88±0.23	0.24±0.01	0.33±0.01
		+D-C-P	19.35±0.92	0.37±0.01	0.44±0.00	5.88±0.30	0.23±0.00	0.32±0.01
		-D+C+P	17.96±0.76	0.36±0.01	0.42±0.01	3.79±0.34	0.18±0.01	0.25±0.01
		-D+C-P	2.32±0.37	0.12±0.01	0.16±0.01	0.12±0.21	0.06±0.00	0.09±0.01
		-D-C+P	19.22±0.75	0.38±0.01	0.43±0.02	3.96±0.62	0.18±0.01	0.26±0.02
		-D-C-P	19.81±0.77	0.37±0.01	0.42±0.01	3.17±0.33	0.17±0.01	0.24±0.01
T	bpe	+D+C+P	8.96±2.07	0.26±0.02	0.36±0.01	3.90±1.03	0.21±0.02	0.32±0.01
		+D+C-P	12.89±3.52	0.33±0.02	0.44±0.02	3.57±1.10	0.21±0.02	0.33±0.01
		+D-C+P	15.41±2.46	0.36±0.02	0.46±0.01	5.39±0.68	0.24±0.01	0.36±0.00
		+D-C-P	20.04±0.60	0.38±0.00	0.48±0.01	7.05±1.00	0.27±0.02	0.38±0.01
		-D+C+P	19.34±4.59	0.41±0.03	0.49±0.02	6.10±1.42	0.24±0.03	0.36±0.02
		-D+C-P	13.60±2.37	0.36±0.02	0.46±0.01	2.86±0.74	0.19±0.01	0.32±0.01
		-D-C+P	22.39±1.57	0.44±0.01	0.51±0.00	7.08±0.71	0.27±0.02	0.37±0.02
		-D-C-P	20.87±1.16	0.42±0.01	0.50±0.01	5.47±0.63	0.24±0.01	0.35±0.00
	char	+D+C+P	13.39±0.37	0.27±0.00	0.37±0.00	8.69±1.33	0.29±0.01	0.43±0.01
		+D+C-P	15.45±0.50	0.31±0.00	0.43±0.01	8.02±0.40	0.28±0.01	0.42±0.01
		+D-C+P	13.73±0.40	0.31±0.00	0.43±0.01	8.21±0.95	0.28±0.01	0.42±0.01
		+D-C-P	13.06±1.22	0.29±0.01	0.42±0.01	7.18±0.88	0.27±0.00	0.42±0.00
		-D+C+P	16.06±2.91	0.33±0.04	0.43±0.03	7.63±2.23	0.28±0.03	0.42±0.03
		-D+C-P	17.75±0.41	0.34±0.01	0.44±0.01	6.16±1.13	0.26±0.01	0.41±0.00
		-D-C+P	15.73±1.19	0.33±0.02	0.43±0.02	6.97±1.40	0.26±0.02	0.41±0.02
		-D-C-P	11.26±4.63	0.24±0.09	0.34±0.10	4.04±3.64	0.17±0.09	0.29±0.12
T+G	word	+D+C+P	2.77±0.57	0.16±0.01	0.22±0.02	4.76±0.38	0.20±0.01	0.28±0.02
		+D+C-P	3.65±0.54	0.19±0.02	0.27±0.02	4.23±1.00	0.19±0.02	0.27±0.03
		+D-C+P	5.15±0.82	0.23±0.01	0.31±0.01	6.04±0.30	0.22±0.01	0.30±0.01
		+D-C-P	4.42±0.52	0.20±0.01	0.28±0.01	4.81±0.64	0.20±0.01	0.27±0.02
		-D+C+P	2.93±0.73	0.17±0.01	0.24±0.00	3.59±0.38	0.18±0.00	0.24±0.00
		-D+C-P	2.70±0.48	0.14±0.01	0.20±0.01	2.58±0.67	0.14±0.02	0.20±0.01
		-D-C+P	3.51±0.77	0.16±0.02	0.23±0.02	2.57±0.27	0.16±0.02	0.22±0.02
		-D-C-P	3.63±0.89	0.17±0.01	0.24±0.02	2.99±0.80	0.16±0.01	0.23±0.01
	bpe	+D+C+P	2.72±0.73	0.19±0.01	0.30±0.01	4.71±0.38	0.23±0.01	0.34±0.01
		+D+C-P	3.38±1.35	0.20±0.04	0.32±0.03	3.21±1.43	0.19±0.04	0.31±0.03
		+D-C+P	7.10±1.10	0.28±0.02	0.39±0.02	7.52±1.10	0.28±0.02	0.37±0.01
		+D-C-P	5.68±1.21	0.26±0.02	0.37±0.02	5.78±1.38	0.25±0.02	0.35±0.01
		-D+C+P	3.56±0.52	0.21±0.01	0.34±0.01	4.47±1.21	0.22±0.02	0.35±0.01
		-D+C-P	4.45±1.02	0.22±0.02	0.33±0.01	4.60±1.36	0.22±0.02	0.34±0.02
		-D-C+P	7.10±0.40	0.27±0.00	0.37±0.01	7.42±0.70	0.26±0.01	0.36±0.01
		-D-C-P	6.69±0.77	0.26±0.01	0.36±0.01	5.93±1.35	0.25±0.01	0.36±0.01
	char	+D+C+P	7.82±0.44	0.26±0.01	0.38±0.01	9.38±0.22	0.30±0.01	0.44±0.01
		+D+C-P	8.36±0.51	0.29±0.01	0.42±0.01	8.65±0.90	0.28±0.01	0.42±0.01
		+D-C+P	**7.28±0.49**	**0.29±0.01**	**0.42±0.01**	**10.03±0.37**	**0.31±0.01**	**0.44±0.01**
		+D-C-P	7.04±0.14	0.27±0.00	0.42±0.00	7.34±0.88	0.27±0.01	0.41±0.01
		-D+C+P	7.48±0.74	0.29±0.01	0.43±0.01	8.85±0.78	0.29±0.01	0.43±0.01
		-D+C-P	7.99±1.57	0.27±0.01	0.41±0.01	7.96±0.69	0.27±0.01	0.42±0.01
		-D-C+P	5.98±0.59	0.27±0.02	0.41±0.02	8.25±0.94	0.29±0.02	0.43±0.02
		-D-C-P	5.33±1.89	0.23±0.05	0.37±0.05	5.20±3.06	0.24±0.05	0.38±0.05

Table 2: Overall NMT results.

5.4 Manual Revision

We present now some analysis of actual generated cases. Figure 4 shows the AMR graph, the reference, and the output generated by the three approaches for the sentences "He/She does not want it" ("*não quer*") and "He/She attended excellent schools, and majored in economics at Yale." ("*frequentou excelentes escolas, e se formou em economia por Yale.*"). We can see some mistakes for each approach associated with hidden subjects (highlighted in red), wrong conjugation (blue), fluency/concordance (green), repetitions (purple), random words (yellow), and entity copying (pink).

The first example is simple, and the three approaches present similar outputs. SMT produces almost the same reference; however,

			DEV			TEST		
			BLEU	METEOR	chrF++	BLEU	METEOR	chrF++
G	word	+D	0.00 ±0.00	0.03 ±0.01	0.02 ±0.01	0.00 ±0.00	0.03 ±0.01	0.02 ±0.01
		-D	0.00 ±0.00	0.03 ±0.01	0.02 ±0.01	0.00 ±0.00	0.03 ±0.01	0.02 ±0.01
	bpe	+D	0.00 ±0.00	0.03 ±0.01	0.03 ±0.02	0.00 ±0.00	0.03 ±0.02	0.03 ±0.02
		-D	0.00 ±0.00	0.02 ±0.01	0.01 ±0.00	0.00 ±0.00	0.02 ±0.01	0.01 ±0.00
	char	+D	0.00 ±0.00	0.09 ±0.01	0.13 ±0.01	1.59 ±0.47	0.09 ±0.01	0.14 ±0.01
		-D	0.00 ±0.00	0.05 ±0.00	0.09 ±0.00	0.00 ±0.00	0.05 ±0.00	0.09 ±0.00
T	word	+D	14.88 ±4.17	0.32 ±0.06	0.38 ±0.06	4.66 ±1.50	0.18 ±0.04	0.26 ±0.05
		-D	10.41 ±4.20	0.24 ±0.06	0.30 ±0.07	1.95 ±1.74	0.13 ±0.04	0.19 ±0.05
	bpe	+D	8.44 ±1.60	0.23 ±0.02	0.29 ±0.01	2.60 ±0.37	0.14 ±0.00	0.22 ±0.01
		-D	21.04 ±1.09	0.42 ±0.01	0.48 ±0.00	6.75 ±0.51	0.26 ±0.01	0.36 ±0.01
	char	+D	11.46 ±1.67	0.25 ±0.02	0.32 ±0.03	6.07 ±2.02	0.23 ±0.04	0.35 ±0.05
		-D	7.09 ±2.24	0.18 ±0.03	0.24 ±0.02	1.43 ±0.78	0.12 ±0.03	0.23 ±0.03
T+G	word	+D	3.52 ±2.14	0.17 ±0.05	0.23 ±0.05	3.80 ±2.01	0.16 ±0.04	0.23 ±0.05
		-D	1.00 ±1.74	0.10 ±0.04	0.15 ±0.05	1.00 ±1.72	0.09 ±0.04	0.15 ±0.06
	bpe	+D	1.37 ±0.35	0.12 ±0.01	0.18 ±0.00	1.82 ±0.32	0.12 ±0.01	0.19 ±0.01
		-D	5.62 ±0.43	0.26 ±0.01	0.36 ±0.01	6.44 ±0.79	0.26 ±0.01	0.36 ±0.01
	char	+D	5.21 ±1.25	0.22 ±0.03	0.33 ±0.05	6.09 ±1.50	0.22 ±0.04	0.34 ±0.05
		-D	2.53 ±1.63	0.17 ±0.04	0.28 ±0.05	2.63 ±1.94	0.17 ±0.04	0.29 ±0.05

Table 3: Overall G2S results.

			DEV			TEST		
			BLEU	METEOR	chrF++	BLEU	METEOR	chrF++
SMT	word	+C+P	25.66	0.43	0.56	12.92	0.31	0.48
		+C-P	24.72	0.42	0.55	12.52	0.31	0.48
		-C+P	22.09	0.41	0.54	14.69	0.34	0.50
		-C-P	22.29	0.41	0.54	10.03	0.30	0.48
NMT	word	+D+C+P	11.21 ±1.36	0.25 ±0.01	0.32 ±0.02	5.38 ±1.03	0.22 ±0.02	0.30 ±0.02
		+D+C-P	14.25 ±0.92	0.31 ±0.01	0.39 ±0.01	4.95 ±0.52	0.21 ±0.01	0.29 ±0.01
		+D-C+P	16.82 ±0.81	0.34 ±0.01	0.42 ±0.00	6.70 ±0.79	0.24 ±0.01	0.32 ±0.01
		+D-C-P	17.10 ±0.47	0.34 ±0.00	0.42 ±0.00	6.68 ±0.20	0.23 ±0.00	0.32 ±0.01
		-D+C+P	14.88 ±1.42	0.32 ±0.01	0.38 ±0.01	3.94 ±0.64	0.19 ±0.01	0.26 ±0.01
		-D+C-P	14.98 ±1.48	0.31 ±0.01	0.37 ±0.01	3.25 ±0.51	0.17 ±0.01	0.24 ±0.01
		-D-C+P	17.64 ±0.74	0.35 ±0.01	0.41 ±0.01	4.76 ±0.44	0.20 ±0.01	0.28 ±0.01
		-D-C-P	16.87 ±0.47	0.33 ±0.01	0.39 ±0.01	4.48 ±0.31	0.19 ±0.00	0.26 ±0.01
	bpe	+D+C+P	11.81 ±0.43	0.28 ±0.01	0.37 ±0.01	6.65 ±1.10	0.25 ±0.01	0.35 ±0.01
		+D+C-P	14.32 ±0.87	0.33 ±0.01	0.43 ±0.01	5.09 ±0.54	0.24 ±0.01	0.35 ±0.01
		+D-C+P	16.98 ±3.23	0.37 ±0.02	0.47 ±0.02	7.70 ±1.53	0.27 ±0.01	0.38 ±0.01
		+D-C-P	16.32 ±2.56	0.36 ±0.02	0.45 ±0.01	6.15 ±0.87	0.26 ±0.01	0.36 ±0.01
		-D+C+P	13.80 ±3.03	0.35 ±0.03	0.46 ±0.01	5.61 ±0.82	0.24 ±0.02	0.36 ±0.02
		-D+C-P	14.53 ±3.18	0.35 ±0.02	0.45 ±0.01	4.79 ±1.55	0.22 ±0.02	0.34 ±0.02
		-D-C+P	21.38 ±0.93	0.41 ±0.01	0.48 ±0.01	7.80 ±0.77	0.27 ±0.01	0.37 ±0.01
		-D-C-P	20.25 ±1.06	0.40 ±0.01	0.49 ±0.01	6.38 ±1.16	0.26 ±0.01	0.38 ±0.01
	char	+D+C+P	12.61 ±0.50	0.27 ±0.00	0.37 ±0.00	9.42 ±0.47	0.30 ±0.00	0.44 ±0.00
		+D+C-P	14.59 ±0.43	0.31 ±0.00	0.43 ±0.01	9.07 ±0.80	0.29 ±0.02	0.43 ±0.01
		+D-C+P	13.20 ±0.16	0.31 ±0.00	0.43 ±0.01	9.83 ±0.88	0.30 ±0.01	0.44 ±0.01
		+D-C-P	12.91 ±0.53	0.30 ±0.01	0.42 ±0.01	8.49 ±0.88	0.29 ±0.01	0.42 ±0.01
		-D+C+P	17.18 ±0.54	0.35 ±0.00	0.45 ±0.00	10.14 ±0.38	0.30 ±0.01	0.44 ±0.01
		-D+C-P	16.65 ±0.72	0.33 ±0.01	0.44 ±0.01	8.10 ±0.88	0.28 ±0.01	0.42 ±0.01
		-D-C+P	12.19 ±4.38	0.27 ±0.08	0.37 ±0.09	5.93 ±3.48	0.24 ±0.10	0.36 ±0.12
		-D-C-P	14.58 ±0.58	0.31 ±0.01	0.43 ±0.00	7.61 ±0.82	0.27 ±0.01	0.42 ±0.00
G2S	word	+D	16.84 ±1.88	0.36 ±0.02	0.43 ±0.02	7.70 ±1.74	0.26 ±0.03	0.34 ±0.03
		-D	9.73 ±5.58	0.23 ±0.09	0.29 ±0.09	2.73 ±2.17	0.14 ±0.05	0.20 ±0.06
	bpe	+D	7.59 ±1.97	0.22 ±0.02	0.28 ±0.02	3.28 ±0.74	0.15 ±0.01	0.23 ±0.01
		-D	**20.85 ±1.21**	**0.41 ±0.02**	**0.48 ±0.02**	**8.69 ±0.59**	**0.29 ±0.02**	**0.38 ±0.02**
	char	+D	11.10 ±1.95	0.25 ±0.03	0.32 ±0.02	7.03 ±2.46	0.24 ±0.04	0.35 ±0.05
		-D	7.94 ±1.22	0.22 ±0.02	0.30 ±0.02	4.00 ±0.49	0.19 ±0.01	0.32 ±0.02

Table 4: Results of adding *translated* development set to the *gold* one. It is called T + G train/dev.

this includes the pronoun "*ele*" ("he/she") that is treated as a hidden subject in the reference. Conversely, NMT and G2S omit the pronoun, making the generated sentence more natural; nevertheless, both approaches generate the verb "*querer*" ("want") in a different conjugation (1st person). A possible explanation is that NMT and G2S are trained on char and bpe-level representations, this way, they can generate different conjugations easily. In addition, NMT generates the word "*dizer*" ("to say") that is not part of the AMR graph.

The second one is a harder example with more relations and concepts such as named entities ("university"), co-references ("e1 / *ele*" or "he/she") and connectors ("*e*"). In this case, none of the approaches can omit the pronoun "*ele*" as the reference does. Another common problem in all approaches is the lack of agreement/fluency. For example, the expression "na yale" should be replaced by "*em yale*" in order to be more fluent.

Analyzing other issues, SMT tries to generate sentences with all possible concepts included in the graph, even if the generated text is not fluent. On the other hand, neural models suffer from classical problems such as repetition and random word generation (the hallucination problem).

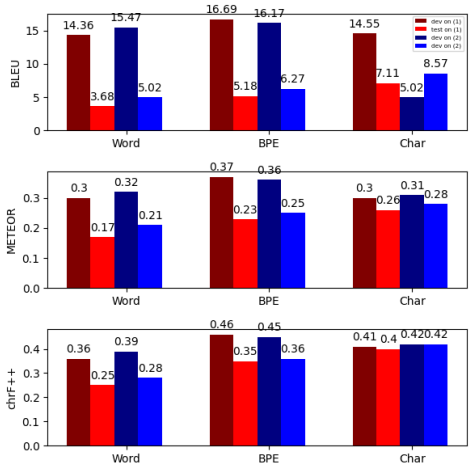


Figure 3: Difference between development and test performance for experiments on (1) T and (2) T + G train/dev.

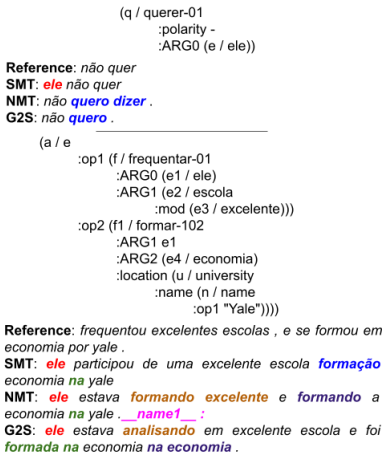


Figure 4: Outputs generated by the different approaches.

6 Conclusion and future work

This work presented a study of different strategies for tackling low-resource AMR-to-text generation for Brazilian Portuguese. We explore the helpfulness of additional translated corpus, different granularity levels in input representation, and three preprocessing strategies. It is worth noting this study can be helpful for work in other languages or meaning representations, mainly, when there is no pretrained models available.

Concerning the use of *translated* corpus, we can confirm its helpfulness. However, there are different contexts for each approach in which we can better leverage it. SMT improves its performance when the model is trained on the *translated* and *gold* corpora together. Neural models benefit from *translated* corpus more than SMT, even when these are trained on it solely. However, its join with the *gold* corpus can produce different results. In particular, G2S showed that there are structural divergences between *translated* and *gold* AMR graphs that can harm the performance when models are trained on both corpora. However, adding *translated* corpus to the development set allows to make the model more robust and achieve better performance.

About the representation levels, we highlight the use of finer-grained representations such as subwords and characters. Char-level seems to be the best option for NMT and bpe for G2S. However, it is worth noting that our study focuses on sentences of 23 tokens at maximum. This way, if we extend the work to longer sentences, bpe would probably performs better than char for NMT.

Finally, different combinations of preprocessing strategies are helpful for each approach, being preordering the best strategy for both machine translation approaches and delexicalisation for NMT. In the case of G2S, delexicalisation produces mixed results, being important just for word and char-level representations.

As future work, we plan to explore state-of-the-art approaches that are usually based on transformers, such as T5 (Ribeiro et al., 2020), or GPT-2 (Mager et al., 2020). Besides such issues, given some divergences between the *translated* and *gold* corpora that can harm the performance, it would be interesting to explore transfer learning for leveraging the knowledge learned from the *translated* corpus instead of training on both corpora together.

To the interested reader, more details about this work may be found at the web portal of the POeTiSA project at `https://sites.google.com/icmc.usp.br/poetisa`.

Acknowledgments

The authors are grateful to CAPES and the Center for Artificial Intelligence (C4AI - `http://c4ai.inova.usp.br/`) of the University of São Paulo, sponsored by IBM and FAPESP (grant #2019/07665-4). Besides, this research has been carried out using the computational resources of the Center for Mathematical Sciences Applied to Industry (CeMEAI) funded by FAPESP (grant 2013/07375-0).

References

Anchiêta, R. and T. 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.

Bahdanau, D., K. Cho, and Y. Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Banarescu, L., C. Bonial, S. Cai, M. Georgescu, K. Griffitt, U. Hermjakob, K. Knight, P. Koehn, M. Palmer, and N. 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.

Beck, D., G. Haffari, and T. Cohn. 2018. Graph-to-sequence learning using gated graph neural networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 273–283. Association for Computational Linguistics.

Carmo, D., M. Piau, I. Campiotti, R. Nogueira, and R. Lotufo. 2020. Ptt5: Pretraining and validating the t5 model on brazilian portuguese data. *arXiv preprint arXiv:2008.09144*.

Castro Ferreira, T., I. Calixto, S. Wubben, and E. 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.

Duran, M. S. and S. M. 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.

Fan, A. and C. Gardent. 2020. Multilingual AMR-to-text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2889–2901, Online, November. Association for Computational Linguistics.

Flanigan, J., C. Dyer, N. A. Smith, and J. Carbonell. 2016. Generation from abstract meaning representation using tree transducers. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 731–739, San Diego, California, June. Association for Computational Linguistics.

Hartmann, N., E. Fonseca, C. Shulby, M. Treviso, J. Silva, and S. 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.

Heafield, K. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, July. Association for Computational Linguistics.

Hedderich, M. A., L. Lange, H. Adel, J. Strötgen, and D. Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, Online, June. Association for Computational Linguistics.

Hieber, F., T. Domhan, M. J. Denkowski, D. Vilar, A. Sokolov, A. Clifton, and M. Post. 2017. Sockeye: A toolkit for neural machine translation. *ArXiv*, abs/1712.05690.

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens,

C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Koehn, P., F. J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127–133. Association for Computational Linguistics.

Konstas, I., S. Iyer, M. Yatskar, Y. Choi, and L. Zettlemoyer. 2017. Neural AMR: Sequence-to-sequence models for parsing and generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 146–157, Vancouver, Canada, July. Association for Computational Linguistics.

Lavie, A. and A. 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.

Lerner, U. and S. Petrov. 2013. Source-side classifier preordering for machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 513–523, Seattle, Washington, USA, October. Association for Computational Linguistics.

Levi, F. W. 1942. Finite geometrical systems.

Li, X., T. H. Nguyen, K. Cao, and R. Grishman. 2015a. Improving event detection with abstract meaning representation. In *Proceedings of the First Workshop on Computing News Storylines*, pages 11–15, Beijing, China, July. Association for Computational Linguistics.

Li, Y., D. Tarlow, M. Brockschmidt, and R. Zemel. 2015b. Gated graph sequence neural networks. *arXiv preprint arXiv:1511.05493*.

Liao, K., L. Lebanoff, and F. Liu. 2018. Abstract meaning representation for multi-document summarization. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1178–1190, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Luong, T., H. Pham, and C. D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal, September. Association for Computational Linguistics.

Mager, M., R. Fernandez Astudillo, T. Naseem, M. A. Sultan, Y.-S. Lee, R. Florian, and S. 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, July. Association for Computational Linguistics.

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

Okazaki, N. 2007. Crfsuite: a fast implementation of conditional random fields (crfs).

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

Papineni, K., S. Roukos, T. Ward, and W.-J. 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.

Popović, M. 2017. chrF++: words helping character n-grams. In *Proceedings of*

the Second Conference on Machine Translation, pages 612–618, Copenhagen, Denmark, September. Association for Computational Linguistics.

Pourdamghani, N., K. Knight, and U. Hermjakob. 2016. Generating English from abstract meaning representations. In *Proceedings of the 9th International Natural Language Generation conference*, pages 21–25, Edinburgh, UK, September 5-8. Association for Computational Linguistics.

Ribeiro, L. F. R., M. Schmitt, H. Schütze, and I. Gurevych. 2020. Investigating pretrained language models for graph-to-text generation. *CoRR*, abs/2007.08426.

Sennrich, R., B. Haddow, and A. Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August. Association for Computational Linguistics.

Sobrevilla Cabezudo, M. A., S. Mille, and T. 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, November. Association for Computational Linguistics.

Sobrevilla Cabezudo, M. A. and T. 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, August. Association for Computational Linguistics.

Song, L., D. Gildea, Y. Zhang, Z. Wang, and J. Su. 2019. Semantic neural machine translation using amr. *Transactions of the Association for Computational Linguistics*, 7:19–31.

Souza, F., R. Nogueira, and R. Lotufo. 2020. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23*.

Xue, N., O. Bojar, J. Hajič, M. Palmer, Z. Urešová, and X. Zhang. 2014. Not an interlingua, but close: Comparison of English AMRs to Chinese and Czech. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1765–1772, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

5.3 Comparison of Cross-lingual strategies for AMR-to-Brazilian Portuguese Generation

This section encompasses the paper below.

CABEZUDO, M. A. S.; ANCHIÊTA, R.T.; PARDO, T. Comparison of Cross-lingual strategies for AMR-to-Brazilian Portuguese Generation, submitted to the Language Resources and Evaluation journal, 2022.

Contributions:

- Empirical evaluation of diverse cross-lingual strategies for low-resource AMR-to-Text generation on three approaches: Neural Machine Translation, Graph-to-Sequence, and pre-trained models.
- Manual analysis of the outputs produced by the different strategies, highlighting the main findings.

Comparison of Cross-lingual strategies for AMR-to-Brazilian Portuguese Generation

Marco Antonio Sobrevilla Cabezudo[1] · Rafael Torres Anchiêta[2] · Thiago Alexandre Salgueiro Pardo[1]

Received: date / Accepted: date

Abstract This work presents a study of different strategies for tackling low-resource AMR-to-text generation for Brazilian Portuguese on three approaches. In particular, we explore the use of English AMR as an interlingua and transfer learning (TL). The results suggest that using AMR as an interlingua can be a strong baseline. However, we need to consider the bilingual dictionary used in the concept translation, as it can harm text generation. On the other hand, TL seems to be a promising strategy for generating accurate outputs, but its contribution is not significant. Finally, we present a transformer-based model with the best performance, surpassing all baselines.

Keywords Abstract Meaning Representation · Natural Language Generation · Low-resource setting · Brazilian Portuguese

1 Introduction

Abstract Meaning Representation (AMR) is a semantic formalism that encodes the meaning of a sentence as a rooted, acyclic, labeled, and directed graph Banarescu et al (2013). It has been a relevant research topic in recent years with diverse applications such as semantic parsing (Flanigan et al, 2014; Xu et al, 2020; Bevilacqua et al, 2021), and automatic summarization (Vilca and Cabezudo, 2017; Inácio and Pardo, 2021), among others.

Concerning AMR-to-Text generation, diverse works and approaches have been applied for English (Pourdamghani et al, 2016; Beck et al, 2018; Ribeiro et al, 2020). However, for other languages, as far as we know, there are only the work of Sobrevilla Cabezudo et al (2019) (for Brazilian Portuguese) and, recently, the works of Fan and Gardent (2020) and Ribeiro et al (2021) that focus on multilingual text generation. However, the latter ones treat AMR as an interlingua. So, they explore

**Corresponding author: Marco Antonio Sobrevilla Cabezudo E-mail: msobrevillac@usp.br
1 Instituto de Ciências Matemáticas e de Computação, University of São Paulo, Brazil
2 Instituto Federal de Educação, Ciência e Tecnologia do Piauí, Brazil

multilingual AMR-to-text generation by using English AMR as input, disregarding some possible structural divergences.

One problem that constrains the research in other languages is getting high-quality corpora. Several efforts have been made to build AMR corpora for approaching this problem. Some works explore the potential of AMR as an interlingua to automatically build corpora for their language (Xue et al, 2014; Damonte and Cohen, 2018; Anchiêta and Pardo, 2018). Other works adopt the English AMR guidelines to their languages (Migueles-Abraira et al, 2018; Sobrevilla Cabezudo and Pardo, 2019) with some particular linguistic adequations. Nevertheless, the latter ones report small corpora. These smaller corpora prevent state-of-the-art methods from being replicated and/or achieving similar performance to those obtained in English, harming mainly neural models.

There are different applied strategies to deal with low-resource scenarios (Hedderich et al, 2021). A classic approach is cross-lingual projection. Its basic form consists of training a task-specific model in a high-resource language and, using parallel corpora, aligning unlabeled low-resource data to its equivalent in the high-resource language where labels can be obtained using a classifier. These labels can then be projected back to the corpus in the low-resource language (Yarowsky et al, 2001). Alternatively, methods can translate high-resource labeled datasets into low-resource ones by using an off-the-shelf machine translation system (Khalil et al, 2019).

Corpora translation seems to be a promising approach as we can get new corpora and then study particular linguistic phenomena. However, it has to face two limitations (at least): relying on the translation quality, which is not the same for all languages, and the possible domain-level divergence between the translated and the gold corpora, which can prevent leveraging all its potential. Concerning AMR, we also can suffer from some structural divergences in the translated AMR graphs (Sobrevilla Cabezudo et al, 2019; Wein and Schneider, 2021) as, originally, AMR is highly biased to English (Banarescu et al, 2013).

Another alternative is to leverage knowledge from a high-resource language/domain. For example, Transfer Learning (TL) aims to use knowledge from a task to improve the performance of a related one, reducing the amount of required training data (Torrey and Shavlik, 2010). TL has shown success in several tasks, being Neural Machine Translation (NMT) one of the most known, conveying knowledge from a high-resource language pair to low-resource language pair (Zoph et al, 2016). Similarly, it has been applied in domain adaptation with some success (Dethlefs, 2017).

With the advent of the pre-trained models (Radford et al, 2019; Raffel et al, 2020), Transfer Learning has become a hot topic in recent years. These models can be trained on large corpora of unannotated text, then fine-tuned for specific tasks on smaller amounts of supervised data, relying on the induced language model structure to facilitate generalization beyond the annotations. In particular, models such as BERT (Devlin et al, 2019), GPT-2 (Radford et al, 2019) or T5 (Raffel et al, 2020) have achieved state-of-the-art results in several NLP tasks such as Named-Entity Recognition, Sentiment Analysis, and even AMR-to-text generation for English (Mager et al, 2020; Ribeiro et al, 2020). However, as far as we know, its ability to deal with low-resource AMR-to-Text generation has not been proven.

In this work, we explore and compare some strategies for low-resource AMR-to-text generation for Brazilian Portuguese. Mainly, we try to answer the question *what is the best strategy for improving Brazilian Portuguese (BP) performance if we only have a similar corpus in a high-resource language (English)?* To do this, we explore the use of AMR as an interlingua, the translation of English AMR to Brazilian Portuguese (BP) and vice-versa, and transfer learning from an English AMR corpus to a gold BP AMR one on Sequence-to-Sequence and Graph-to-Sequence models.

In order to perform a broad study, we start disregarding the existence of pre-trained models such as GPT-2 or T5, aiming to suggest some directions for languages in which there are no pre-trained models available. After that, we evaluate the same experiments starting from a pre-trained model, achieving impressive results.

In general, our contributions are:

- We conduct a thorough empirical evaluation of diverse strategies for low-resource AMR-to-Text generation on three approaches: Sequence-to-Sequence, Graph-to-Sequence, and pre-trained models.
- We present a manual analysis of the outputs produced by the different strategies, highlighting the main findings.

2 Abstract Meaning Representation (AMR)

AMR aims to encode the meaning of a sentence in a directed, labeled, acyclic, and rooted graph (Banarescu et al, 2013). This representation includes (but not exclusively) semantic information related to semantic roles, named entities, spatial-temporal information, and co-references.

AMR has become an important research topic in the meaning/semantic representation field (Bos, 2016) and has been proven helpful in many NLP tasks like automatic text summarization (Liao et al, 2018; Inácio and Pardo, 2021), event detection (Li et al, 2015a) and machine translation (Song et al, 2019). Part of its success is based on its attempt to abstract away from syntactic idiosyncrasies and surface forms and its wide use of mature linguistic resources, such as PropBank (Palmer et al, 2005).

Figure 1 presents an example of an AMR graph for the sentence "*The girl adjusted the machine.*". It is worth noting that as AMR abstracts away the syntactic information, multiple possible sentences can correspond to this graph. This way, another possible sentences that represent the graph could be "*The girl made adjustments to the machine.*" and "*The machine was adjusted by the girl.*".

The current AMR-annotated corpus for English contains 59,255 instances[1]. For Non-English languages, there are some efforts to build corpora leveraging the alignments and existing parallel corpora by using AMR as an interlingua (Xue et al, 2014; Anchiêta and Pardo, 2018). On the other hand, some works adapt the AMR guidelines to their languages (Sobrevilla Cabezudo and Pardo, 2019). However, most corpora are far from presenting a size similar to the English one.

For Brazilian Portuguese, as far as we know, there are three AMR corpora, one focused on annotating the sentences of the "The Little Prince" book (Anchiêta and

[1] Available at `https://catalog.ldc.upenn.edu/LDC2020T02`.

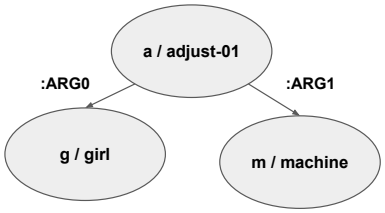


Fig. 1: AMR example for the sentence “The girl adjusted the machine.”

Pardo, 2018), another one that contains manually annotated news text sentences (Sobrevilla Cabezudo and Pardo, 2019), and one focused on opinative sentences (Inácio et al, 2022). Similarly to Banarescu et al (2013), some concepts of all corpora were annotated using Verbo-Brasil (Duran and Aluísio, 2015), a lexical resource analogous to PropBank (Palmer et al, 2005). About the corpora size, the first-mentioned corpus comprises 1,527 annotated sentences, the second one 870 sentences, and the last one 404 sentences.

3 Related Work

Several AMR-to-Text generation methods for English have been proposed in recent years. Methods based on Statistical Machine Translation (SMT) (Pourdamghani et al, 2016), and tree and graph-to-string transducers (Flanigan et al, 2016; Song et al, 2017) have been proposed at the beginning. In parallel, neural models like sequence-to-sequence (Neural Machine Translation or NMT) (Castro Ferreira et al, 2017; Konstas et al, 2017) and graph-to-sequence (Beck et al, 2018) have obtained comparable and even better results. Recently, pre-trained-based approaches (Mager et al, 2020) have emerged, beating all previous approaches.

As far as we know, Sobrevilla Cabezudo et al (2019) present the only work focused on AMR-to-Text generation for a Non-English language. In this work, the authors evaluate the helpfulness of translated AMR corpus when we only have a few hundred instances in the gold AMR corpus in the test set and no training set is given. The results prove its helpfulness; however, it highlights that there is bias to the translated corpus. In addition, another study shows that there are structural divergences between translated AMR and gold AMR, suggesting that translated corpus could serve as a starting point (Sobrevilla Cabezudo and Pardo, 2022).

Other non-English work (Fan and Gardent, 2020; Ribeiro et al, 2021) have tried to generate sentences in various languages from English AMR graphs, using it as an interlingua. Results are promising. However, this work does not consider specific linguistic phenomena as the previous ones.

4 Experimental Setup

4.1 Data

The AMR-to-Text generation task is evaluated on the news section of the Brazilian Portuguese (BP) AMR corpus released by Inácio et al (2022) (we will refer to this as *GOLD* dataset). This manually annotated corpus comprises 870 instances divided into 402, 224, and 244 instances for training, development, and testing, respectively.

In order to perform the cross-lingual study, we use an automatically generated AMR corpus for Portuguese (we will refer to this as *TRANSLATED* dataset). This corpus is generated by translating both AMR graphs and sentences from the English AMR corpus to BP and inheriting the alignments between node/edges and surface tokens[2] (Sobrevilla Cabezudo et al, 2019). Overall, this corpus comprises 18,219 and 1,027 instances for training and development, respectively. In addition, we use the original English version of this corpus for some experiments (we will refer to this as *ENGLISH* dataset).

4.2 Approaches

Initial experiments are performed using Sequence-to-Sequence (S2S) and Graph-to-Sequence (G2S) architectures to evaluate their behavior when we have just a few resources in different languages. After this, we explore using a pre-trained model under the same settings. Details about the hyperparameters may be found in Appendix A.

Sequence-to-Sequence S2S receives a flattened AMR version as input that can be obtained in various ways. For example, using the PENMAN notation (Matthiessen and Bateman, 1991) or applying the depth-first search (DFS) algorithm to generate the flattened version. On the other hand, Castro Ferreira et al (2017) explore some preprocessing strategies:

- Delexicalisation: that anonymises some entities;
- Compression: that determines which nodes and relations should be included in the flattened version[3];
- Preordering: that determines the order of the nodes and relations in the flattened graph[4].

In our experiments, we only use the preordering method proposed by Castro Ferreira et al (2017) as this setting produced the best results in previous experiments. Besides, we study machine translation based on neural models (Bahdanau et al, 2015) and subwords (byte-pair encoding) as input representations that are generated by using Bertimbau's vocabulary (Souza et al, 2020).

[2] In this work, we use the LDC2016E25 corpus to conduct the experiments.

[3] When no compression is applied, all nodes and relations are included in the flattened version.

[4] DFS is used when no preordering is applied.

Graph-to-Sequence (G2S) We use the approach proposed by Beck et al (2018) that applies Gated Graph Neural Network (GGNN) (Li et al, 2015b) for encoding AMR graphs, avoiding possible errors generated by the linearisation in sequence models.

This approach receives a Levi (AMR) Graph as input instead of an original one. A Levi graph is an unlabeled graph originated by the conversion of the labeled edges in a labeled graph into nodes linked to the actual nodes (Levi, 1942).

In experiments, we use subwords (byte-pair encoding) as input representation - in the same way as the NMT approach - and the lexicalised version of the output because this produced the best performance on the development set in previous experiments.

Transformer-based models Similar to Ribeiro et al (2020), we explore T5 (Raffel et al, 2020), which is a Transformer-based architecture (Vaswani et al, 2017) that uses a text-to-text approach.

Particularly, we use the HuggingFace implementation (Wolf et al, 2020) and the T5 base model for Brazilian Portuguese (Carmo et al, 2020)[5] as starting point and then we fine-tune on our AMR-to-Text datasets. The input for the model is the same one used in the NMT approach, however, in order to imitate the T5 setup, we add the prefix "*traduzir grafo a sentença:*" ("Translate graph into sentence:") before the flattened AMR graph.

4.3 English AMR-to-X

Initial experiments aim to use English AMR graphs as inputs and English (or BP) sentences as outputs during training, assuming that models can overcome structural differences (Damonte and Cohen, 2018). This way, we can leverage the knowledge from a bigger corpus without modifications to AMR graphs. Experiments are detailed in Figure 2.

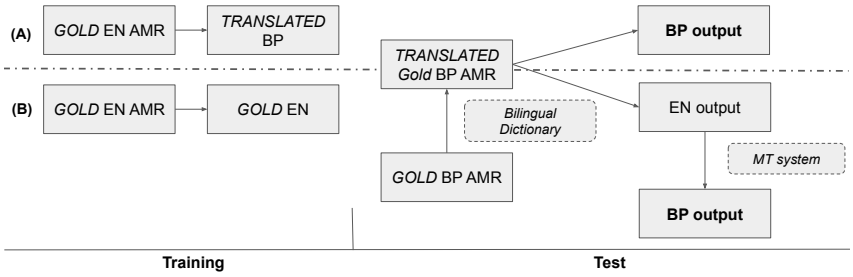


Fig. 2: English AMR-to-X strategies overview. (A) Training on the *SENT-TRANSLATED* corpus and testing on the translations of the BP AMR graphs. (B) Training on the *ENGLISH* corpus and testing in a similar way to (A); however, translating outputs into BP using an available Machine Translation model.

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

4.3.1 English AMR-to-BP

One of the AMR project statements is that AMR is not an interlingua and is strongly biased toward English. However, Damonte and Cohen (2018) explore cross-lingual AMR parsing strategies. Specifically, they exploit an AMR parser for English and parallel corpora to learn AMR parsers for several languages. The results show that it is possible to use AMR annotations for English as a semantic representation for sentences written in other languages. Furthermore, the generated parsers overcome structural differences between the languages.

Inspired by this cross-lingual work, we train models on a modified version of the *ENGLISH* corpus (named *SENT-TRANSLATED* corpus). This version contains the English AMR graphs on the source and the Portuguese translations on the target side[6] (Figure 2-A). For evaluation, we translate the nodes in the BP AMR graphs into their corresponding English ones. We use the bilingual dictionary provided by the MUSE project[7], however, we corrected some word translations. On the other hand, we translate BP frames using alignments between the Verbo-Brasil and the PropBank repository. Finally, if no frames are found, we use the bilingual dictionary to generate the translation.

Specifically, we explore S2S and G2S approaches. Models share the source and target representation, and we use subwords as input representations that are generated by using the Multilingual BERT's vocabulary[8].

4.3.2 English AMR-to-English

For AMR parsing, Uhrig et al (2021) propose a simple baseline that consists in translating the sentences into English and then projecting their AMR with a monolingual AMR parser. Results show a significant improvement in comparison with some State-of-the-Art models.

Motivated by this work, we explore a similar strategy that consists of *training* a AMR-to-Text generation model on the ENGLISH AMR corpus, and evaluating by (1) *translating GOLD* BP AMR graphs into their corresponding English version, (2) *generating* English outputs by using the trained model, and (3) *translating* the English outputs into their corresponding BP ones (Figure 2-B). The nodes in AMR graphs are translated similarly to the previous experiment, and the English to BP translation is performed by using MarianNMT (Junczys-Dowmunt et al, 2018).

We explore S2S, G2S, and Transformer-based approaches. Particularly, S2S and G2S models share the source and target representation, and we use subwords generated by using the BERT's vocabulary[9] as input representations.

[6] These translations are the same as the ones from the *TRANSLATED* corpus.

[7] Available at `https://github.com/facebookresearch/MUSE`.

[8] Vocabulary is available at `https://github.com/google-research/bert/blob/master/multilingual.md`.

[9] Vocabulary is available at `https://storage.googleapis.com/bert_models/2020_02_20/uncased_L-12_H-768_A-12.zip`.

4.4 Transfer Learning

As mentioned before, Transfer Learning aims to use knowledge from a task or domain - for which there is a large amount of available data - to improve the performance on a related one with a smaller amount of required training data (Torrey and Shavlik, 2010; Ruder, 2019).

Its application has been successful in different scenarios. For example, Zoph et al (2016) explore Neural Machine Translation in a low-resource language pair by re-utilising the learned knowledge from a high-resource language pair, producing improvements in the target language pair. On the other hand, Dethlefs (2017) explores how linguistic knowledge from a source domain in NLG, for which labeled data is available, can be adapted to a target domain by reusing training data across domains. The results show that learned representations can be transferred across domains, overcoming some lexical-syntactic divergences.

In our context, we observe a difference in text generation performance between *TRANSLATED* and *GOLD* corpora that structural AMR divergences can produce as we inherit English AMR structures. Besides, we translate sentences from English into Portuguese, opening the possibility of generating paraphrases. This way, we aim to extract as much as possible knowledge from the *TRANSLATED* corpus and leverage it into the *GOLD* corpus.

Firstly, we train a model on the *TRANSLATED* AMR corpus and, then, fine-tune it on the *GOLD*[10]. In addition, we explore to train models on both *TRANSLATED* and *GOLD* corpora together (named *MERGED* corpus). Previous experiments show that training on *MERGED* corpus produces improvements in the text generation on *GOLD* instances (Sobrevilla Cabezudo and Pardo, 2022). However, there is a bias to the *TRANSLATED* one due to the unbalancing (*TRANSLATED* corpus size is 40 times the size of the *GOLD* one). Details about the hyperparameters are found in the Appendix B.

4.5 Baseline

We use a Phrase-based Statistical Machine Translation system (SMT) (Koehn et al, 2003) as a baseline as this has demonstrated to be a strong baseline in the AMR-to-Text generation task. The training parameters are the same of Castro Ferreira et al (2017) and a 5-gram language model trained on the Brazilian Portuguese corpus provided by Hartmann et al (2017) by using KenLM (Heafield, 2011). Besides, we use Moses (Koehn et al, 2007) for training the translation models.

In addition to this baseline, we train models for the three approaches on the *TRANSLATED* and the *MERGED* corpora to measure the performance before the fine-tuning.

[10] Fine-tuning can be seen as the simplest way to perform Transfer Learning as this last one is an broader concept.

5 Results and Analysis

Table 1 shows the overall results for Sequence-to-Sequence (S2S), Graph-to-Sequence (G2S), and Transformer-based approaches (T5) in terms of BLEU (B) (Papineni et al, 2002), METEOR (M) (Lavie and Agarwal, 2007), chrF++ (C++) (Popović, 2017), and BERTScore (BS) (Zhang et al, 2020) evaluation metrics [11][12]. The Table contains the results of applying the two proposed strategies (English-to-X and Transfer Learning) on all approaches. In addition, we add the results obtained by a Statistical Machine Translation model (SMT) on *TRANSLATED*, *GOLD*, and *MERGED* corpora.

In general, results show that the SMT trained on the *MERGED* corpus beats all strategies and approaches applied when we do not regard the transformer-based approach. However, the difference decreases when transferring knowledge from the *MERGED* to the *GOLD* corpus ("*+ FT ON GOLD*" rows in Table 1), obtaining comparable results in terms of BERTScore ($\sim 0.80 - 0.81$).

In the case of the transformer-based model, we can see that T5's performance largely surpasses the result obtained by the SMT model when it is trained on the *TRANSLATED* corpus (~ 23.97 BLEU) and the *MERGED* corpora (~ 25.12 BLEU). Besides, the model trained on the *GOLD* corpus obtains comparable results with the SMT approach trained on the *MERGED* one.

Concerning the English-to-X experiments, we can see that training on the *ENGLISH* or the *SENT-TRANSLATED* corpora gets the worst performance in the *GOLD* test set for all metrics (except for T5). A possible explanation is that translations of BP AMR nodes can occasionally be associated with words that are not part of the vocabulary, are rare, or are incorrect, making the text generation harder and producing hallucinations.

On the other hand, in the case of the training on the *ENGLISH* corpus, we raise two possible additional issues: the first one is the need to translate English outputs produced by the generation model into BP as its quality depends on the quality of the text produced originally. The second one is related to the machine translation system because this can create paraphrases or make translation mistakes, producing quite different outputs from the expected ones. We expect this does not happen because the performance of MariaNMT is usually good for various language pairs. However, a manual revision is needed to confirm this.

It is worth noting that the performance of T5 trained on the *ENGLISH* corpus is similar to the obtained one by the SMT model trained on the *MERGED* corpus (which is the BP version of the *ENGLISH* one) in terms of BERTScore. However, the other metrics show a significant difference. It is expected because the SMT baseline does not use a BP-to-English step as T5 does (for *ENGLISH* corpus); thus, the outputs are more token-level accurate.

About the BERTScore values, we could say that the metric is more robust, overcoming any difference in terms of tokens. Nonetheless, another possibility is that

[11] We execute 4 runs for each experiment and show the mean and standard deviation.

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

BERTScore assigns high values even if outputs contain strange (less-related) translations because of the nature of the metric. Similar to the issues in previous paragraphs, it is necessary to perform a manual revision to find out the cause.

About the Transfer Learning strategy, we can see a performance gain in S2S and G2S approaches, mainly when the model is firstly trained on the *TRANSLATED* corpus. This is expected as there are divergences between both *TRANSLATED* and *GOLD* corpora and the *MERGED* corpus contains some *GOLD* instances. On the other hand, the best performance for S2S and G2S is similar, achieving comparable results with SMT in BERTScore ($\sim 0.80 - 0.81$). In the case of T5, we can see a not significant improvement, suggesting that the model learns stable representations.

Another point to note is that the main gain happens when evaluating BLEU. However, the other metrics do not show the same behavior, suggesting that the only difference between the parent model and the fine-tuned model is that the latter uses the exact words or n-grams as the reference ones.

Analysing the different strategies, we can see the same behavior as the previous approaches when the model is trained on *ENGLISH* corpora, i.e., lower values in BLEU, METEOR, and chrF++ but a higher value in the BERTScore metric. Besides, this strategy achieves similar results (BERTScore) to those obtained by SMT on *MERGED* corpus. It is expected as SMT tends to generate more accurate outputs.

Finally, we aim to know the best way to use the English AMR corpus: applying Transfer Learning or adding its *TRANSLATED* version to the GOLD? Comparing both approaches (training on *TRANSLATED* + fine-tune on *GOLD* vs training on *MERGED* corpus in Table 1), we can see different results. For example, S2S produces better results when Transfer Learning is applied, mainly for BLEU, in which the difference is around 2.2 points. However, there is no difference for T5, even using the input produced by the same preprocessing. In general, T5 is more robust as it was trained on more data; this way, we could expect that S2S performance changes if we pre-train the models on more data and then fine-tune on the *MERGED* corpus in a similar way to T5.

In the case of G2S, we can see that Transfer Learning is only a bit better than using the *MERGED* corpus for BLEU. However, METEOR and chrF++ produce a difference of almost 3 points in favor of the *MERGED* corpus. An explanation is that G2S better handles divergences between *TRANSLATED* and *GOLD* corpora but cannot realise outputs correctly. Furthermore, it is worth noting that G2S is trained on BPE-level inputs, so it is prone to produce some morphological concordance errors due to the small dataset. Therefore, we could suggest that increasing the number of instances could produce better performance in all metrics.

5.1 Ablation Study

In order to verify if the model trained on *TRANSLATED* and *MERGED* corpora are really overfitted and prevents learning on the *GOLD* corpus, we explore training on both corpora a different number of epochs and then apply the same fine-tuning process. Tables 3, 4 and 5 show the results for each approach (Appendix C).

	Training set	B	M	C++	BS
SMT	GOLD	6.05	0.24	0.43	0.76
	MERGED	14.69	0.34	0.50	0.82
S2S	ENGLISH	3.24 ±0.39	0.17	0.29	0.75
	SENT-TRANSLATED	2.45 ±0.52	0.16	0.24	0.74
	TRANSLATED	7.28 ±0.33	0.28	0.37	0.79
	+ FT on GOLD	10.04 ±0.28	0.29	0.38	0.80
	MERGED	7.80 ±0.77	0.27	0.37	0.79
	+ FT on GOLD	8.91 ±0.29	0.29	0.38	0.81
G2S	ENGLISH	3.63 ±0.68	0.18	0.29	0.76
	SENT-TRANSLATED	3.11 ±0.53	0.15	0.23	0.73
	TRANSLATED	5.75 ±0.69	0.24	0.33	0.78
	+ FT on GOLD	9.73 ±0.32	0.26	0.34	0.79
	MERGED	8.69 ±0.59	0.29	0.38	0.80
	+ FT on GOLD	10.17 ±0.25	0.30	0.39	0.81
T5	ENGLISH	10.73 ±0.54	0.31	0.44	0.83
	GOLD	12.52 ±0.85	0.31	0.42	0.82
	TRANSLATED	23.97 ±0.58	0.46	0.57	0.87
	+ FT on GOLD	24.39 ±0.69	0.46	0.57	0.87
	MERGED	25.12 ±0.33	0.47	0.58	0.87
	+ FT on GOLD	25.66 ±0.13	0.47	0.57	0.87

Table 1: S2S, G2S, and T5 performance on the test set. English-to-X strategy comprises models trained on *ENGLISH* or *SENT-TRANSLATED* corpora. Transfer Learning results start with "+ FT on X", which means that the model is trained on *TRANSLATED* or *MERGED* corpora and then fine-tune on the *GOLD* corpus. Standard Deviation for METEOR, chrF++, and BERTScore are omitted because their range is constant (0.00-0.02)

Results show that S2S and G2S approaches do not need more epochs than 15 and 17, respectively; instead, few ones can help obtain similar results in some cases. Conversely, in the case of the transformer-based model, we can see that results are similar when the model is trained on the *TRANSLATED* corpus. However, the performance increases when the model is trained on the *MERGED* corpus by a few more epochs than 3 but then seems to stop improving and then decreases, overfitting to the corpora.

5.2 Manual Revision

Aiming to understand some results, we conduct a manual revision. We select 65 instances from the *GOLD* test set and verify what the main mistakes or phenomena that generators produce are.

We define 6 categories for evaluating the English-to-X strategy: (1) total hallucinations, when the output is different from the reference, (2) partial hallucination, when the output contains the main point of the reference, but some tokens are not part of the reference, (3) concept translation errors, when a BP AMR concept is translated into a non-related English one, (4) Machine Translation error, when the English output is valid but not the BP one, (5) paraphrase, when the output is a paraphrase of the reference, and (6) valid, when output is acceptable and accurate to the reference. Besides, we only evaluate the training on the *ENGLISH* corpus and the outputs of the best approach (T5) as the other approaches present inferior results.

	Concept Translation	
	Weird/Incorrect	Adequate
	31	34
Valid	3	19
⇒ Paraphrase	3	9
Partial Hallucination	10	4*
Hallucination	18	11
BERTScore	0.80	0.85

Table 2: Distribution of the categories analised for instances with weird/incorrect concept translations and adequate concept translation. Outputs are generated by the T5 model trained on *ENGLISH* corpus. * We note that 2 of these 4 partial hallucinations correspond to problems in the Machine Translation system, i.e., the English output is correct but the translation is incorrect. In addition, the problem is associated to Named-entity translation.

Table 2 presents the results. As it can be seen, the main problem is the error in concept translation, i.e., the bilingual dictionary used for the translation. It produces problems like the one shown in Figure 3. Although the T5's output can be valid (possible paraphrase), there are some cases (18) in which the output generated is a complete hallucination.

Among other results, we note that, as we expected, the model produces paraphrases in most cases (3 for incorrect and 9 for correct concept translations). Also, when AMR graphs contain incorrect/weird concept translations, the model is prone to produce more hallucinations. At last, about the high BERTScore value obtained in experiments, we can see that, even when we have several hallucinations (weird/incorrect concept translation column), the BERTScore is 0.80. It can suggest that due to its nature, BERTScore scores high to adequate outputs and hallucinations that can share some relation with the reference.

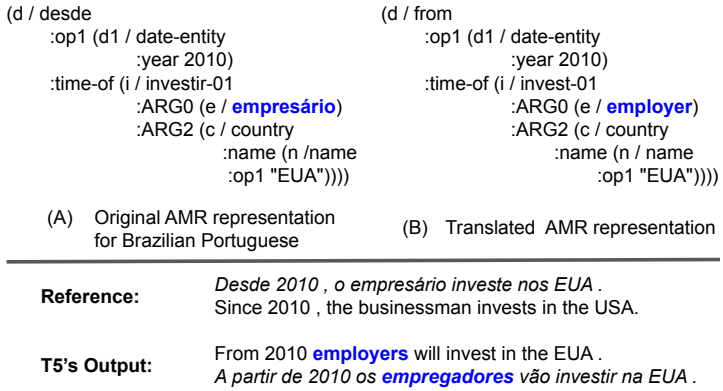


Fig. 3: Example of error because the word "*empresário*" (businessperson) was changed by employer in the AMR graph."

In the case of transfer learning, we only regard 4 categories: total and partial hallucinations, paraphrases, and valid outputs. Besides, we are more flexible in identifying a label; thus, we assume that there are many morphological concordance errors for diverse approaches and try to avoid their influence on the annotation.

In general, for S2S and G2S, we find out that the performance does not significantly improve when fine-tuning is applied after training on the *TRANSLATE* corpus. This way, a better option seems to be to use the *MERGED* corpus and, optionally, then fine-tune on the *GOLD* corpus as these experiments decrease the number of total hallucinations, generating partial hallucinations and valid outputs.

Finally, in the case of T5, we can see a slight improvement when fine-tuning is applied after training on the *TRANSLATED* or the *MERGED* corpora (number of valid outputs increases). However, comparing the strategy of training on the *TRANSLATED* corpus + fine-tune on the *GOLD* vs. training on the *MERGED* corpus, we find that the latter is prone to produce more accurate outputs and even more partial hallucinations in the worst case. Figure 4 shows the outputs for each model in the transfer learning setting.

Reference: *mas não da forma que eles estão fazendo .* — but not in the way they are doing.

	pre-training	**fine-tuning**
S2S - TRANSLATED	*não é assim , mas eles fizeram . . .* — not so , but they did . . .	*mas não é de forma .* — but it is in no way.
S2S - MERGED	*não é uma forma de forma que eles fazem .* — it's not a way the way they do.	*mas não é de forma que eles fazem .* — but that's not how they do it.
G2S - TRANSLATED	*não é uma forma que eles fizeram .* — it's not a way they did it.	*não é uma forma que eles fazer .* — it's not a way they do it.
G2S - MERGED	*mas não há forma que eles fizeram .* — but there is no way they did.	*mas não de fazer .* — but not to do
T5 - TRANSLATED	*não é de forma alguma que eles fazem .* — it is not at all what they do.	*mas não da forma que eles fizeram .* — but not the way they did.
T5 - MERGED	*não da forma que eles fizeram .* — not the way they did.	*não há forma de fazer isso .* — there is no way to do this.

Fig. 4: Outputs for each model when only the pre-training is performed and when the model is fine-tuned on the **GOLD** corpus.

6 Conclusion and future work

This work presented a study of different strategies for tackling low-resource AMR-to-text generation for Brazilian Portuguese on S2S and G2S approaches. In particular, we explore the use of English AMR as an interlingua and transfer learning from a *TRANSLATED* corpus into a *GOLD* one. In addition, we fine-tune a transformer-based pre-trained model, T5, on the same corpora and explore the same strategies. Finally, we manually revised the outputs, obtaining some insights.

Results suggest that using AMR as an interlingua is a strong baseline, as highlighted by work in semantic parsing. However, we need to consider the bilingual dictionary used in the concept translation, as it can produce non-sense concepts, generating hallucinations. On the other hand, transfer learning (TL) seems to be a good strategy to generate outputs similar to the references (in terms of n-grams); nonetheless, a semantic metric suggests that there is no difference between using (or not) TL. Finally, about the transformer-based model, we could see that it obtains the best performance, surpassing all the other approaches, even the baseline. However, TL does not seem to significantly contribute in this case.

As future work, we plan to explore data augmentation strategies, starting from the *TRANSLATED* corpus, as we have demonstrated its helpfulness in the text generation task.

References

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

Bahdanau D, Cho K, Bengio Y (2015) Neural machine translation by jointly learning to align and translate. In: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings

Banarescu L, Bonial C, Cai S, Georgescu M, Griffitt K, Hermjakob U, Knight K, Koehn P, Palmer M, Schneider N (2013) Abstract meaning representation for sembanking. In: Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, Association for Computational Linguistics, Sofia, Bulgaria, pp 178–186

Beck D, Haffari G, Cohn T (2018) Graph-to-sequence learning using gated graph neural networks. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, pp 273–283

Bevilacqua M, Blloshmi R, Navigli R (2021) One spring to rule them both: Symmetric amr semantic parsing and generation without a complex pipeline. Proceedings of the AAAI Conference on Artificial Intelligence 35(14):12,564–12,573

Bos J (2016) Squib: Expressive power of Abstract Meaning Representations. Computational Linguistics 42(3):527–535

Carmo D, Piau M, Campiotti I, Nogueira R, Lotufo R (2020) Ptt5: Pretraining and validating the t5 model on brazilian portuguese data. arXiv preprint arXiv:200809144

Castro Ferreira T, Calixto I, Wubben S, Krahmer E (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, Association for Computational Linguistics, Santiago de Compostela, Spain, pp 1–10

Damonte M, Cohen SB (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), Association for Computational Linguistics, New Orleans, Louisiana, pp 1146–1155

Dethlefs N (2017) Domain transfer for deep natural language generation from abstract meaning representations. IEEE Computational Intelligence Magazine 12(3):18–28

Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. 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), Association for Computational Linguistics, Minneapolis, Minnesota, pp 4171–4186

Duran MS, Aluísio SM (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, Association for Computational Linguistics, Denver, Colorado, pp 216–221

Fan A, Gardent C (2020) Multilingual AMR-to-text generation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, pp 2889–2901

Flanigan J, Thomson S, Carbonell J, Dyer C, Smith NA (2014) A discriminative graph-based parser for the Abstract Meaning Representation. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Baltimore, Maryland, pp 1426–1436

Flanigan J, Dyer C, Smith NA, Carbonell J (2016) Generation from abstract meaning representation using tree transducers. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, San Diego, California, pp 731–739

Hartmann N, Fonseca E, Shulby C, Treviso M, Silva J, Aluísio S (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, Sociedade Brasileira de Computação, Uberlândia, Brazil, pp 122–131

Heafield K (2011) KenLM: Faster and smaller language model queries. In: Proceedings of the Sixth Workshop on Statistical Machine Translation, Association for

Computational Linguistics, Edinburgh, Scotland, pp 187–197
Hedderich MA, Lange L, Adel H, Strötgen J, Klakow D (2021) A survey on recent approaches for natural language processing in low-resource scenarios. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, pp 2545–2568
Hieber F, Domhan T, Denkowski MJ, Vilar D, Sokolov A, Clifton A, Post M (2017) Sockeye: A toolkit for neural machine translation. ArXiv abs/1712.05690
Inácio M, Pardo T (2021) Semantic-based opinion summarization. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), INCOMA Ltd., Held Online, pp 619–628
Inácio ML, Cabezudo MAS, Ramisch R, Di Felippo A, Pardo TAS (2022) The amr-pt corpus and the semantic annotation of challenging sentences from journalistic and opinion texts. SciELO Preprints
Junczys-Dowmunt M, Grundkiewicz R, Dwojak T, Hoang H, Heafield K, Neckermann T, Seide F, Germann U, Aji AF, Bogoychev N, Martins AFT, Birch A (2018) Marian: Fast neural machine translation in C++. In: Proceedings of ACL 2018, System Demonstrations, Association for Computational Linguistics, Melbourne, Australia, pp 116–121
Khalil T, Kiełczewski K, Chouliaras GC, Keldibek A, Versteegh M (2019) Cross-lingual intent classification in a low resource industrial setting. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, pp 6419–6424
Koehn P, Och FJ, Marcu D (2003) Statistical phrase-based translation. In: Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, pp 127–133
Koehn P, Hoang H, Birch A, Callison-Burch C, Federico M, Bertoldi N, Cowan B, Shen W, Moran C, Zens R, Dyer C, Bojar O, Constantin A, Herbst E (2007) Moses: Open source toolkit for statistical machine translation. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, Association for Computational Linguistics, Prague, Czech Republic, pp 177–180
Konstas I, Iyer S, Yatskar M, Choi Y, Zettlemoyer L (2017) Neural AMR: Sequence-to-sequence models for parsing and generation. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vancouver, Canada, pp 146–157
Lavie A, Agarwal A (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, Association for Computational Linguistics, Prague, Czech Republic, pp 228–231
Levi FW (1942) Finite geometrical systems
Li X, Nguyen TH, Cao K, Grishman R (2015a) Improving event detection with abstract meaning representation. In: Proceedings of the First Workshop on Comput-

ing News Storylines, Association for Computational Linguistics, Beijing, China, pp 11–15

Li Y, Tarlow D, Brockschmidt M, Zemel R (2015b) Gated graph sequence neural networks. arXiv preprint arXiv:151105493

Liao K, Lebanoff L, Liu F (2018) Abstract meaning representation for multi-document summarization. In: Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics, Santa Fe, New Mexico, USA, pp 1178–1190

Luong T, Pham H, Manning CD (2015) Effective approaches to attention-based neural machine translation. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Lisbon, Portugal, pp 1412–1421

Mager M, Fernandez Astudillo R, Naseem T, Sultan MA, Lee YS, Florian R, Roukos S (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, Association for Computational Linguistics, Online, pp 1846–1852

Matthiessen C, Bateman JA (1991) Text Generation and Systemic-Functional Linguistics: Experiences from English and Japanese. Pinter Publishers

Migueles-Abraira N, Agerri R, Diaz de Ilarraza A (2018) Annotating abstract meaning representations for Spanish. In: Proceedings of the 11th International Conference on Language Resources and Evaluation, European Languages Resources Association, Miyazaki, Japan, pp 3074–3078

Palmer M, Gildea D, Kingsbury P (2005) The proposition bank: An annotated corpus of semantic roles. Computational Linguistics 31(1):71–106

Papineni K, Roukos S, Ward T, Zhu WJ (2002) Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pp 311–318

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

Pourdamghani N, Knight K, Hermjakob U (2016) Generating English from abstract meaning representations. In: Proceedings of the 9th International Natural Language Generation conference, Association for Computational Linguistics, Edinburgh, UK, pp 21–25

Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I (2019) Language models are unsupervised multitask learners

Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu PJ (2020) Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research 21(140):1–67

Ribeiro LFR, Schmitt M, Schütze H, Gurevych I (2020) Investigating pretrained language models for graph-to-text generation. CoRR abs/2007.08426, 2007.08426

Ribeiro LFR, Pfeiffer J, Zhang Y, Gurevych I (2021) Smelting gold and silver for improved multilingual AMR-to-Text generation. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pp 742–

750
Ruder S (2019) Neural transfer learning for natural language processing. PhD thesis, NUI Galway
Sobrevilla Cabezudo MA, Pardo T (2019) Towards a general abstract meaning representation corpus for Brazilian Portuguese. In: Proceedings of the 13th Linguistic Annotation Workshop, Association for Computational Linguistics, Florence, Italy, pp 236–244
Sobrevilla Cabezudo MA, Pardo TA (2022) Low-resource amr-to-text generation: A study on brazilian portuguese. Procesamiento del Lenguaje Natural
Sobrevilla Cabezudo MA, Mille S, Pardo T (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), Association for Computational Linguistics, Hong Kong, China, pp 94–103
Song L, Peng X, Zhang Y, Wang Z, Gildea D (2017) AMR-to-text generation with synchronous node replacement grammar. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Vancouver, Canada, pp 7–13
Song L, Gildea D, Zhang Y, Wang Z, Su J (2019) Semantic neural machine translation using amr. Transactions of the Association for Computational Linguistics 7:19–31
Souza F, Nogueira R, Lotufo R (2020) BERTimbau: pretrained BERT models for Brazilian Portuguese. In: 9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23
Torrey L, Shavlik J (2010) Transfer learning. In: Handbook of research on machine learning applications and trends: algorithms, methods, and techniques, IGI Global, pp 242–264
Uhrig S, Garcia Y, Opitz J, Frank A (2021) Translate, then parse! a strong baseline for cross-lingual AMR parsing. In: Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies (IWPT 2021), Association for Computational Linguistics, Online, pp 58–64
Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA, NIPS'17, p 6000–6010
Vilca GCV, Cabezudo MAS (2017) A study of abstractive summarization using semantic representations and discourse level information. In: Text, Speech, and Dialogue, Springer-Verlag, pp 482–490
Wein S, Schneider N (2021) Classifying divergences in cross-lingual AMR pairs. In: Proceedings of The Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop, Association for Computational Linguistics, Punta Cana, Dominican Republic, pp 56–65
Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf R, Funtowicz M, Davison J, Shleifer S, von Platen P, Ma C, Jernite Y, Plu J, Xu C, Le Scao T, Gugger S, Drame M, Lhoest Q, Rush A (2020) Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference

on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, pp 38–45

Xu D, Li J, Zhu M, Zhang M, Zhou G (2020) Improving AMR parsing with sequence-to-sequence pre-training. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, pp 2501–2511

Xue N, Bojar O, Hajič J, Palmer M, Urešová Z, Zhang X (2014) Not an interlingua, but close: Comparison of English AMRs to Chinese and Czech. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), European Language Resources Association (ELRA), Reykjavik, Iceland, pp 1765–1772

Yarowsky D, Ngai G, Wicentowski R (2001) Inducing multilingual text analysis tools via robust projection across aligned corpora. In: Proceedings of the First International Conference on Human Language Technology Research

Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y (2020) Bertscore: Evaluating text generation with bert. In: International Conference on Learning Representations, URL https://openreview.net/forum?id=SkeHuCVFDr

Zoph B, Yuret D, May J, Knight K (2016) Transfer learning for low-resource neural machine translation. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Austin, Texas, pp 1568–1575

A Model Hyperparameters

Our Sequence-to-Sequence and Graph-to-Sequence model are based on Sockeye[13] (Hieber et al, 2017). It is worth noting that parameters that are not detailed in this section are defined by Sockeye. We describe them here for reproducibility:

A.1 Sequence-to-Sequence (S2S)

- The encoder and the decoder are a 1-layer RNN, and a 2-layers RNN with LSTM, each with a 512D hidden unit.
- RNN decoder uses bilinear attention (Luong et al, 2015).
- Vocabulary shared between source and target and weight tying between source, target and output layers.
- Source and target embeddings are 512D.
- Maximum sequence length in the decoder is 80
- We use a 0.25 dropout in source embeddings.
- We use Adam optimizer with 0.0003 as the initial learning rate.
- Learning rate is halved every time DEV perplexity does not improve for 3 checkpoints/epochs.
- We use mini-batches of size 16.
- Early stopping is used based on perplexity scores. Training stops if a model does not improve on the DEV set for more than 8 checkpoings/epochs.

[13] https://github.com/beckdaniel/sockeye/

A.2 Graph-to-Sequence

- The number of layers in the encoder is 8.
- All dimensionalities are fixed at 512D except for the encoder (576D).
- The remaining parameters (for training and the decoder) are the same as the NMT approach.

A.3 Transformer-based model

- We use mini-batches of size 8 and gradient accumulation of 4.
- Maximum sequence length in the decoder is 80.
- We set the maximum number of epochs as 12.
- We use AdamW optimizer and a learning rate of 0.0005.
- Early stopping is used based on perplexity scores. Training stops if a model does not improve on the DEV set for more than 4 epochs.

B Fine-tuning Hyperparameters

In general, we use the same hyperparameters as the parent models. However, we explore to modify some of them obtaining the best results.

B.1 S2S and G2S

- We use a 0.30 dropout in source and target embeddings and the hidden layers of the RNN decoder.
- We use Adam optimizer and a learning rate of 0.00003.
- Learning rate is halved every time DEV perplexity does not improve for 2 checkpoints/epochs.
- Early stopping is used based on perplexity scores. Training stops if a model does not improve on the DEV set for more than 5 checkpoings/epochs.

B.2 Transformer-based model

- We set the maximum number of epochs as 10.
- We use AdamW optimizer and a learning rate of 0.00005.
- Early stopping is used based on perplexity scores. Training stops if a model does not improve on the DEV set for more than 3 epochs.

C Ablation Study

This section present the ablation study. Tables 3, 4 and 5 show the results for S2S, G2S, and Transformer-based approaches on *TRANSLATED* (T) and *MERGED* (M) corpora in terms of BLEU (B) (Papineni et al, 2002), METEOR (M) (Lavie and Agarwal, 2007), chrF++ (C++) (Popović, 2017), and BERTScore (Zhang et al, 2020) evaluation metrics.

Set	Epoch	DEV				TEST			
		B	M	C++	BS	B	M	C++	BS
T	5	1.84	0.15	0.21	0.74	1.82	0.14	0.20	0.74
	8	6.75	0.23	0.31	0.77	5.51	0.23	0.30	0.78
	11	7.18	0.26	0.35	0.79	7.46	0.26	0.35	0.79
	14	8.35	0.27	0.37	0.79	8.81	0.28	0.37	0.80
	17	9.24	0.27	0.37	0.80	9.45	0.29	0.38	0.80
	20	8.65	0.28	0.38	0.80	9.65	0.28	0.38	0.80
	Best (15)	7.95	0.27	0.37	0.80	10.04	0.29	0.38	0.80
M	5	2.62	0.16	0.22	0.74	3.03	0.15	0.21	0.75
	8	7.26	0.25	0.33	0.78	6.81	0.25	0.33	0.78
	11	8.59	0.29	0.38	0.80	8.47	0.27	0.36	0.80
	14	9.33	0.30	0.39	0.80	8.88	0.29	0.39	0.81
	17	8.53	0.30	0.40	0.80	8.65	0.29	0.39	0.81
	20	8.95	0.30	0.41	0.81	9.60	0.29	0.39	0.81
	Best (13)	9.24	0.29	0.39	0.80	8.91	0.29	0.38	0.81

Table 3: S2S performance on the *GOLD* corpus after a number of epochs in pre-training

Set	Epoch	DEV				TEST			
		B	M	C++	BS	B	M	C++	BS
T	5	2.00	0.12	0.16	0.73	0.41	0.10	0.15	0.73
	8	5.41	0.18	0.24	0.76	3.53	0.17	0.22	0.76
	11	6.72	0.23	0.30	0.78	6.43	0.21	0.28	0.78
	14	7.91	0.25	0.32	0.79	9.01	0.25	0.32	0.79
	17	8.51	0.26	0.34	0.79	9.73	0.26	0.34	0.79
	20	7.16	0.26	0.34	0.79	8.84	0.26	0.34	0.79
	Best (17)	8.51	0.26	0.34	0.79	9.73	0.26	0.34	0.79
M	5	4.59	0.18	0.22	0.75	2.26	0.15	0.21	0.75
	7	6.43	0.25	0.32	0.78	6.15	0.23	0.31	0.78
	11	8.65	0.28	0.36	0.79	9.01	0.28	0.37	0.80
	14	8.73	0.29	0.38	0.80	8.91	0.28	0.38	0.80
	17	9.19	0.30	0.40	0.80	10.17	0.30	0.39	0.81
	20	8.94	0.30	0.40	0.80	9.63	0.30	0.40	0.80
	Best (17)	9.19	0.30	0.40	0.80	10.17	0.30	0.39	0.81

Table 4: G2S performance on the *GOLD* corpus after a number of epochs in pre-training

Set	Epoch	DEV				TEST			
		B	M	C++	BS	B	M	C++	BS
T	1	20.17	0.44	0.54	0.86	20.18	0.43	0.54	0.86
	3	25.56	0.48	0.57	0.87	24.15	0.47	0.57	0.87
	5	23.76	0.46	0.57	0.86	23.53	0.47	0.56	0.87
	7	25.47	0.48	0.58	0.86	23.56	0.46	0.56	0.87
	Best (4)	23.94	0.47	0.57	0.87	24.39	0.46	0.57	0.87
M	1	22.96	0.47	0.56	0.87	22.70	0.45	0.55	0.87
	3	27.27	0.50	0.59	0.87	25.66	0.47	0.57	0.87
	5	28.70	0.50	0.59	0.87	28.46	0.48	0.58	0.88
	7	25.33	0.49	0.58	0.87	27.59	0.50	0.59	0.88
	Best (3)	27.27	0.50	0.59	0.87	25.66	0.47	0.57	0.87

Table 5: T5 performance on the *GOLD* corpus after a number of epochs in pre-training

CHAPTER

6

KNOWLEDGE-LEVERAGING APPROACHES

This chapter presents works that aim to introduce/leverage knowledge from other resources or tasks. This way, the works aim to answer the following research questions:

- *What is the best strategy for dealing with data sparsity in AMR-to-text generation?*
- *How does data augmentation methods behave on AMR-to-text generation in low-resource settings and what is the best way to augment data?*

The chapter is divided in two sections. The first section brings a paper about the use of a pipeline approach for tackling the low-resource AMR-to-Text generation task. Finally, the second section brings a paper that focus on evaluating the helpfulness of paraphrase for improving the text generation task.

6.1 Exploring a POS-based Two-stage Approach for Improving Low-Resource AMR-to-Text Generation

This section comprehends the paper below.

CABEZUDO, M. A. S.; PARDO, T. Exploring a POS-based Two-stage Approach for Improving Low-Resource AMR-to-Text Generation, accepted at the Generation, Evaluation and Metrics workshop (GEM) at Empirical Methods in Natural Language Processing, 2022.

Contributions:

- A simple two-stage method that consists of generating masked surface realization and infilling the masked tokens with a transformer-based architecture, leveraging the capabilities of the transformer architecture for filling masked tokens.
- Manual revision on the outputs of the best approaches and the end-to-end approach.

Exploring a POS-based Two-stage Approach for Improving Low-Resource AMR-to-Text Generation

Marco Antonio Sobrevilla Cabezudo and **Thiago Alexandre Salgueiro Pardo**
Interinstitutional Center for Computational Linguistics (NILC)
Institute of Mathematical and Computer Sciences, University of São Paulo
São Carlos/SP, Brazil
msobrevillac@usp.br, taspardo@icmc.usp.br

Abstract

This work presents a two-stage approach for tackling low-resource AMR-to-text generation for Brazilian Portuguese. Our approach consists of (1) generating a masked surface realization in which some tokens are masked according to its Part-of-Speech class and (2) infilling the masked tokens according to the AMR graph and the previous masked surface realization. Results show a slight improvement over the baseline, mainly in BLEU (1.63) and METEOR (0.02) scores. Moreover, we evaluate the pipeline components separately, showing that the bottleneck of the pipeline is the masked surface realization. Finally, the human revision suggests that models still suffer from hallucinations, and some strategies to deal with the problems found are proposed.

1 Introduction

Abstract Meaning Representation (AMR) is a semantic formalism that encodes the meaning of a sentence into a labeled directed and rooted graph (Banarescu et al., 2013). This representation comprises semantic information related to semantic roles, named entities, and co-references, among others.

AMR is a widely-studied research topic in the semantic representation field and has been proven helpful in many Natural Language Processing tasks (Liao et al., 2018; Song et al., 2019). Its success is partially based on its broad use of mature linguistic resources, such as PropBank (Palmer et al., 2005), and its attempt to abstract away from syntax. Figure 1 shows an example of an AMR graph and its corresponding PENMAN notation for the sentence "*The boy must not go.*".

AMR-to-text generation aims to "translate" an Abstract Meaning Representation graph into its corresponding text. This task has been widely tackled by diverse approaches, starting from statistical, transducer-based and transition-based ones (Pourdamghani et al., 2016; Flanigan et al., 2016; Lampouras and Vlachos, 2017), until end-to-end neural ones (Mager et al., 2020; Ribeiro et al., 2021a), recently.

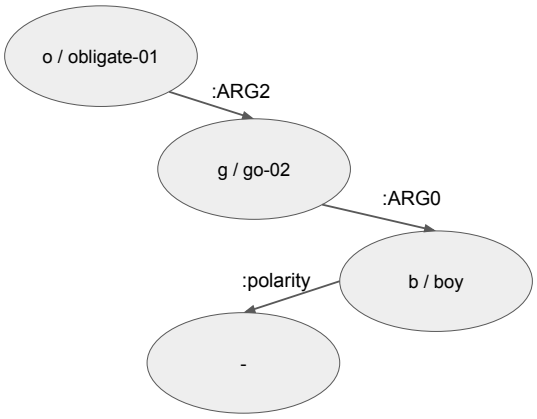


Figure 1: AMR example for the sentence "The boy must not go."

In particular, end-to-end neural models -mainly those based on pre-trained models- have largely outperformed the initial methods, achieving state-of-the-art results (Ribeiro et al., 2021b). These models can generate fluent text. However, they are prone to generate hallucinations, i.e., texts that are irrelevant to or contradicted with the input (Reiter, 2018).

Another drawback is that these models are usually data-hungry, i.e., they need to be trained on a large dataset to achieve a good performance. It can be a problem when we deal with low-resource domains, languages, or tasks (Sobrevilla Cabezudo and Pardo, 2022). Even when the results may be better than those obtained by statistical methods,

they are still far from good results. For example, Ribeiro et al. (2021b) show that fine-tuning T5 (Raffel et al., 2020) on a small portion of a big dataset (∼500 instances) produces a ∼10-15 BLEU score.

In general, an approach to have more control over the decoding process (and avoid hallucinations) is to use a pipeline-based method in which the model of each pipeline's module is implemented with neural models (Castro Ferreira et al., 2019; Ma et al., 2019; Puduppully and Lapata, 2021). Another alternative is to use templates, infill concepts in these templates, and then define a strategy to transform them into sentences/paragraphs (Kasner and Dušek, 2020; Mota et al., 2020). Both approaches have proven to be helpful in text generation tasks. However, the main issue for the latter one is that it only can be applied in restricted domains as it is necessary to define a set of templates.

In this work, we approach the AMR-to-text generation task in two stages. Firstly, generating a masked surface realization in which some tokens are masked according to its Part-of-Speech (POS) classes. Then, finally, infilling the masked tokens according to the AMR graph and the previous masked surface realization[1].

The intuition for masking some tokens this way is that some POS classes are more difficult to be predicted during text decoding and can harm the performance. On the other hand, filling-in-the-blank is commonly used on current SotA architectures, such as T5 (Raffel et al., 2020) during the pre-training phase. This way, we can leverage the learned knowledge to infill the masked tokens in the previous stage adequately.

Experiments are conducted on low-resource an AMR-to-text generation task for Brazilian Portuguese (Inácio et al., 2022) to show how this method behaves even when a large dataset is unavailable.

In general, our main contributions are:

- we propose a simple two-stage method that consists of generating masked surface realization and infilling the masked tokens with a transformer-based architecture;
- we conduct a manual revision on the outputs of the best approaches and the end-to-end approach.

[1]The code is available at https://github.com/msobrevillac/DICO-AMR2Text.

2 Related Work

AMR-to-Text generation Modular approaches have been mainly focused on converting AMR graphs into syntax trees via transition-based methods (Lampouras and Vlachos, 2017), end-to-end methods (Cao and Clark, 2019) or rule-based graph-transducers (Mille et al., 2017) and use an off-the-shelf method (neural or statistical) to generate the text. These methods usually have got a low performance on test sets (May and Priyadarshi, 2017).

AMR is more open-ended than other datasets such as WebNLG (Gardent et al., 2017). This way, extracting templates can be a complex task. Some attempts to get templates in the form of rules are presented by Flanigan et al. (2016) and Song et al. (2017). However, these approaches need some manually created rules and have been surpassed by current models.

On the other hand, current neural models have achieved SotA results. However, they need a large dataset to get high performance. On the contrary, a small portion of an extensive dataset produces lower scores (Ribeiro et al., 2021a,b)[2].

Data-to-Text generation Currently, most data-to-text methods are based on end-to-end neural approaches. In particular, methods that fine-tunes a pre-trained model, such as BART (Lewis et al., 2020) or T5 (Raffel et al., 2020), on its specific generation task have achieved SotA results.

Other works have tried to approach this kind of tasks using pipeline approaches (Castro Ferreira et al., 2019; Ma et al., 2019; Puduppully and Lapata, 2021) and template-based approaches (Kasner and Dušek, 2020; Mota et al., 2020). In particular, pipeline approaches have advantages in low-resource settings and unseen domains. On the other hand, template-based approaches tend to infill the templates with concepts and then use them to generate the complete sentence.

3 Experimental Setup

3.1 Dataset

We conduct all experiments on the journalistic section of the AMR-PT corpus (Inácio et al., 2022) (named AMRNews)[3]. The AMRNews corpus comprises 870 sentences with up to 23 tokens each from

[2]Ribeiro et al. (2021b) show an impressive improvement using structural adapters. However, this is not part of this study.

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

Brazilian news texts manually annotated according to adapted AMR guidelines (Sobrevilla Cabezudo and Pardo, 2019). Besides, it is divided into 402, 224, and 244 instances for training, development, and testing, respectively.

3.2 Architecture

Aiming to leverage the "fill-in-the-blank" potential of current pre-trained neural models, we propose a two-stage approach consisting of generating a masked surface realization and then infilling the masked tokens using a pre-trained model. Figure 2 shows an example of the whole process.

3.2.1 Masked Surface Realisation

The first stage involves generating a sentence corresponding to an AMR graph in which some tokens are masked. The idea behind this is that some tokens can be more difficult to be predicted. This way, we can mask them and let the next stage complete the masked tokens.

To decide what tokens should be masked, we use Part-of-Speech-based criteria. This way, we group all Part-of-Speech (POS) classes into main classes according to their function. For example, pronouns, nouns, and proper nouns are usually actors/places in a sentence, while verbs represent relations. This way, the main classes are: "*substantivos*" (nouns), "*verbos*" (verbs), "*qualificadores*" (qualifiers), and "*outros*" (others). Table 1 shows the main and POS classes included in each.

Main Class	**Part-of-Speech**
Substantivos (nouns)	pronoun, noun, proper noun
Verbos (verbs)	auxiliary verb, verb
Qualificadores (qualifiers)	adverb, adjective
Outros (others)	other Part-of-Speech

Table 1: Main and POS classes used in experiments

We train a model for each main class separately. Besides, we train a model for all main classes together. The input consists of a prefix and an AMR graph in the PENMAN notation (eliminating the frameset numbers). We use the expression "*mascarar **X** desde amr:*" ("Mask **X** from amr:") as prefix for each instance, where "***X***" is an specific main class. The output is the corresponding sentence, but words that belong to the target main class are masked. Box 1 from Figure 2 shows an example of this sub-task.

For experiments, we fine-tune the Portuguese T5 (PTT5) (Carmo et al., 2020)[4] on our corpus. Among the hyperparameters, we use AdamW optimizer with a learning rate of 5e-4, a max source and target length of 120 and 80 tokens, 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.

3.2.2 Word Infilling

The second stage in the pipeline consists of infilling the masked tokens. In general, the task can be defined as follows: given an AMR graph in a similar format to the one used at the previous stage and a masked sentence, the model predicts the masked words.

Each instance in the corpus is formatted as follows: a prefix, the AMR graph similar to the one used in the previous stage, the word "*contexto*:" (context), and the masked sentence. Box 2 from Figure 2 shows an example of the input and output. We use the expression "*preencher amr:*" ("fill amr:") as prefix and train a model for each main class separately and another model for all main classes together.

Similar to the previous stage, we fine-tune PTT5 on our task. The main reason use PTT5 is that it was pre-trained for a similar task ('filling-in-the-blank') (Carmo et al., 2020; Raffel et al., 2020). This way, we aim to leverage the learned knowledge in our use case. We use the same hyperparameters as the used ones in the first stage; however, we modify the source length to 200 tokens.

4 Results and Discussion

Table 2 shows the overall results for all the trained models on test set in terms of BLEU (Papineni et al., 2002), METEOR (Lavie and Agarwal, 2007), chrF++ (Popović, 2017), and BERTScore (Zhang et al., 2020) evaluation metrics[5][6]. In addition, we report the results for a baseline that generates sentences with no masked tokens. This baseline is obtained by fine-tuning PPT5 on our task. However, the input consists of a prefix "*gerar texto desde amr:*" ("generate text from amr:"), followed

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

[5]We execute 4 runs for each experiment and show the mean and standard deviation.

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

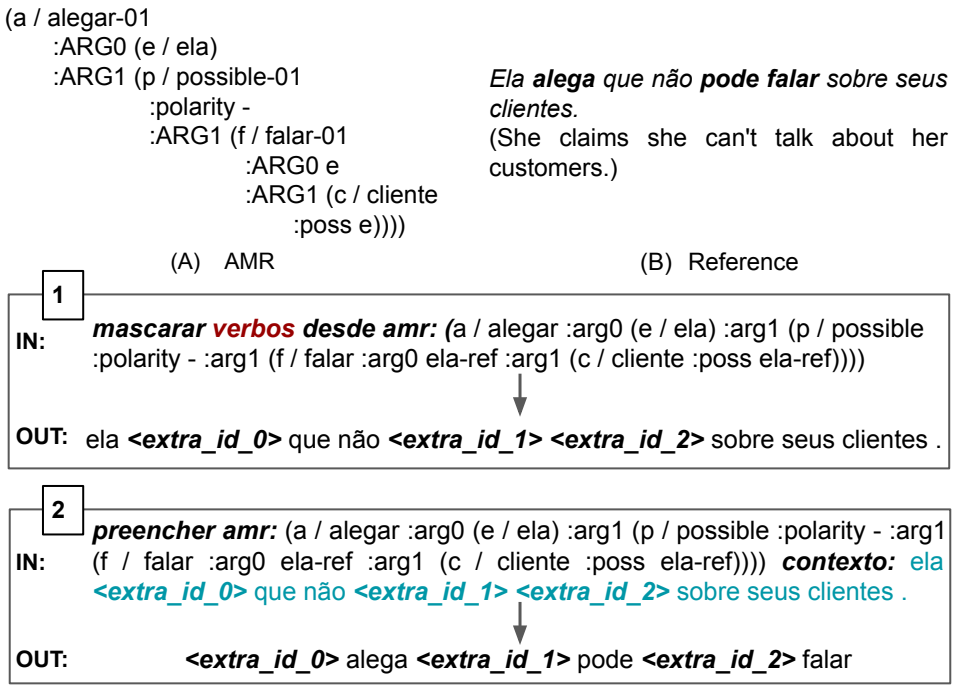


Figure 2: Pipeline Example. Box 1 describes the input and output for the masked surface realization module, and Box 2 illustrates the input and output for the word infilling module.

by an AMR graph represented by the PENMAN notation in a similar way as all already mentioned models, and the output is the original sentence.

Overall, results show a slight improvement over the baseline when we use the model trained on all main classes, mainly in BLEU (+1.63) and METEOR (+0.02) scores. Moreover, the best main classes to be masked seem to be “verb" and “qualifier". On the other hand, masking nouns and other POS classes harm the decoding performance. We might interpret this result as the characters in a sentence, and some connections between chunks are the most important in the realization of a sentence.

Another point to note is that it is better to train models on all main classes together instead of separately. A possible explanation is that more data can lead to better results. Also, examples from other main classes serve as negative examples for a specific main class, and it helps to improve its performance.

In order to verify which stage of the pipeline is affecting the overall performance, we evaluate each module separately. Table 3 and 4 shows the performance on both modules in terms of BLEU, METEOR and chrF++. However, for word infilling, we only evaluate BLEU-2 and BLEU-3, as the number of tokens to be predicted is three as most. In addition, we evaluate METEOR.

Concerning the Mask Surface Realization task, Table 3 indicates that verb masking leads to the best performance. A possible explanation for this result is that, as mentioned before, participants, situations, or locations in a sentence and connections between chunks are the most important and the easiest classes to predict during decoding. Also, it is worth noting that the verbs and qualifiers are less frequent in our dataset, as we can find 1.37-1.47 verbs/qualifiers per sentence. Therefore, it can make decoding easier than nouns (2.23 nouns per sentence).

Table 4 shows the opposite result, as the verb infilling is the most challenging task. However, we note that the values for BLEU-3 in the case of nouns and others are small. This way, it can confuse the infilling order in sentences with more tokens belonging to these classes. Moreover, we note that METEOR score for verbs less penalizes the performance (in comparison with BLEU), suggesting that the model can predict a different conjugation of the expected word.

It is worth noting that, in general, the bottleneck of the whole pipeline is the masked surface realization task, as values are similar to the overall performance. Even the verb-focused decoding,

		BLEU	METEOR	chrF++	BERTScore
Baseline		10.39 ± 0.48	0.29 ± 0.01	0.41 ± 0.01	0.82 ± 0.00
SEP	Noun	5.32 ± 0.56	0.22 ± 0.01	0.35 ± 0.01	0.80 ± 0.01
	Verb	8.95 ± 1.46	0.27 ± 0.01	0.39 ± 0.02	0.81 ± 0.00
	Qualifier	9.44 ± 0.87	0.27 ± 0.01	0.39 ± 0.01	0.81 ± 0.00
	Other	8.21 ± 0.99	0.27 ± 0.01	0.39 ± 0.02	0.81 ± 0.01
ALL	Noun	8.87 ± 0.69	0.28 ± 0.01	0.40 ± 0.02	0.81 ± 0.01
	Verb	**12.02 ± 2.13**	**0.31 ± 0.03**	0.42 ± 0.03	0.83 ± 0.01
	Qualifier	10.34 ± 1.34	0.30 ± 0.02	0.42 ± 0.02	0.83 ± 0.01
	Other	7.74 ± 1.71	0.28 ± 0.01	0.42 ± 0.02	0.81 ± 0.00

Table 2: Overall Results on test set. Experiments in block “SEP" are the ones in which a model is trained on each main class separately, and “ALL" are the ones in which a model is trained on all main classes together, but we evaluate it individually.

having the worst performance on the word infilling task, achieves the highest performance because the previous task gets the best one. A possible explanation for this problem is how the generation is performed. We use an encoder-decoder architecture in which the generation of a token depends on the previously generated tokens. This way, adding mask tokens in training could make it more difficult as the pre-trained model never saw these tokens in a generation task (these were used for training the blank infilling task). Among the alternatives to solve this issue, we could explore other strategies to determine the less confident tokens in a generated sentence and mask them for the next stage. Also, we could try a non-autoregressive model that can overcome the problem of dependency mentioned before (Su et al., 2021).

		BLEU	METEOR	chrF++
SEP	Noun	6.90 ± 1.05	0.45 ± 0.02	0.48 ± 0.02
	Verb	10.91 ± 0.48	0.42 ± 0.02	0.47 ± 0.02
	Qualifier	8.43 ± 0.82	0.30 ± 0.01	0.39 ± 0.01
	Other	10.11 ± 0.74	0.53 ± 0.03	0.56 ± 0.03
ALL	Noun	9.41 ± 1.65	0.49 ± 0.03	0.51 ± 0.03
	Verb	12.31 ± 1.52	0.45 ± 0.03	0.49 ± 0.04
	Qualifier	10.31 ± 1.27	0.32 ± 0.03	0.40 ± 0.03
	Other	10.22 ± 2.67	0.54 ± 0.04	0.56 ± 0.04

Table 3: Results on Mask Surface Realisation on dev test.

5 Manual Revision

We conduct a manual revision of the outputs for each model in order to check the main and most common errors. In particular, we select the two best models in our experiments, i.e., the ones trained on all main classes but focusing on masking/filling verbs and qualifiers.

		BLEU-2	BLEU-3	METEOR
SEP	Noun	33.80 ± 3.83	11.45 ± 3.33	0.46 ± 0.03
	Verb	18.53 ± 2.22	-	0.41 ± 0.01
	Qualifier	44.98 ± 9.12	-	0.57 ± 0.01
	Other	40.35 ± 3.99	18.48 ± 4.02	0.52 ± 0.02
ALL	Noun	41.20 ± 3.07	22.20 ± 3.43	0.57 ± 0.02
	Verb	20.95 ± 3.77	-	0.50 ± 0.02
	Qualifier	39.05 ± 10.21	-	0.65 ± 0.01
	Other	40.90 ± 4.70	19.55 ± 4.13	0.53 ± 0.03

Table 4: Results on Word Infilling on dev set

We analyze 35 instances from the test set and classify the outputs into four classes: (1) Accurate (“Acc"), for accurate outputs,(2) Hallucination (“Hall"), for outputs that are not related to the reference, (3) Cut chunk, for outputs that contains only a portion of the reference, and (4) Small Changes, for outputs with slightly different from the reference (some tokens are different). Table 5 shows the frequency of each class for all evaluated models.

In general, the model trained on all main classes, but focusing on verbs got the best results. It is worth noting the high number of hallucinations in all models, mainly when longer sentences are evaluated. Also, the cut chunks happen in the same cases. Moreover, there are several instances where only changing a simple word (or two) would be necessary to make the output similar to the reference. This problem happens mainly with connectors such as “*em*" (“in" or “at") or “*de*" (“of") (words highlighted in red in Figure 3) and with bad conjugations in the case of the verbs.

Figure 3 shows three examples. The first example shows that the model focused on verbs gets an accurate output (example 1). The second example shows that the outputs for models focused on verbs and qualifiers can generate paraphrases instead of

	Acc	Hall	Cut Chunk	Small Changes
Baseline	9	16	4	6
ALL-Verb	14	12	1	8
ALL-Qualifier	9	18	1	7

Table 5: Number of accurate outputs ("Acc") and errors in the human evaluation.

the same sentence; however, these are accurate too. Finally, the third example is a case in which the models generate hallucinations ("*eua investiram em 2010.*"), outputs with cut chunks ("*investir em os eua.*") or small changes.

Reference	*nada disso é criminoso .* none of this is criminal.
Baseline	*nada de isso .* none of that.
ALL-Verb	***nada de isso é criminoso .*** none of this is criminal.
ALL-Qualifier	*nada de criminoso .* nothing criminal.
Reference	*no vestiário , passou mal .* in the locker room , he felt sick .
Baseline	*passou mal .* he was feeling sick
ALL-Verb	***ele passou mal no vestiário .*** he got sick in the locker room.
ALL-Qualifier	***passou mal no vestiário .*** he got sick in the locker room.
Reference	*desde 2010 , o empresário investe nos eua .* since 2010, the businnessman invests in the usa.
Baseline	***eua investiram em 2010 .*** usa invested in 2010.
ALL-Verb	*investir em os eua .* invest in the usa.
ALL-Qualifier	***em** 2010 , o empresário **não** investiu no eua .* in 2010, the businessman did not invest in the usa.

Figure 3: Outputs comparison between the reference, the baseline, and the two best models in our experiments. The first lines for each model are the sentences generated in Brazilian Portuguese, and the next ones are the corresponding English translations.

6 Conclusion and Further Work

This work presents a simple two-stage approach to the low-resource AMR-to-text generation task. The approach consists of generating a masked surface realization in which some tokens are masked according to a POS class criteria and infilling the masked tokens according to the AMR graph and the previous masked surface realization.

Results show a slight improvement over the baseline, mainly in BLEU (1.63) and METEOR (0.02) scores. However, it is necessary to fine-tune the model on all the sub-corpus created together. Besides, we can note that verb masking seems to be the best strategy in this approach.

On the other hand, we note that the bottleneck of this approach is the masked surface realization model, as some generated tokens are different and unrelated to the original reference (hallucinations), and some tokens are omitted from the original reference. Some possible explanations for this problem are how the generation is performed -as each output word is conditioned on previously generated outputs-and the need to constrain the decoding process.

As further work, we plan to explore strategies to enforce the model to cover all the AMR concepts in the masked generated sentence and non-autoregressive text generation with pre-trained models, similar to Su et al. (2021). Besides, we plan to explore other strategies to mask tokens according to its confidence in decoding instead of using a POS-based one as the later can add more complexity to the task. Finally, we plan to extend this work to English AMR corpus, in order to make a better comparison in terms of generalization.

Limitations

This work tackles the AMR-to-Text generation task with a pipeline approach, and the results are similar to those obtained for previous work with the same amount of data (∼10-15 BLEU score). However, the performance could be different as the lengths of the sentences in our task are up to 23 tokens, and the sentences evaluated in works for English are longer.

Other limitation is related to the criteria used for masking some tokens as it can introduce more complexity, mainly for low-resource languages.

Acknowledgments

The authors of this work would like to thank the Center for Artificial Intelligence (C4AI-USP) and the support from the São Paulo Research Foundation (FAPESP grant 2019/07665-4) and from the IBM Corporation. Besides, this research carried out using the computational resources of the Center for Mathematical Sciences Applied to Industry (CeMEAI) funded by FAPESP (grant 2013/07375-0).

References

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.

Kris Cao and Stephen Clark. 2019. Factorising AMR generation through syntax. 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 2157–2163, Minneapolis, Minnesota. 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, Chris van der Lee, Emiel van Miltenburg, and Emiel Krahmer. 2019. Neural data-to-text generation: A comparison between pipeline and end-to-end architectures. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 552–562, Hong Kong, China. Association for Computational Linguistics.

Jeffrey Flanigan, Chris Dyer, Noah A. Smith, and Jaime Carbonell. 2016. Generation from abstract meaning representation using tree transducers. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 731–739, San Diego, California. 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.

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*.

Zdeněk Kasner and Ondřej Dušek. 2020. Data-to-text generation with iterative text editing. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 60–67, Dublin, Ireland. Association for Computational Linguistics.

Gerasimos Lampouras and Andreas Vlachos. 2017. Sheffield at SemEval-2017 task 9: Transition-based language generation from AMR. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 586–591, Vancouver, Canada. 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.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Kexin Liao, Logan Lebanoff, and Fei Liu. 2018. Abstract meaning representation for multi-document summarization. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1178–1190, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Shuming Ma, Pengcheng Yang, Tianyu Liu, Peng Li, Jie Zhou, and Xu Sun. 2019. Key fact as pivot: A two-stage model for low resource table-to-text generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2047–2057, Florence, Italy. 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.

Jonathan May and Jay Priyadarshi. 2017. SemEval-2017 task 9: Abstract Meaning Representation parsing and generation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 536–545, Vancouver, Canada. Association for Computational Linguistics.

Simon Mille, Roberto Carlini, Alicia Burga, and Leo Wanner. 2017. FORGe at SemEval-2017 task 9: Deep sentence generation based on a sequence of graph transducers. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 920–923, Vancouver, Canada. Association for Computational Linguistics.

Abelardo Vieira Mota, Ticiana Linhares Coelho da Silva, and José Antônio Fernandes De Macêdo. 2020. Template-based multi-solution approach for data-to-text generation. In *Advances in Databases and Information Systems: 24th European Conference, ADBIS 2020, Lyon, France, August 25–27, 2020, Proceedings*, page 157–170, Berlin, Heidelberg. Springer-Verlag.

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.

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.

Nima Pourdamghani, Kevin Knight, and Ulf Hermjakob. 2016. Generating English from abstract meaning representations. In *Proceedings of the 9th International Natural Language Generation conference*, pages 21–25, Edinburgh, UK. Association for Computational Linguistics.

Ratish Puduppully and Mirella Lapata. 2021. Data-to-text Generation with Macro Planning. *Transactions of the Association for Computational Linguistics*, 9:510–527.

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.

Ehud Reiter. 2018. Hallucination in neural NLG.

Leonardo F. R. Ribeiro, Martin Schmitt, Hinrich Schütze, and Iryna Gurevych. 2021a. 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.

Leonardo F. R. Ribeiro, Yue Zhang, and Iryna Gurevych. 2021b. Structural adapters in pretrained language models for AMR-to-Text generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4269–4282, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marco A. Sobrevilla Cabezudo and Thiago A.S. Pardo. 2022. Low-resource amr-to-text generation: A study on brazilian portuguese. *Procesamiento del Lenguaje Natural*, 68.

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.

Linfeng Song, Daniel Gildea, Yue Zhang, Zhiguo Wang, and Jinsong Su. 2019. Semantic neural machine translation using amr. *Transactions of the Association for Computational Linguistics*, 7:19–31.

Linfeng Song, Xiaochang Peng, Yue Zhang, Zhiguo Wang, and Daniel Gildea. 2017. AMR-to-text generation with synchronous node replacement grammar. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 7–13, Vancouver, Canada. Association for Computational Linguistics.

Yixuan Su, Deng Cai, Yan Wang, David Vandyke, Simon Baker, Piji Li, and Nigel Collier. 2021. Non-autoregressive text generation with pre-trained language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 234–243, Online. Association for Computational Linguistics.

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*.

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

This section encompasses the paper below.

CABEZUDO, M. A. S.; INÁCIO, M. L.; PARDO, T. Investigating Paraphrase Generation as a Data Augmentation Strategy for Low-Resource AMR-to-Text Generation, submitted to the Northern European Journal of Language Technology (NEJLT), 2023.

Contributions:

- We investigate two paraphrase generation approaches (monolingual and cross-lingual) to generate multiple references in an AMR-to-Text generation task.
- Experiments and analysis to prove the helpfulness of paraphrases for Low-Resource AMR-to-Text Generation.
- Release of a paraphrase-focused (multi-reference) AMR corpus for Brazilian Portuguese.

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

Marco Antonio Sobrevilla Cabezudo, University of São Paulo, São Carlos, Brazil msobrevillac@usp.br

Márcio Lima Inácio, University of Coimbra, Coimbra, Portugal mlinacio@dei.uc.pt

Thiago Alexandre Salgueiro Pardo, University of São Paulo taspardo@icmc.usp.br

Abstract Abstract Meaning Representation (AMR) has become a popular meaning representation (MR). One of its main statements is that it tries to abstract away from syntax information. This way, two similar sentences can be associated with the same AMR graph but expressed differently in the syntax. Curiously, the current AMR corpora associate one AMR graph with only one reference. On the other hand, other MRs usually include multiple references as it can help to deal with potential noise in data. This paper investigates the helpfulness of paraphrase generation in low-resource AMR-to-Text generation. We evaluate different ways to generate paraphrases and until what point they can be helpful. Automatic results show that this strategy largely surpasses the baseline and a classical data augmentation method, even using fewer training instances. Furthermore, the human evaluation shows that this strategy is more prone to generate syntactic-based paraphrases and can overcome the previous approaches. Finally, we release a Paraphrase-extended version of our AMR corpus.

1 Introduction

Abstract Meaning Representation (AMR) is one of the most popular semantic representations in recent years. AMR encodes the whole meaning of a sentence into a labeled directed and rooted graph, including information such as semantic roles, named entities, and co-references, among others (Banarescu et al., 2013). Moreover, it has been successfully used in diverse applications/tasks such as semantic parsing (Flanigan et al., 2014), automatic summarization (Vilca and Cabezudo, 2017), and paraphrase detection (Issa et al., 2018).

Part of its popularity is due to its broad use of mature linguistic resources, such as the PropBank (Palmer et al., 2005), and its attempt to abstract away from syntax. Figure 1 shows the AMR graph (Sub-figure A) and the PENMAN notation (Matthiessen and Bateman, 1991) (Sub-figure B) for the sentence "*The boy must go.*" as well as another alternative surface forms. We note that all surface forms convey the same meaning but are syntactically and lexically different.

Curiously, and as far as we know, AMR corpora only contain one reference per each AMR graph, not taking advantage of their syntax-independent nature. Conversely, other semantic representations such as the proposed one at WebNLG challenge (Gardent et al., 2017) or the E2E dataset (Dušek et al., 2020) usually present multiple references for each representation. Besides, having multiple references is beneficial for developing Natural Language Generation systems since it helps systems deal with potential noise by increasing the data diversity (Dušek et al., 2020).

On the other hand, manually producing additional references can be an expensive task. In particular, words included in surface forms are highly linked to the concepts in an AMR graph (Banarescu et al., 2013). This way, references created for an AMR graph must include only its concepts in their canonical form or possible derivations (in addition to the relation realization) as much as possible. For example, the concept "boy" in Figure 1 should not be replaced by "guy" in a possible surface form (even if both words can be interchangeable). An alternative to the manual annotation task is automatically generating new references using paraphrase generation models. However, we still need to satisfy the statement mentioned previously.

Paraphrase generation has proven helpful for data augmentation in diverse tasks such as natural language understanding (Okur et al., 2022), question answering, and task-oriented dialog systems (Gao et al., 2020). However, as far as we know, this strategy has yet to be studied for improving the performance of AMR-to-Text

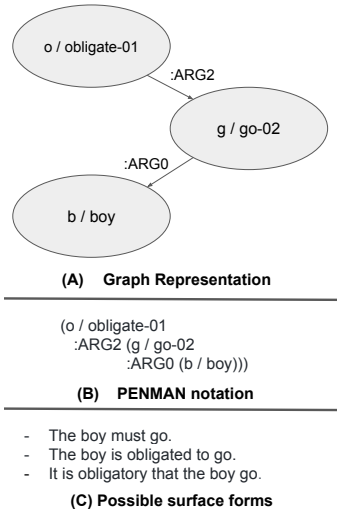


Figure 1: AMR example for the sentence "The boy must go."

generation nor for creating more robust AMR corpora. Besides, Other approaches used in the literature use AMR parsers for generating new instances (Castro Ferreira et al., 2017; Mager et al., 2020; Ribeiro et al., 2021) might easily surpass it. However, we are interested in low-resource settings in which AMR parsing might negatively affect the performance of the AMR-to-Text generation task.

This work aims to evaluate the helpfulness of paraphrases for the Low-resource AMR-to-text generation task for Brazilian Portuguese (BP). In general, we explore two different ways of generating paraphrases. The first uses a paraphrasing model for Portuguese (Pellicer et al., 2022). The other one uses English as a pivot language and is divided into two sub-approaches: one that only uses machine translation models (for translating and back-translating), whereas the other one also includes a paraphrase generation model for English.

Due to the possibility of introducing noise into the models by adding unrelated paraphrases, we explore using three selection criteria. These criteria can help to select a specific number of high-quality paraphrases. Finally, we explore if the use of the added paraphrases can benefit when they are added into 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 an 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 last 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 paraphraser 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 paraphraser 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 (Sennrich et al., 2016) 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 (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 *English-to-Portuguese*) 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 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 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, respectively.

3.2 Settings

We evaluate different criteria such as the number of paraphrases 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 the paraphrases only to the training set, i.e., we use it as a paraphrase-based data augmentation strategy.
- Train-Dev (B): We add the paraphrases to the training and development sets. It aims to verify if increasing the diversity in the development set can help converge to a better performance. Besides, It also aims to create a multi-reference AMR corpus.

Finally, the new multi-reference AMR corpus consists of 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 As we aim to evaluate the helpfulness of paraphrasing for increasing the number of references, the baseline model is obtained by fine-tuning PPT5 (Carmo et al., 2020) on the original AMRNews, i.e., with 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 the works of the 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 correferences.

This approach suffers from problems such as the lack of parentheses or correferences. 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.

3.4 Hyperparameters

Training Models are generated by fine-tuning the Portuguese T5 (PTT5)[7] 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.

[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`.

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

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

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

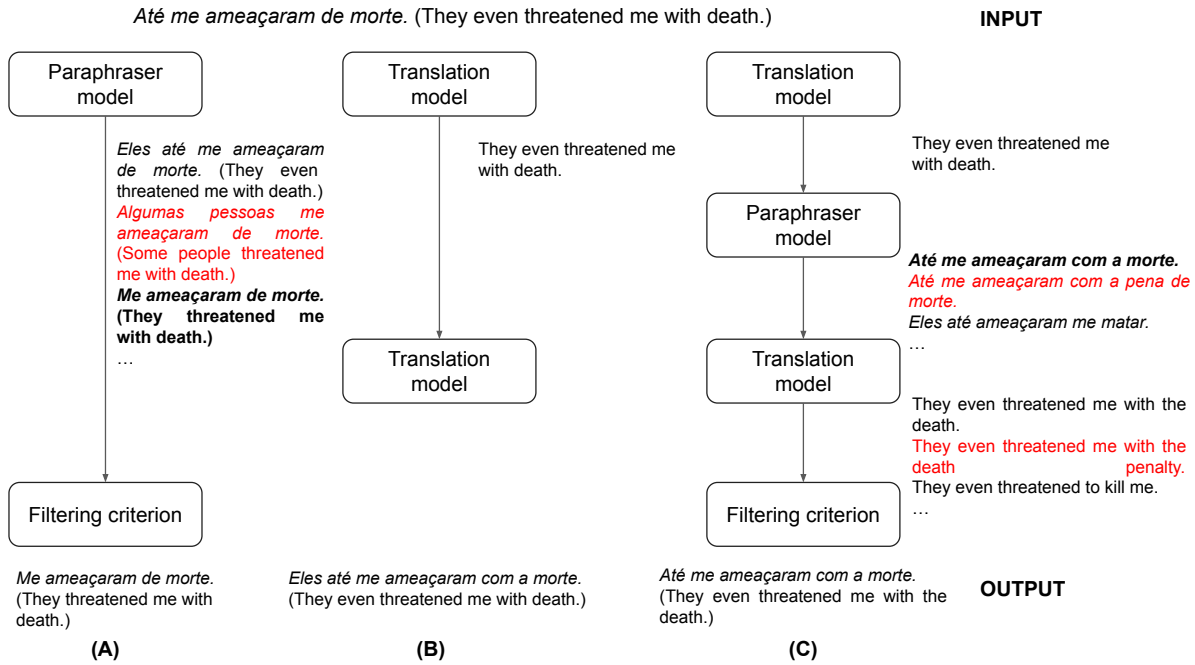


Figure 2: Pipeline Example for Paraphrase Generation. (A) Portuguese approach: A sentence written in Brazilian Portuguese (BP) is given to a Portuguese paraphrase model, and it generates the paraphrases. (B) English-pivot approach: A sentence written in BP is given to a machine translation model that 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 the 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 for selecting the best paraphrases.

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.

4 Results and Discussion

Table 1 shows the overall results for the models on the test set from the original AMR corpus[8]. The results are reported for each approach, i.e., data augmentation by parsing, using Portuguese and English-pivot paraphrases, and for each paraphrase selection criteria. In addition, the results are obtained by training the models in the setting T (data augmentation strategy). The results are reported in terms of BLEU (Papineni et al., 2002), METEOR (Lavie and Agarwal, 2007), chrF++ (Popović, 2017), and BERTScore (Zhang et al., 2020)[9][10].

Overall, we can see that all the paraphrase-based models surpass the baseline in all the metrics, being the largest difference of 3.81 for BLEU, 0.04 points for METEOR, 0.05 points for chrF++ and 0.02 points for BERTScore[11], proving the helpfulness of augmenting data via paraphrases.

Concerning the paraphrase generation strategy, we note that, as expected, paraphraser models (both for Portuguese and English-pivot approaches) produce better results than only translation models (except for TER where the results are larger). In addition, it is not clear what is the best paraphrase selection criteria for each approach. However, it is worth noting that each selection criterion gets the best result using a different number of paraphrases. For example, adding a few paraphrases (5-6) produces the best results for the Portuguese-based approach when we use BLEU and METEOR as criteria. In contrast, we need

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

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

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

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

			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. The best models for each selection criteria are shown. 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.

to add eight paraphrases to get the best results in the case of TER. On the other hand, all the models that follow the English-pivot approach need more paraphrases (7-9) to achieve their best performance.

We also note that all approaches outperform the results obtained by the classic data augmentation approach (Bosque-Augmented in Table 1). Besides, we highlight that paraphrase-based approaches need fewer instances to achieve better performance. In particular, the maximum number of instances the paraphrase-based approach adds is up to 4000 (adding ten paraphrases per instance). However, the Portuguese approach only needs half to achieve higher performance. Surprisingly, we can see that even adding only one paraphrase per instance (BACK-TRANSLATION 1-1 experiment in Table 1), i.e., almost duplicating the dataset, can achieve comparable results.

On the other hand, the main drawback is that the performance does not seem to increase more with more than 8 paraphrases and even it starts decreasing in some cases (see Figure 3 and Figure 6 in Appendix A). Therefore, it would be worth evaluating if increasing the number of instances in the classic data augmentation approach would produce a bigger improvement or introduce more noisy data (due to the extremely low-resource setting), harming the 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? Concerning the setting T (only adding instances to the training set), Figure 3 and 6 show how the performance on the development set changes according to the number of paraphrases used for augmenting the data.

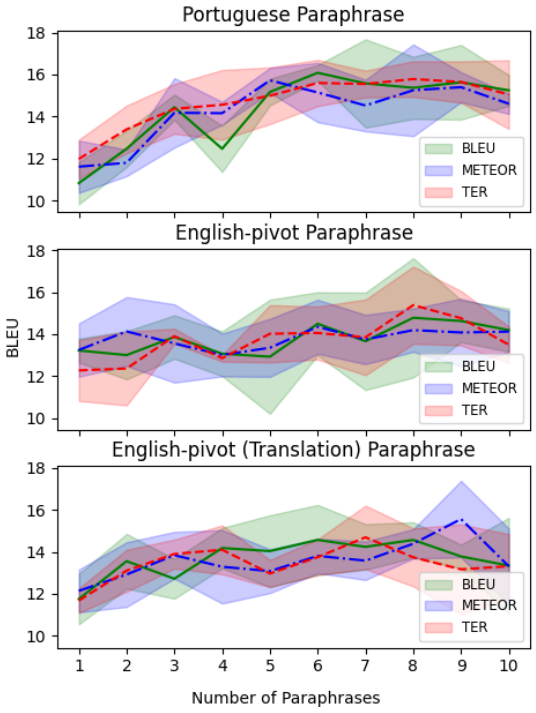


Figure 3: BLEU scores per selection criterion and per number of selected paraphrases in the T setting. Results are shown on the development set.

In general, the highest performance in all metrics can be achieved by adding a few paraphrases (up to 5-6) in the case of the Portuguese paraphrasing approach. However, it is necessary to add more paraphrases for the English-pivot approaches (7-9). A possible explanation is the existence of a quantity-quality trade-off, i.e., English-pivot approaches can generate lower-quality paraphrases (although not too low). However, when we

increase the number of paraphrases, the higher diversity can improve the performance.

Another point to highlight is that the back-translation strategy performance (*English-pivot (Translation) Paraphrase* in Figures 3 and 6), presents the steepest drop in all metrics when more data is added (in particular, when 10 paraphrases are added), showing that we need to select the instances when following this strategy carefully. On the other hand, the other approaches suffer from a soft drop, BERTScore being the less affected metric. The semantic nature of this metric can explain that it is not affected by synonyms/paraphrases in the outputs.

In addition, we can see that the standard deviation for most metrics increases when more paraphrases are added, harming mainly the BLEU score. It is expected as BLEU can be seen as a more restrictive metric. A plausible explanation is that adding more paraphrases in training makes the model more prone to generate various paraphrases.

Figure 4 and 7 show the results when the models are trained on the setting B (adding instances to both training and development sets). In general, adding more paraphrases produces better results (7-9 paraphrases) in all metrics for all approaches. On the other hand, adding 10 paraphrases leads to a decrease in the performance, being that both the Portuguese and the English-pivot ones (the later in the back-translation strategy - English-pivot (Translation) Paraphrase in Figures 4 and 7) the most affected.

What are the best paraphrase selection criteria? With respect to the setting T, we note mixed behaviors (Figures 3 and 6) that mainly depend on the paraphrase generation approach. In the case of the Portuguese approach, we can see that METEOR and BLEU seem to be better options when we add 5-6 paraphrases; however, more paraphrases produce a smooth decrease. A plausible explanation is that these metrics are good for selecting the best instances quickly when the analyzed paraphrases are good enough, assuming that the Portuguese approach introduces less noise in the paraphrase generation. In this way, the initial paraphrases contain more words overlapping with the original sentence, serving as an oversampling strategy for dealing with infrequent words/n-grams. Conversely, TER needs more paraphrases (7-9) to get the same behavior.

Concerning the English-pivot approaches, all the selection criteria tend to achieve better results when more paraphrases are included. The standard deviation in the results is higher than the one obtained by the Portuguese approach. It can confirm that using English as a pivot slightly harms the paraphrase quality and generates more diverse and possibly less-related paraphrases, making the models produce more diverse

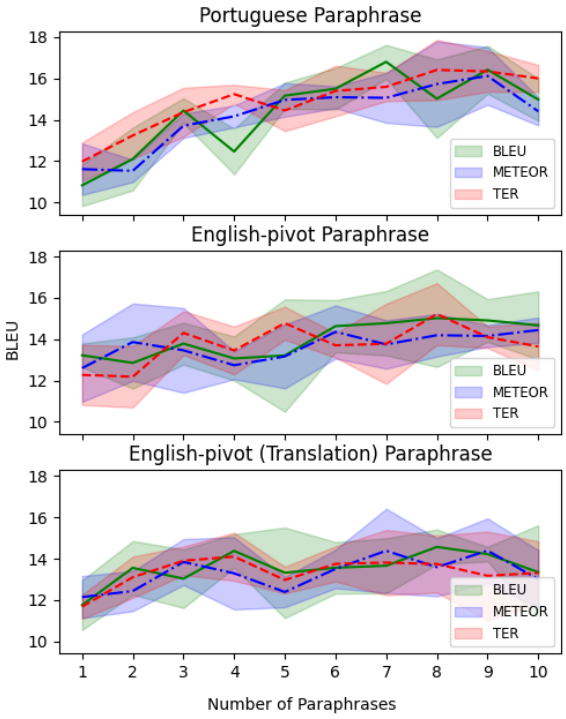


Figure 4: BLEU scores per selection criterion and per number of selected paraphrases in the B setting. Results are shown on the development set.

outputs.

It is also worth noting that, even though all selection criteria are affected by the English-pivot approach, TER produces different results/trends. In particular, we can see a notorious drop when we only use back-translation, showing that it is more sensitive to output quality. However, it can achieve comparable results to the Portuguese approach in the test set when the English paraphrase generation is inserted in the pipeline instead of generating diverse back-translations. It proves the usefulness of the English paraphrase generation in cases where no paraphrase generation models exist in a non-English language. On the other hand, METEOR also seems to be the better option when we only have English-to-X and X-to-English machine translation systems.

With respect to the analysis on setting B, Figures 4 and 7) show a different result on the Portuguese approach, being BLEU and TER the best selection criteria. However, they (in particular, TER) present high standard deviations. This way, this could lead us to wrong conclusions. In order to get better conclusions, we also evaluate the models on the test set. The first row in Table 3 (one reference evaluation) shows the results in both settings (T and B) on the development and test sets. We can see that although TER on setting B produces the highest performance on the development set,

it presents a decrease on the test in terms of BLEU (from 14.02 to 13.77). This finding exhibits this metric's nature that does not prioritize the exact words/n-grams in its evaluation.

Concerning the English-pivot approaches, we note a similar behavior to the setting T regarding the use of back-translation and back-translation + English paraphrase generation. Again, however, we can see that BLEU and METEOR produce the best results.

How much does the paraphrase's quality affect the performance? In order to measure how much the paraphrase's quality affects the AMR-to-Text performance, we train a model under one of the best settings, but we change the training data. This way, we use the worst paraphrases instead of the best ones. In particular, we use the Portuguese approach, the METEOR criterion, and 5 paraphrases. To get the worst 5 paraphrases, we select the last 5 paraphrases from the experiment with 10 paraphrases[12].

Table 2 shows the results of the development set along with some similarity metrics between the paraphrases and the original instance in the training set. In particular, we calculate the cosine similarity and the three selection metrics used in the experiments. As can be seen, all the similarity metrics present a high drop, being that the cosine similarity is the less affected. It can be explained by the semantic nature of this metric, which can overcome problems associated with using synonyms or semantic-related words and penalizes some changes less than the other metrics.

Among the results, we note that the performance in all metrics decreases in general. However, BLEU seems less affected (a drop of 0.34 points). A point to highlight is that, opposite to the slight drop in terms of mean, the standard deviation increases twice its value. 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 even non-sense) words.

How much does the inclusion of paraphrases in the development set contribute? Given that the current corpus includes only one reference per instance, we create a multi-reference version of the test set. This version is created by applying one of the best previously evaluated strategies on the test set and modifying/discarding some instances. Specifically, we use the model trained following the Portuguese approach, which includes five paraphrases per instance and uses METEOR as the selection criteria. This way, we create a multi-reference test set that includes 1-6 references per instance.

Table 3 summarizes the performance of the Portuguese-based model trained in both settings (T and B) for each selection criterion. The performance is also evaluated on the one-reference and multi-reference test sets. Concerning the one-reference evaluation, we note that adding paraphrases to the development set produces mixed results and increases the standard deviation, making the BLEU score the most affected metric. It suggests that this strategy might be helpful but can add more noise and instability to the models. In particular, we note that the most benefited selection criterion is BLEU as the performance increases 1.24 in terms of BLEU (from 13.01 to 14.25). On the contrary, TER produces a performance drop in BLEU, 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).

Analyzing 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, we note that the performance difference for the BLEU and METEOR selection criteria is similar to the obtained in the one-reference evaluation.

5 Manual Revision

Aiming to understand some results, we conduct a manual revision. We select 112 instances from the development set and verify the main mistakes or phenomena generators produce.

We define 2 main categories for evaluating the diverse models: valid and invalid outputs. Valid outputs include 3 sub-categories: "equivalent", which means that the system output and the reference are the "same" (it can happen with some minor modifications such as the use of determinants); "semantic", when the system output is equivalent to the reference but using different words or non-syntax paraphrases; and "syntactic", when the output is equivalent to the reference but there are some syntax differences (e.g., changing active voice to passive voice).

On the other hand, Invalid outputs include 3 subcategories: "missing", when the system output is similar to the reference, but some few words are omitted; "partial hallucination", when the system output contains part of the reference and part of extra information not related to the input/reference; and "total hallucination", when the output is unrelated or different from the reference.

[12] 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.

EVALUATION	SETTING		DEV				TEST			
	PARAPHRASE	CRITERIA	BLEU	METEOR	chrF++	BERTScore	BLEU	METEOR	chrF++	BERTScore
One Reference	T	BLEU	16.08 ± 0.38	0.36 ± 0.01	0.45 ± 0.01	0.83 ± 0.00	13.01 ± 0.45	0.32 ± 0.01	0.44 ± 0.01	0.83 ± 0.00
		METEOR	15.73 ± 0.59	0.37 ± 0.01	0.46 ± 0.01	0.84 ± 0.00	14.20 ± 0.41	0.33 ± 0.01	0.46 ± 0.01	0.84 ± 0.01
		TER	15.79 ± 0.85	0.35 ± 0.01	0.45 ± 0.01	0.83 ± 0.01	14.02 ± 1.48	0.33 ± 0.02	0.44 ± 0.01	0.84 ± 0.01
	B	BLEU	16.81 ± 0.82	0.36 ± 0.02	0.45 ± 0.02	0.83 ± 0.01	14.25 ± 1.61	0.33 ± 0.01	0.45 ± 0.02	0.83 ± 0.01
		METEOR	16.12 ± 1.39	0.36 ± 0.01	0.45 ± 0.01	0.84 ± 0.01	14.75 ± 1.35	0.33 ± 0.02	0.46 ± 0.01	0.84 ± 0.00
		TER	16.41 ± 1.46	0.36 ± 0.02	0.46 ± 0.01	0.83 ± 0.01	13.77 ± 1.14	0.33 ± 0.01	0.45 ± 0.01	0.84 ± 0.00
Multi-reference	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 for the Portuguese approach on setting T and B using one reference and multi-references in the test set. The results are shown for each criteria.

The analyzed approaches are described as follows:

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

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, 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 in the output; 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 behavior 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 introduce more noise, thus, being prone to generate more hallucinations. In other words, the model generates some common n-grams (or paraphrases) but adds extra non-related words in several cases.

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

"literally". On the other hand, paraphrase approaches are less prone to generate total hallucinations, being the best the Portuguese approach and the worst the English-pivot approach that applies Back-translation and Paraphrase generation. We can see another example in Figure 5.

		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. 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 5: 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.

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

Models are expected to produce hallucinations as they are trained on an extremely small corpus (402-4020 instances); however, generating bad paraphrases can exacerbate this behavior. For example, we show the paraphrases generated by one of our approaches 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 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 dialog 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 work presents a study of the helpfulness of paraphrases for the AMR-to-text generation task for Brazilian Portuguese (BP). First, we explore two strategies for generating paraphrases: one that uses a model trained on the target language (Brazilian Portuguese) and the other that uses English as a pivot (English-pivot approach). Also, to ensure the quality of the outputs, we evaluate three diverse criteria and explore how the number of added paraphrases can affect the model performance. All these experiments are conducted by training in two settings: one that focuses on adding the paraphrases only to the training set (data augmentation strategy) and the other that creates a paraphrased-extended AMR corpus (adding paraphrases to both training and development sets).

Overall, we show that paraphrase generation is an exciting and straightforward data augmentation strategy that largely surpasses the baseline and a classic data augmentation strategy used in AMR-to-text generation in an extremely low-resource setting; however, not all the metrics work in the same way, and it is necessary to select the paraphrases carefully. On the other hand, the paraphrase-extended AMR corpus shows a slight improvement, and adding more paraphrases per instance is better for better performance.

The manual revision shows that Portuguese paraphrase-based models are better generators of valid outputs, where syntactic variations are the most common effect. On the other hand, English-pivot presents lower performance. Also, paraphrase-based models are also prone to generate hallucinations and missing words (mainly the English-pivot ones), making it necessary to curate the corpus as some paraphrases are non-related and can harm the overall performance.

Among the further work, we plan to curate the AMR corpus that includes paraphrases and explore other strategies to generate syntax-focused paraphrases. Also, although our method obtains better results than a baseline and a classical data augmentation approach, the main limitation of our approach is that it can only add a limited number of paraphrases. Although classic data augmentation approaches depend on the quality of the AMR parser, they can increase the number of instances instead of adding variability in the outputs. In this way, we would like to combine the proposed paraphrase-based strategy with the classical data augmentation methods to create a bigger AMR corpus, as it presents some limitations in terms of generating augmented data.

Finally, The AMR corpus for Brazilian Portuguese that includes paraphrases and the code is freely available at `url`[13].

Considerations

This work is intended to investigate the helpfulness of paraphrase generation as a data augmentation strategy for low-resource AMR-to-Text generation and try to understand what is the best way to combine different criteria such as the paraphrase approach (Portuguese or English-pivot), the selection of paraphrases, number of added paraphrases, and the use of the paraphrases in different settings. In this way, we do not consider current large language models in the evaluation for comparison. It is well-known that these models can achieve impressive results on diverse tasks. However, this work aims to study a particular approach rather than compare it with current possible SotA models.

References

Afonso, Susana, 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).

Banarescu, Laura, 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.

Carmo, Diedre, 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*.

Castro Ferreira, Thiago, 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.

Dušek, Ondřej, 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.

Edunov, Sergey, Myle Ott, Marc'Aurelio Ranzato, and Michael Auli. 2020. On the evaluation of machine translation systems trained with back-translation. In

[13]The data and code will be available after the acceptance for publishing.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 2836–2846, Online. Association for Computational Linguistics.

Flanigan, Jeffrey, Sam Thomson, Jaime Carbonell, Chris Dyer, and Noah A. Smith. 2014. A discriminative graph-based parser for the Abstract Meaning Representation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1426–1436, Baltimore, Maryland. Association for Computational Linguistics.

Gao, Silin, 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.

Gardent, Claire, 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.

Hou, Yutai, 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.

Inácio, Marcio Lima, 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*.

Issa, Fuad, 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.

Junczys-Dowmunt, Marcin, 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.

Lavie, Alon 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.

Mager, Manuel, 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.

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

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

van Noord, Rik 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.

Okur, Eda, 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.

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

Papineni, Kishore, 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.

Pellicer, Lucas Francisco Amaral Orosco, 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.

Popović, Maja. 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.

Quan, Jun 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.

Raffel, Colin, 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.

Rentschler, Sophie, 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.

Ribeiro, Leonardo F. R., 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.

Scherrer, Yves. 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.

Sennrich, Rico, 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.

Snover, Matthew, 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.

Sobrevilla Cabezudo, Marco Antonio, 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.

Sobrevilla Cabezudo, Marco Antonio 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.

Vilca, Gregory César Valderrama 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.

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

Zhang, Yuan, 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.

Zheng, Renjie, 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.

Zhou, Jianing 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

Table 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.

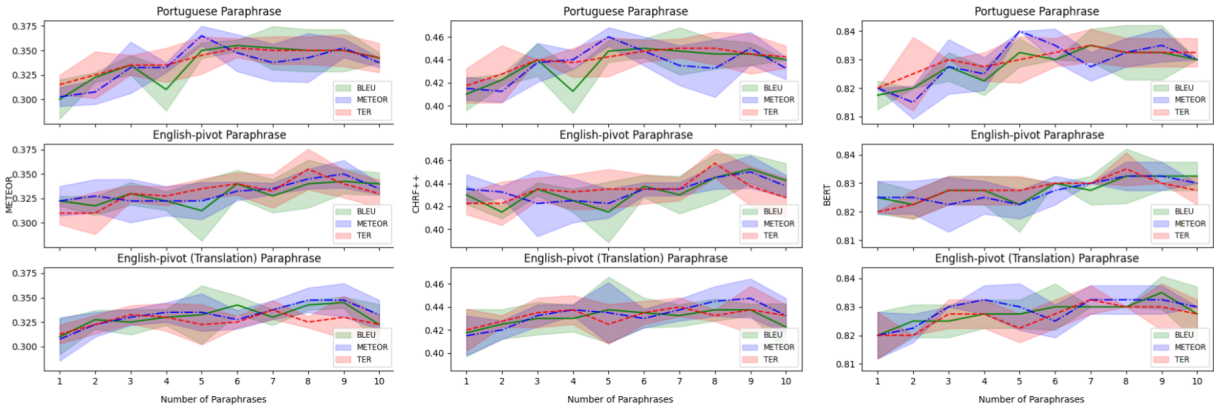


Figure 6: METEOR, chrF++ and BERT scores per selection criterion and per number of selected paraphrases in the T setting. Results are shown on the development set.

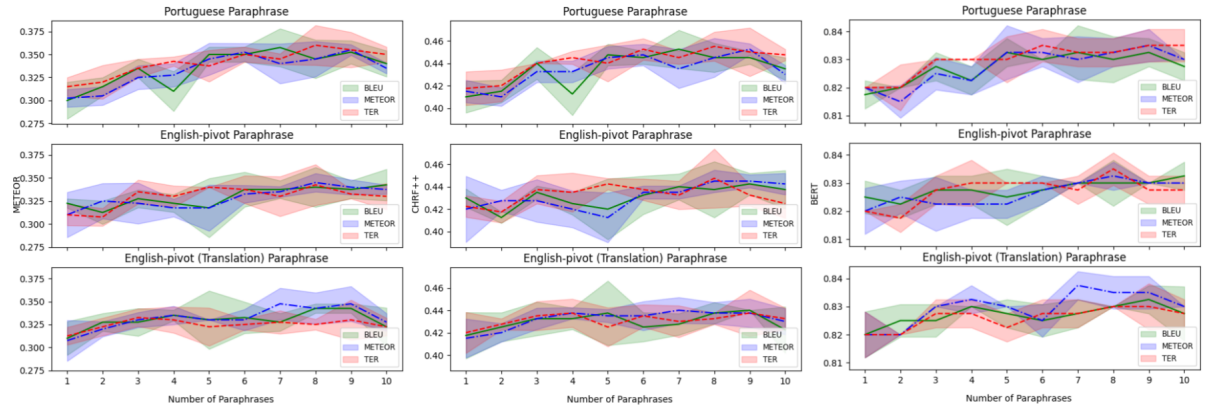


Figure 7: METEOR, chrF++ and BERT scores per selection criterion and per number of selected paraphrases in the B setting. Results are shown on the development set.

CHAPTER

7

CONCLUDING REMARKS

This chapter presents the concluding remarks, lessons learned, answers to research questions and hypotheses, contributions, research limitations, and potential future work based on what was developed in this thesis.

7.1 Conclusions and Contributions

The main goal of this work was to explore, develop, adapt and evaluate NLG methods for Brazilian Portuguese from AMR. The main hypothesis guiding this goal is that it is possible to develop natural language generation methods for Portuguese from AMR with satisfactory accuracy. Overall, the main research's goal was achieved after a series of studies, culminating with the papers presented in Chapter 6.

In particular, we defined and adapted the original AMR guidelines for annotating sentences in Brazilian Portuguese, built the first general-purpose AMR corpus for Brazilian Portuguese, and it served as motivation to extend the annotation to the opinative domain (INÁCIO, 2021) [1]. Besides, some linguistic phenomena, classified as hard cases, were studied. All this work served to achieve the specific goal "*Creating an AMR corpus for the development and evaluation of NLG methods in Brazilian Portuguese*".

On the other hand, various studies were conducted to "*evaluate the potential of using English-focused AMR corpus for improving the AMR-to-Text generation task for Brazilian Portuguese*". Starting from using the translation of the English AMR corpus with no extra data (Section 5.1 in Chapter 5) until evaluating it in diverse cross-lingual scenarios (Section 5.3 in Chapter 5), we could verify the helpfulness of this resource, proving the hypothesis "*English AMR corpus, even though linguistic phenomena differences, improves the AMR-to-text generation task for Brazilian Portuguese*".

[1] The corpus is available at <https://github.com/nilc-nlp/AMR-BP>.

In order to achieve the specific goal "*Comparing pipeline-based methods with end-to-end neural methods for AMR-to-Brazilian Portuguese generation task*", this research explored a two-stage approach that tries to generate masked utterances and leverage the potential of pre-trained models such as T5 for filling the masked tokens (Section 6.1 in Chapter 6). Results showed a slight performance improvement. However, it confirmed the hypothesis "*Pipeline approaches lead to improvements in low-resource AMR-to-text generation.*".

Concerning the goal "*Evaluating the performance of data augmentation methods in AMR-to-text generation and applying strategies for better selecting the augmented data*", this thesis proposed to evaluate the helpfulness of paraphrase generation for improving the AMR-to-Text generation performance (Section 6.2 in Chapter 6). Results confirmed the hypothesis "*Data augmentation improves the performance of Low-resource AMR-to-text generation.*", even in a extremely low-resource setting (starting 402 instances in the training set). In addition, a novel paraphrase-based multi-reference AMR corpus for Brazilian Portuguese was released. This corpus aims to improve the evaluation of AMR-to-Text generation and parsing, and the process to get it can be replicated for other languages, opening the possibility to improve the evaluation in this area.

With regarding the research questions, this thesis answered them successfully:

- *How different is English AMR corpus from Portuguese AMR corpus in terms of linguistic phenomena?*
  Answer: In general, we can easily use the original English AMR guidelines for annotating corpora and overcome some argument/adjunct identification issues as there is a semantic repository for Brazilian Portuguese (Verbo-Brasil (DURAN; ALUÍSIO, 2012; DURAN; ALUÍSIO, 2015)). However, some cases in both news and opinative texts are different and need to be treated carefully. Some examples of our work are the indeterminate subjects, NP Ellipsis, and diminutives (mainly for opinative) (works described in subsections 4.1 and 4.2).

- *Is it possible to leverage the cross-linguistic potential of the English AMR corpus for increasing the size of the Portuguese AMR corpus and the performance of AMR-to-text generation?*
  Answer: Yes, it is. In general, we can infer that current methods can overcome possible structural divergences (work described in subsection 5.3). However, in our particular case, we faced some issues related to domain divergence that can be beaten by domain adaptation strategies (such as fine-tuning). However, the main issue is the size of our AMR corpus, as it prevents leveraging the knowledge more.

- *What is the best strategy for dealing with data sparsity in AMR-to-text generation?*
  Answer: Earlier experiments with models only trained on the AMR corpus for Brazilian Portuguese showed that the best strategy is to generate a linearized version of the AMR

graph and use a Statistical Machine Translation system to generate the surface form (work described in subsection 5.1). However, when the translated version of the AMR English corpus is used, byte-pair encoding-based representation in conjunction with Graph-to-Sequence architectures produced the best performance on the AMR-to-Text generation (work described in subsection 5.2). Currently, the T5-based model performs best in both settings (with translated English AMR corpus and without) (work described in subsection 5.3). On the other hand, we verify that pipeline approaches can help dealing with data sparsity by using the knowledge provided by pre-trained language models (work described in subsection 6.1).

- *What is the best way to leverage the knowledge provided by the English AMR corpus?*
  Answer: In principle, the best way to leverage the English AMR corpus is to use it for annotating more data, as it leads to task improvements (works described in subsections 5.1 and 5.2). On the other hand, we can use the English AMR corpus as a base to train a general model and then fine-tune on our gold AMR corpus (work described in subsection 5.3).

- *How does data augmentation methods behave on AMR-to-text generation in low-resource settings and what is the best way to augment data?*
  Answer: Experiments showed that paraphrase generation can be useful for improving the performance in Low-resource AMR-to-Text generation, overcoming some classical data augmentation approaches (work described in subsection 6.2). However, it has some limitations as this strategy is linked to the corpus. This way, it should be interesting to combine classical data augmentation with paraphrase-based approaches in order to gain more diversity.

7.2 Limitations and Considerations

The major limitation of this work is the size of the AMR-BR corpus. It has 870 instances divided into 402, 224, and 244 instances for training, development, and testing, respectively. Even using diverse strategies to approach this low-resource setting, the performance achieves a limit and makes it necessary to increase the corpus size. On the other hand, the sentences annotated in this work are short (up to 23 tokens). In this way, increasing annotated sentences (if longer sentences are annotated) could change/decrease the overall performance. It is also worth noting that this research mainly focused on improving and evaluating the accuracy of the diverse strategies and models, giving less priority to other aspects such as the readability.

On the other hand, although current neural generation systems have shown an impressive performance on the Low-Resource AMR-to-Text generation task, they still present some challenges, such as the hallucination generation, as these models usually tend to generate hallucinations, mainly in out-of-domain sets or, like ours, disperse small datasets.

Another limitation of this research is associated to the difficulties of dealing with some linguistic challenges such as irony and metaphor generation. For example, in our study, we show that metaphor can be difficult to encode an we proposed a simple way to do it. However, this thesis did not take into account it during generation. This way, more studies about the metaphor encoding, parsing and generation should be addressed.

Finally, it is worth asking what AMR's future is and its application to tasks such as language generation in front of the current large language models (LLMs). This thesis presented two big directions: using AMR as a semantic representation for corpora building in Brazilian Portuguese and exploring diverse methods and strategies to overcome some limitations in low-resource AMR-to-Text generation in the same language. However, we only included experiments with T5. About it, it is possible to say as follows:

- Concerning semantic representations, Abstract Meaning Representation presents a way to explicitly represent the knowledge or the "meaning" of a sentence. This way, it intends to focus on how to represent diverse semantic phenomena instead of competing against language models that work more like a black box.
- Concerning the language generation, current large language models (LLMs) such as GPT-3 (BROWN *et al.*, 2020), or their derivatives (e.g., chatGPT) have proven to be useful in language generation tasks, largely outperforming various next-generation models. This way, it might be seen as a tool for improving AMR-to-text generation and combining a symbolic approach (starting from a semantic representation) with a neural approach (provided by the LLMs).

7.3 Future Work

Some topics that can be potentially investigated in future works are described below:

- As mentioned in the previous section, the AMR corpus for Brazilian Portuguese is smaller (870 instances) than their analogous corpora for English and Chinese. Although reasonable/comparable performance in the AMR-to-Text generation task in this setting was achieved, extending the AMR corpus is an interesting direction as it can serve semantics (with a linguistic bias) and computational studies.
- As mentioned in the limitations, another interesting direction is to explore how to encode other linguistic phenomena, such as metaphors, and evaluate how the current models perform on them.
- In Chapter 5, the English AMR corpus was translated into Brazilian Portuguese, and the concepts, relations, and alignments were inherited from the English version. This resource produced improvements in the AMR-to-Text generation task for Brazilian Portuguese.

However, more work can be done by analyzing the quality of the translated AMR graphs to measure the impact on diverse tasks such as AMR parsing. Besides, it could serve for measuring/quantifying the actual divergences between English and Brazilian Portuguese and finding some linguistic phenomena for Brazilian Portuguese that AMR cannot cover, as a complement of the work of Anchiêta and Pardo (2018).

- In Chapter 6, the pipeline strategy adopted in Section 6.1 showed at least two directions about how the decoding process is performed. The first consists of using constrained decoding to force the model to produce all input tokens and then determine the tokens for which the model is less confident, mask them, and leverage the capabilities of T5 to fill the masks. The other direction is to explore alternative decoding processes. In particular, some non-autoregressive text generation models could be studied as they try to eliminate the dependency on previous tokens.

- Section 6.2 shows other directions that can be studied. Firstly, the paraphrase-extended AMR corpus must be curated because some paraphrases contain unrelated words or bad translations. Furthermore, it can negatively affect the performance of the models and the evaluation. Lastly, syntax-focused strategies to create paraphrases can be studied as AMR graphs are more associated with syntax.

- Finally, other data augmentation strategies similar to the used in the literature must be studied. Besides, semi-supervised approaches are an interesting direction as they have been proven helpful in AMR-to-Text generation for English (KONSTAS *et al.*, 2017; LEE *et al.*, 2022). Finally, exploring few-shot approaches and comparing them with the methods reviewed in this thesis would be another interesting topic to be investigated.

7.4 Publications: Published and Submitted

Overall, 8 papers were written along this work, 5 have been published, 1 has been accepted for publication, 1 is in the review process in a journal, and 1 have been submitted to a Natural Language Processing journal. Table 5 presents in chronological order all the papers published and submitted during this research.

In addition, Table 6 presents a list of publications closely related to this thesis.

Table 5 – List of papers published/submitted to conferences and journals.

Publication
CABEZUDO, M. A. S.; PARDO, T. Natural language generation: Recently learned lessons, directions for semantic representation-based approaches, and the case of Brazilian Portuguese language. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. Florence, Italy: Association for Computational Linguistics, 2019. p. 81–88.
CABEZUDO, M. A. S.; PARDO, T. Towards a General Abstract Meaning Representation Corpus for Brazilian Portuguese. In: Proceedings of the 13th Linguistic Annotation Workshop. Florence, Italy: Association for Computational Linguistics, 2019. p. 236–244.
CABEZUDO, M. A. S.; MILLE, S.; PARDO, T. 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). Hong Kong, China: Association for Computational Linguistics, 2019. p. 94–103.
CABEZUDO, M. A. S.; PARDO, T. Low-resource AMR-to-Text Generation: A Study on Brazilian Portuguese. Procesamiento del Lenguaje Natural, Vol 68. p. 85-97. Sociedad Española para el Procesamiento del Lenguaje Natural. 2022.
INÁCIO, M. L.; **CABEZUDO, M. A. S.**; RAMISCH, R.; DI FELIPPO, A.; PARDO, T. A. S. The AMR-PT corpus and the semantic annotation of challenging sentences from journalistic and opinion texts. In SciELO Preprints. https://doi.org/10.1590/1678-460x202255159. 2022.
CABEZUDO, M. A. S.; ANCHIÊTA, R.T.; PARDO, T. Comparison of Cross-lingual strategies for AMR-to-Brazilian Portuguese Generation, submitted to the Language Resources and Evaluation journal, 2022.
CABEZUDO, M. A. S.; PARDO, T. Exploring a POS-based Two-stage Approach for Improving Low-Resource AMR-to-Text Generation, accepted at the Generation, Evaluation and Metrics workshop (GEM) at Empirical Methods in Natural Language Processing, 2022.
CABEZUDO, M. A. S.; INÁCIO, M. L.; PARDO, T. Investigating Paraphrase Generation as a Data Augmentation Strategy for Low-Resource AMR-to-Text Generation, submitted to the Northern European Journal of Language Technology, 2023.

Table 6 – List of additional papers published/submitted to conferences and journals.

Publication
CABEZUDO, M. A. S.; PARDO, T. NILC-SWORNEMO at the Surface Realization Shared Task: Exploring Syntax-Based Word Ordering using Neural Models. In Proceedings of the First Workshop on Multilingual Surface Realisation. Melbourne, Australia. Association for Computational Linguistics, 2018. p. 58-64.
MONSALVE, F.; RIVAS ROJAS, K.; **CABEZUDO, M.A.S.**; ONCEVAY, A. 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. Florence, Italy: Association for Computational Linguistics, 2019. p. 81–89.
ANCHIÊTA, R. T.; **CABEZUDO, M. A. S.**; PARDO, T. A. S. SEMA: An extended semantic evaluation for AMR. Proceedings of the 20th Computational Linguistics and Intelligent Text Processing. Springer International Publishing. 2019.
CABEZUDO, M.A.S.; PARDO T.A.S. NILC at SR'20: Exploring Pre-Trained Models in Surface Realisation. In Proceedings of the Third Workshop on Multilingual Surface Realisation. Online: Association for Computational Linguistics, 2020. p. 50–56.
CABEZUDO, M.A.S.; PARDO T.A.S. NILC at WebNLG+: Pretrained Sequence-to-Sequence Models on RDF-to-Text Generation. In Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+). Online: Association for Computational Linguistics, 2020. p. 131–136.

BIBLIOGRAPHY

ABEND, O.; RAPPOPORT, A. Ucca: A semantics-based grammatical annotation scheme. In: **Proccedings of the 10th International Conference on Computational Semantics**. Potsdam, Germany: Association for Computer Linguistics, 2013. p. 1–12. Citation on page 24.

____. The state of the art in semantic representation. In: **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics**. Vancouver, Canada: Association for Computer Linguistics, 2017. p. 77–89. Citation on page 23.

AFONSO, S.; BICK, E.; HABER, R.; SANTOS, D. Floresta sintá(c)tica: A treebank for portuguese. In: **Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)**. [S.l.]: European Language Resources Association (ELRA), 2002. Citations on pages 42 and 45.

AMR Biliography. **AMR Bibliography**. 2022. <https://nert-nlp.github.io/AMR-Bibliography/>. Accessed: 2022-01-20. Citation on page 25.

ANCHIÊTA, R.; PARDO, T. Towards AMR-BR: A SemBank for Brazilian Portuguese language. In: **Proceedings of the Eleventh International Conference on Language Resources and Evaluation**. Miyazaki, Japan: European Languages Resources Association, 2018. p. 974–979. Citations on pages 25, 26, 29, and 181.

____. Semantically inspired AMR alignment for the Portuguese language. In: **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. Online: Association for Computational Linguistics, 2020. p. 1595–1600. Citations on pages 29, 44, and 46.

Anchieta, R. T.; Cabezudo, M. A. S.; Pardo, T. A. S. SEMA: an Extended Semantic Evaluation Metric for AMR. **arXiv e-prints**, May 2019. Citation on page 43.

ANCHIÊTA, R. T.; PARDO, T. A. S. A rule-based amr parser for portuguese. In: SIMARI, G. R.; FERMÉ, E.; SEGURA, F. G.; MELQUIADES, J. A. R. (Ed.). **Advances in Artificial Intelligence - IBERAMIA 2018**. Cham: Springer International Publishing, 2018. p. 341–353. ISBN 978-3-030-03928-8. Citations on pages 29, 38, and 43.

____. Exploring the potentiality of semantic features for paraphrase detection. In: QUARESMA, P.; VIEIRA, R.; ALUÍSIO, S.; MONIZ, H.; BATISTA, F.; GONÇALVES, T. (Ed.). **Computational Processing of the Portuguese Language**. [S.l.]: Springer International Publishing, 2020. p. 228–238. Citation on page 24.

ANCHIêTA, R. T.; PARDO, T. A. S. Análise semântica com base em amr para o português. **Linguamática**, v. 14, n. 1, p. 33–48, Jul. 2022. Citations on pages 43 and 44.

AZIN, Z.; ERYIĞIT, G. Towards Turkish Abstract Meaning Representation. In: **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop**. Florence, Italy: Association for Computational Linguistics, 2019. p. 43–47. Citation on page 38.

BAI, X.; CHEN, Y.; ZHANG, Y. Graph pre-training for AMR parsing and generation. In: **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. Dublin, Ireland: Association for Computational Linguistics, 2022. p. 6001–6015. Citation on page 57.

BAI, X.; SONG, L.; ZHANG, Y. Online back-parsing for AMR-to-text generation. In: **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. Online: Association for Computational Linguistics, 2020. p. 1206–1219. Citations on pages 56, 57, and 59.

BAKER, C. F.; FILLMORE, C. J.; LOWE, J. B. The berkeley framenet project. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics**. Montreal, Quebec, Canada, 1998. p. 86–90. Citation on page 23.

BANARESCU, L.; BONIAL, C.; CAI, S.; GEORGESCU, M.; GRIFFITT, K.; HERMJAKOB, U.; KNIGHT, K.; KOEHN, P.; PALMER, M.; SCHNEIDER, N. Abstract meaning representation for sembanking. In: **Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse**. Sofia, Bulgaria: Association for Computational Linguistics, 2013. p. 178–186. Citations on pages 15, 24, 25, 28, 36, and 37.

BASILE, V.; BOS, J.; EVANG, K.; VENHUIZEN, N. Developing a large semantically annotated corpus. In: **Proceedings of the Eighth International Conference on Language Resources and Evaluation**. Istanbul, Turkey: ELRA, 2012. p. 3196–3200. Citation on page 24.

BECK, D.; HAFFARI, G.; COHN, T. Graph-to-sequence learning using gated graph neural networks. In: **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. Melbourne, Australia: Association for Computational Linguistics, 2018. p. 273–283. Citations on pages 26, 28, and 57.

BELTAGY, I.; LO, K.; COHAN, A. SciBERT: A pretrained language model for scientific text. In: **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**. Hong Kong, China: Association for Computational Linguistics, 2019. p. 3615–3620. Citation on page 41.

BENDER, E. **The BenderRule: On Naming the Languages We Study and Why It Matters**. 2019. <https://thegradient.pub/the-benderrule-on-naming-the-languages-we-study-and-why-it-matters/>. [Online; accessed 19-March-2021]. Citation on page 38.

BEVILACQUA, M.; BLLOSHMI, R.; NAVIGLI, R. One spring to rule them both: Symmetric amr semantic parsing and generation without a complex pipeline. **Proceedings of the AAAI Conference on Artificial Intelligence**, v. 35, n. 14, p. 12564–12573, May 2021. Citations on pages 56, 57, and 59.

BHATHIYA, H. S.; THAYASIVAM, U. Meta learning for few-shot joint intent detection and slot-filling. In: **Proceedings of the 2020 5th International Conference on Machine Learning Technologies**. New York, NY, USA: Association for Computing Machinery, 2020. (ICMLT 2020), p. 86–92. Citation on page 41.

BICK, E. The parsing system palavras: Automatic grammatical analysis of portuguese in a constraint grammar framework. 01 2000. Citations on pages 42, 43, and 45.

BLLOSHMI, R.; TRIPODI, R.; NAVIGLI, R. XL-AMR: Enabling cross-lingual AMR parsing with transfer learning techniques. In: **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. Online: Association for Computational Linguistics, 2020. p. 2487–2500. Citations on pages 25, 29, 38, and 40.

BOJANOWSKI, P.; GRAVE, E.; JOULIN, A.; MIKOLOV, T. Enriching word vectors with subword information. **Transactions of the Association for Computational Linguistics**, v. 5, p. 135–146, 2017. Citation on page 40.

BONIAL, C.; DONATELLI, L.; ABRAMS, M.; LUKIN, S. M.; TRATZ, S.; MARGE, M.; ARTSTEIN, R.; TRAUM, D.; VOSS, C. Dialogue-AMR: Abstract Meaning Representation for dialogue. In: **Proceedings of the 12th Language Resources and Evaluation Conference**. Marseille, France: European Language Resources Association, 2020. p. 684–695. Citation on page 24.

BOS, J. Expressive power of abstract meaning representations. **Computational Linguistics**, v. 42, n. 3, p. 527–535, 2016. Citation on page 24.

BROWN, T.; MANN, B.; RYDER, N.; SUBBIAH, M.; KAPLAN, J. D.; DHARIWAL, P.; NEELAKANTAN, A.; SHYAM, P.; SASTRY, G.; ASKELL, A.; AGARWAL, S.; HERBERT-VOSS, A.; KRUEGER, G.; HENIGHAN, T.; CHILD, R.; RAMESH, A.; ZIEGLER, D.; WU, J.; WINTER, C.; HESSE, C.; CHEN, M.; SIGLER, E.; LITWIN, M.; GRAY, S.; CHESS, B.; CLARK, J.; BERNER, C.; MCCANDLISH, S.; RADFORD, A.; SUTSKEVER, I.; AMODEI, D. Language models are few-shot learners. In: LAROCHELLE, H.; RANZATO, M.; HADSELL, R.; BALCAN, M.; LIN, H. (Ed.). **Advances in Neural Information Processing Systems**. [S.l.]: Curran Associates, Inc., 2020. v. 33, p. 1877–1901. Citation on page 180.

CABEZUDO, M. A. S.; PARDO, T. Natural language generation: Recently learned lessons, directions for semantic representation-based approaches, and the case of Brazilian Portuguese language. In: **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop**. Florence, Italy: Association for Computational Linguistics, 2019. p. 81–88. Citation on page 47.

CAI, D.; LAM, W. Graph transformer for graph-to-sequence learning. **Proceedings of the AAAI Conference on Artificial Intelligence**, v. 34, n. 05, p. 7464–7471, Apr. 2020. Citations on pages 56 and 57.

CAI, S.; KNIGHT, K. Smatch: an evaluation metric for semantic feature structures. In: **Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**. Sofia, Bulgaria: Association for Computational Linguistics, 2013. p. 748–752. Citation on page 43.

CAO, K.; CLARK, S. Factorising AMR generation through syntax. 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)**. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. p. 2157–2163. Citations on pages 44, 56, and 57.

CONDORI, R. E. L.; PARDO, T. A. S. Opinion summarization methods: Comparing and extending extractive and abstractive approaches. **Expert Systems with Applications**, v. 78, p. 124–134, 2017. ISSN 0957-4174. Citations on pages 22 and 33.

Corrêa Jr, E. A.; MARINHO, V. Q.; SANTOS, L. B. d.; BERTAGLIA, T. F. C.; TREVISO, M. V.; BRUM, H. B. Pelesent: Cross-domain polarity classification using distant supervision. **6th Brazilian Conference on Intelligent Systems (BRACIS)**, 2017. Citation on page 39.

DALE, R.; EUGENIO, B. D.; SCOTT, D. Introduction to the special issue on natural language generation. **Computational Linguistics**, v. 24, n. 3, p. 345–353, 1998. Citation on page 21.

DAMONTE, M.; COHEN, S. B. 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)**. New Orleans, Louisiana: Association for Computational Linguistics, 2018. p. 1146–1155. Citations on pages 25, 26, 29, 38, 40, 43, and 58.

____. Structural neural encoders for AMR-to-text generation. 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)**. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. p. 3649–3658. Citations on pages 56 and 57.

DEHOUCK, M.; GÓMEZ-RODRÍGUEZ, C. Data augmentation via subtree swapping for dependency parsing of low-resource languages. In: **Proceedings of the 28th International Conference on Computational Linguistics**. Barcelona, Spain (Online): International Committee on Computational Linguistics, 2020. p. 3818–3830. Citation on page 39.

DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. BERT: Pre-training of deep bidirectional transformers for language understanding. 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)**. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. p. 4171–4186. Citations on pages 27 and 41.

DURAN, M. S.; ALUÍSIO, S. M. Propbank-br: a brazilian treebank annotated with semantic role labels. In: **Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)**. Istanbul, Turkey: European Language Resources Association (ELRA), 2012. p. 1862–1867. Citations on pages 42 and 178.

DURAN, M. S.; ALUÍSIO, S. M. 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, *SEM 2015**. Denver, Colorado, USA.: Association for Computational Linguistics, 2015. p. 216–221. Citations on pages 38, 43, and 178.

DURAN, M. S.; MARTINS, J. P.; ALUÍSIO, S. M. Um repositório de verbos para a anotação de papéis semânticos disponível na web (a verb repository for semantic role labeling available in the web) [in portuguese]. In: **Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology**. [S.l.: s.n.], 2013. Citation on page 42.

DUŠEK, O.; NOVIKOVA, J.; RIESER, V. Findings of the E2E NLG Challenge. In: **Proceedings of the 11th International Conference on Natural Language Generation**. Tilburg, The Netherlands: [s.n.], 2018. Citation on page 24.

FADAEE, M.; BISAZZA, A.; MONZ, C. Data augmentation for low-resource neural machine translation. In: **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**. Vancouver, Canada: Association for Computational Linguistics, 2017. p. 567–573. Citation on page 39.

FAN, A.; GARDENT, C. Multilingual AMR-to-text generation. In: **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. Online: Association for Computational Linguistics, 2020. p. 2889–2901. Citations on pages 17, 29, 38, and 58.

FEI, H.; ZHANG, M.; JI, D. Cross-lingual semantic role labeling with high-quality translated training corpus. In: **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**. Online: Association for Computational Linguistics, 2020. p. 7014–7026. Citation on page 40.

FERREIRA, T. C.; CALIXTO, I.; WUBBEN, S.; KRAHMER, E. 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**. Santiago de Compostela, Spain: Association for Computational Linguistics, 2017. p. 1–10. Citations on pages 26, 27, and 28.

FERREIRA, T. C.; LEE, C. van der; MILTENBURG, E. van; KRAHMER, E. Neural data-to-text generation: A comparison between pipeline and end-to-end architectures. In: **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**. Hong Kong, China: Association for Computational Linguistics, 2019. p. 552–562. Citation on page 34.

FINN, C.; ABBEEL, P.; LEVINE, S. Model-agnostic meta-learning for fast adaptation of deep networks. In: PRECUP, D.; TEH, Y. W. (Ed.). **Proceedings of the 34th International Conference on Machine Learning**. [S.l.]: PMLR, 2017. (Proceedings of Machine Learning Research, v. 70), p. 1126–1135. Citation on page 41.

FLANIGAN, J.; DYER, C.; SMITH, N. A.; CARBONELL, J. Generation from Abstract Meaning Representation using tree transducers. In: **Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**. San Diego, California: Association for Computational Linguistics, 2016. p. 731–739. Citations on pages 26 and 56.

FLANIGAN, J.; THOMSON, S.; CARBONELL, J.; DYER, C.; SMITH, N. A. A discriminative graph-based parser for the abstract meaning representation. In: **Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics**. [S.l.]: Association for Computational Linguistics, 2014. p. 1426–1436. Citations on pages 44 and 58.

FREITAS, C.; ROCHA, P.; BICK, E. Floresta sintá(c)tica: Bigger, thicker and easier. In: TEIXEIRA, A.; LIMA, V. L. S. de; OLIVEIRA, L. C. de; QUARESMA, P. (Ed.). **Computational Processing of the Portuguese Language**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008. p. 216–219. ISBN 978-3-540-85980-2. Citation on page 45.

FRIEDRICH, A.; ADEL, H.; TOMAZIC, F.; HINGERL, J.; BENTEAU, R.; MARUSCZYK, A.; LANGE, L. The SOFC-exp corpus and neural approaches to information extraction in the materials science domain. In: **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**. Online: Association for Computational Linguistics, 2020. p. 1255–1268. Citation on page 41.

GAO, Q.; VOGEL, S. Parallel implementations of word alignment tool. In: **Software Engineering, Testing, and Quality Assurance for Natural Language Processing**. Stroudsburg, PA,

USA: Association for Computational Linguistics, 2008. (SETQA-NLP '08), p. 49–57. ISBN 978-1-932432-10-7. Citation on page 44.

GARDENT, C.; SHIMORINA, A.; NARAYAN, S.; PEREZ-BELTRACHINI, L. Creating training corpora for nlg micro-planning. In: **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. [S.l.]: Association for Computational Linguistics, 2017. p. 179–188. Citation on page 24.

GATT, A.; KRAHMER, E. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. **Journal of Artificial Intelligence Research**, AI Access Foundation, El Segundo, CA, USA, v. 61, n. 1, p. 65–170, Jan. 2018. ISSN 1076-9757. Citations on pages 22, 33, and 35.

GOODFELLOW, I. J.; POUGET-ABADIE, J.; MIRZA, M.; XU, B.; WARDE-FARLEY, D.; OZAIR, S.; COURVILLE, A.; BENGIO, Y. Generative adversarial nets. In: **Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2**. Cambridge, MA, USA: MIT Press, 2014. (NIPS'14), p. 2672–2680. Citation on page 41.

GRIEßHABER, D.; VU, N. T.; MAUCHER, J. Low-resource text classification using domain-adversarial learning. **Computer Speech Language**, v. 62, p. 101056, 2020. ISSN 0885-2308. Citation on page 41.

GU, J.; LU, Z.; LI, H.; LI, V. O. Incorporating copying mechanism in sequence-to-sequence learning. In: **Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. Berlin, Germany: Association for Computational Linguistics, 2016. p. 1631–1640. Citation on page 28.

GU, J.; WANG, Y.; CHEN, Y.; LI, V. O. K.; CHO, K. Meta-learning for low-resource neural machine translation. In: **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**. Brussels, Belgium: Association for Computational Linguistics, 2018. p. 3622–3631. Citation on page 41.

GULCEHRE, C.; AHN, S.; NALLAPATI, R.; ZHOU, B.; BENGIO, Y. Pointing the unknown words. In: **Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. Berlin, Germany: Association for Computational Linguistics, 2016. p. 140–149. Citation on page 28.

HARDY, H.; VLACHOS, A. Guided neural language generation for abstractive summarization using Abstract Meaning Representation. In: **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**. Brussels, Belgium: Association for Computational Linguistics, 2018. p. 768–773. Citations on pages 24 and 29.

HARTMANN, N. S.; DURAN, M. S.; ALUÍSIO, S. M. Automatic semantic role labeling on non-revised syntactic trees of journalistic texts. In: **Computational Processing of the Portuguese Language - 12th International Conference, PROPOR 2016, Tomar, Portugal, July 13-15, 2016, Proceedings**. [S.l.: s.n.], 2016. p. 202–212. Citation on page 43.

HEDDERICH, M. A.; LANGE, L.; ADEL, H.; STROTGEN, J.; KLAKOW, D. A survey on recent approaches for natural language processing in low-resource scenarios. **ArXiv**, abs/2010.12309, 2020. Citation on page 39.

HEINZERLING, B.; STRUBE, M. BPEmb: Tokenization-free pre-trained subword embeddings in 275 languages. In: **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**. Miyazaki, Japan: European Language Resources Association (ELRA), 2018. Citation on page 40.

HOYLE, A. M.; MARASOVIĆ, A.; SMITH, N. A. Promoting graph awareness in linearized graph-to-text generation. In: **Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021**. Online: Association for Computational Linguistics, 2021. p. 944–956. Citation on page 57.

INÁCIO, M.; PARDO, T. Semantic-based opinion summarization. In: **Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)**. Held Online: INCOMA Ltd., 2021. p. 619–628. Citation on page 24.

INÁCIO, M. L. **Sumarização de Opinião com base em Abstract Meaning Representation.** 2021. Citation on page 177.

ISSA, F.; DAMONTE, M.; COHEN, S. B.; YAN, X.; CHANG, Y. 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)**. New Orleans, Louisiana: Association for Computational Linguistics, 2018. p. 442–452. Citation on page 24.

JIN, L.; GILDEA, D. **AMR-to-Text Generation with Cache Transition Systems**. [S.l.]: arXiv, 2019. Citations on pages 56 and 57.

____. Generalized shortest-paths encoders for AMR-to-text generation. In: **Proceedings of the 28th International Conference on Computational Linguistics**. Barcelona, Spain (Online): International Committee on Computational Linguistics, 2020. p. 2004–2013. Citations on pages 56 and 57.

JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing (3rd Edition)**. USA: Prentice-Hall, Inc., 2020. ISBN 0131873210. Citation on page 21.

KARAMANOLAKIS, G.; HSU, D.; GRAVANO, L. Leveraging just a few keywords for fine-grained aspect detection through weakly supervised co-training. In: **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**. Hong Kong, China: Association for Computational Linguistics, 2019. p. 4611–4621. Citation on page 39.

KIM, J.-K.; KIM, Y.-B.; SARIKAYA, R.; FOSLER-LUSSIER, E. Cross-lingual transfer learning for POS tagging without cross-lingual resources. In: **Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing**. Copenhagen, Denmark: Association for Computational Linguistics, 2017. p. 2832–2838. Citation on page 41.

KIPPER-SCHULER, K. **VerbNet: A broad-coverage, comprehensive verb lexicon**. 146 p. Phd Thesis (PhD Thesis), 2005. Citation on page 43.

KOEHN, P. Europarl: A parallel corpus for statistical machine translation. In: **Proceedings of Machine Translation Summit X: Papers**. Phuket, Thailand: [s.n.], 2005. p. 79–86. Citation on page 58.

KONSTAS, I.; IYER, S.; YATSKAR, M.; CHOI, Y.; ZETTLEMOYER, L. Neural amr: Sequence-to-sequence models for parsing and generation. In: **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. Vancouver, Canada: Association for Computational Linguistics, 2017. p. 146–157. Citations on pages 26, 28, 57, and 181.

LAMPOURAS, G.; VLACHOS, A. Sheffield at semeval-2017 task 9: Transition-based language generation from amr. In: **Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)**. [S.l.]: Association for Computational Linguistics, 2017. p. 586–591. Citations on pages 26, 45, and 56.

LAVIE, A.; AGARWAL, A. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In: **Proceedings of the Second Workshop on Statistical Machine Translation**. Stroudsburg, PA, USA: Association for Computational Linguistics, 2007. (StatMT '07), p. 228–231. Citation on page 35.

LEE, C. van der; FERREIRA, T. C.; EMMERY, C.; WILTSHIRE, T.; KRAHMER, E. **Neural Data-to-Text Generation Based on Small Datasets: Comparing the Added Value of Two Semi-Supervised Learning Approaches on Top of a Large Language Model**. [S.l.]: arXiv, 2022. Citation on page 181.

LEE, J.; YOON, W.; KIM, S.; KIM, D.; KIM, S.; SO, C. H.; KANG, J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. **Bioinformatics**, 09 2019. ISSN 1367-4803. Citation on page 41.

LEE, S. Natural language generation for electronic health records. **npj Digital Medicine**, v. 1, 12 2018. Citation on page 22.

LEWIS, M.; LIU, Y.; GOYAL, N.; GHAZVININEJAD, M.; MOHAMED, A.; LEVY, O.; STOYANOV, V.; ZETTLEMOYER, L. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**. Online: Association for Computational Linguistics, 2020. p. 7871–7880. Citations on pages 27, 28, 56, and 59.

LI, B.; WEN, Y.; QU, W.; BU, L.; XUE, N. Annotating the little prince with chinese amrs. In: **Proceedings of the 10th Linguistic Annotation Workshop**. Berlin, Germany: Association for Computational Linguistics, 2016. p. 7–15. Citation on page 38.

LI, B.; WEN, Y.; SONG, L.; QU, W.; XUE, N. Building a Chinese AMR bank with concept and relation alignments. In: **Linguistic Issues in Language Technology, Volume 18, 2019 - Exploiting Parsed Corpora: Applications in Research, Pedagogy, and Processing**. [S.l.]: CSLI Publications, 2019. Citation on page 38.

LIAO, K.; LEBANOFF, L.; LIU, F. Abstract Meaning Representation for multi-document summarization. In: **Proceedings of the 27th International Conference on Computational Linguistics**. Santa Fe, New Mexico, USA: Association for Computational Linguistics, 2018. p. 1178–1190. Citation on page 24.

LINH, H.; NGUYEN, H. A case study on meaning representation for Vietnamese. In: **Proceedings of the First International Workshop on Designing Meaning Representations**. Florence, Italy: Association for Computational Linguistics, 2019. p. 148–153. Citation on page 38.

LIU, F.; FLANIGAN, J.; THOMSON, S.; SADEH, N. M.; SMITH, N. A. Toward abstractive summarization using semantic representations. In: **Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**. Denver, Colorado: The Association for Computational Linguistics, 2015. p. 1077–1086. Citation on page 23.

MAGER, M.; ASTUDILLO, R. F.; NASEEM, T.; SULTAN, M. A.; LEE, Y.-S.; FLORIAN, R.; ROUKOS, S. 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**. Online: Association for Computational Linguistics, 2020. p. 1846–1852. Citations on pages 26, 27, 28, 56, 57, and 59.

MANNING, E. A partially rule-based approach to AMR generation. In: **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop**. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. p. 61–70. Citations on pages 56 and 57.

MARCUS, M. P.; MARCINKIEWICZ, M. A.; SANTORINI, B. Building a large annotated corpus of english: The penn treebank. **Computational linguistics**, MIT Press, v. 19, n. 2, p. 313–330, 1993. Citation on page 42.

MATTHIESSEN, C.; BATEMAN, J. A. **Text generation and systemic-functional linguistics: experiences from English and Japanese**. [S.l.]: Pinter Publishers, 1991. Citations on pages 24 and 36.

MAY, J. Semeval-2016 task 8: Meaning representation parsing. In: **Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016**. [S.l.: s.n.], 2016. p. 1063–1073. Citation on page 37.

MAY, J.; PRIYADARSHI, J. SemEval-2017 task 9: Abstract Meaning Representation parsing and generation. In: **Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)**. Vancouver, Canada: Association for Computational Linguistics, 2017. p. 536–545. Citations on pages 25, 35, and 37.

MCDONALD, R.; NIVRE, J.; QUIRMBACH-BRUNDAGE, Y.; GOLDBERG, Y.; DAS, D.; GANCHEV, K.; HALL, K.; PETROV, S.; ZHANG, H.; TÄCKSTRÖM, O.; BEDINI, C.; CASTELLÓ, N. B.; LEE, J. Universal Dependency annotation for multilingual parsing. In: **Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**. Sofia, Bulgaria: Association for Computational Linguistics, 2013. p. 92–97. Citation on page 45.

MI, F.; HUANG, M.; ZHANG, J.; FALTINGS, B. Meta-learning for low-resource natural language generation in task-oriented dialogue systems. In: **Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19**. [S.l.]: International Joint Conferences on Artificial Intelligence Organization, 2019. p. 3151–3157. Citation on page 41.

MIGUELES-ABRAIRA, N.; AGERRI, R.; ILARRAZA, A. Diaz de. Annotating abstract meaning representations for Spanish. In: **Proceedings of the 11th International Conference on Language Resources and Evaluation**. Miyazaki, Japan: European Languages Resources Association, 2018. p. 3074–3078. Citations on pages 25 and 38.

MIKOLOV, T.; SUTSKEVER, I.; CHEN, K.; CORRADO, G. S.; DEAN, J. Distributed representations of words and phrases and their compositionality. In: BURGES, C. J. C.; BOTTOU, L.; WELLING, M.; GHAHRAMANI, Z.; WEINBERGER, K. Q. (Ed.). **Advances in Neural Information Processing Systems**. [S.l.]: Curran Associates, Inc., 2013. v. 26. Citation on page 40.

MILLE, S.; BELZ, A.; BOHNET, B.; WANNER, L. Underspecified Universal Dependency structures as inputs for multilingual surface realisation. In: **Proceedings of the 11th International Conference on Natural Language Generation**. Tilburg University, The Netherlands: Association for Computational Linguistics, 2018. p. 199–209. Citation on page 45.

MILLE, S.; CARLINI, R.; BURGA, A.; WANNER, L. FORGe at SemEval-2017 task 9: Deep sentence generation based on a sequence of graph transducers. In: **Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)**. Vancouver, Canada: Association for Computational Linguistics, 2017. p. 920–923. Citations on pages 45 and 56.

MIRANDA-JIMÉNEZ, S.; GELBUKH, A.; SIDOROV, G. Conceptual graphs as framework for summarizing short texts. **International Journal of Conceptual Structures and Smart Applications (IJCSSA)**, IGI Global, v. 2, n. 2, p. 55–75, 2014. Citation on page 23.

MONSALVE, F.; ROJAS, K. R.; CABEZUDO, M. A. S.; ONCEVAY, A. 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**. Florence, Italy: Association for Computational Linguistics, 2019. p. 81–89. Citation on page 40.

NIVRE, J.; MARNEFFE, M.-C. de; GINTER, F.; GOLDBERG, Y.; HAJIČ, J.; MANNING, C. D.; MCDONALD, R.; PETROV, S.; PYYSALO, S.; SILVEIRA, N.; TSARFATY, R.; ZEMAN, D. Universal Dependencies v1: A multilingual treebank collection. In: **Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)**. Portorož, Slovenia: European Language Resources Association (ELRA), 2016. p. 1659–1666. Citation on page 45.

NOORD, R. van; BOS, J. Neural semantic parsing by character-based translation: Experiments with abstract meaning representations. **Computational Linguistics in the Netherlands Journal**, v. 7, p. 93–108, Dec. 2017. Citation on page 43.

NOVIKOVA, J.; DUŠEK, O.; RIESER, V. The E2E dataset: New challenges for end-to-end generation. In: **Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue**. Saarbrücken, Germany: Association for Computational Linguistics, 2017. p. 201–206. Citation on page 22.

PALMER, M.; GILDEA, D.; KINGSBURY, P. The proposition bank: An annotated corpus of semantic roles. **Computational Linguistics.**, MIT Press, Cambridge, MA, USA, v. 31, n. 1, p. 71–106, Mar. 2005. ISSN 0891-2017. Citations on pages 23, 36, and 42.

PAPINENI, K.; ROUKOS, S.; WARD, T.; ZHU, W.-J. Bleu: A method for automatic evaluation of machine translation. In: **Proceedings of the 40th Annual Meeting on Association for Computational Linguistics**. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002. (ACL '02), p. 311–318. Citation on page 35.

PLANK, B.; AGIĆ, Ž. Distant supervision from disparate sources for low-resource part-of-speech tagging. In: **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**. Brussels, Belgium: Association for Computational Linguistics, 2018. p. 614–620. Citation on page 40.

PONCELAS, A.; WAY, A. Selecting artificially-generated sentences for fine-tuning neural machine translation. In: **Proceedings of the 12th International Conference on Natural Language Generation**. Tokyo, Japan: Association for Computational Linguistics, 2019. p. 219–228. Citation on page 29.

POPOVIĆ, M. chrF++: words helping character n-grams. In: **Proceedings of the Second Conference on Machine Translation**. Copenhagen, Denmark: Association for Computational Linguistics, 2017. p. 612–618. Citation on page 35.

POURDAMGHANI, N.; GAO, Y.; HERMJAKOB, U.; KNIGHT, K. Aligning english strings with abstract meaning representation graphs. In: **Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. [S.l.]: Association for Computational Linguistics, 2014. p. 425–429. Citations on pages 44 and 46.

POURDAMGHANI, N.; KNIGHT, K.; HERMJAKOB, U. Generating english from abstract meaning representations. In: **Proceedings of the Ninth International Natural Language Generation Conference**. Edinburgh, UK: Association for Computational Linguistics, 2016. p. 21–25. Citation on page 26.

RADEMAKER, A.; CHALUB, F.; REAL, L.; FREITAS, C.; BICK, E.; PAIVA, V. de. Universal Dependencies for Portuguese. In: **Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)**. Pisa,Italy: Linköping University Electronic Press, 2017. p. 197–206. Citation on page 45.

RADFORD, A.; WU, J.; CHILD, R.; LUAN, D.; AMODEI, D.; SUTSKEVER, I. Language models are unsupervised multitask learners. 2019. Citations on pages 27, 28, and 56.

RAFFEL, C.; SHAZEER, N.; ROBERTS, A.; LEE, K.; NARANG, S.; MATENA, M.; ZHOU, Y.; LI, W.; LIU, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. **Journal of Machine Learning Research**, v. 21, n. 140, p. 1–67, 2020. Available: <http://jmlr.org/papers/v21/20-074.html>. Citations on pages 27, 28, 56, and 59.

RAO, S.; MARCU, D.; KNIGHT, K.; III, H. D. Biomedical event extraction using Abstract Meaning Representation. In: **BioNLP 2017**. Vancouver, Canada,: Association for Computational Linguistics, 2017. p. 126–135. Citation on page 24.

REGATTE, Y. R.; GANGULA, R. R. R.; MAMIDI, R. Dataset creation and evaluation of aspect based sentiment analysis in Telugu, a low resource language. In: **Proceedings of the 12th Language Resources and Evaluation Conference**. Marseille, France: European Language Resources Association, 2020. p. 5017–5024. ISBN 979-10-95546-34-4. Citation on page 40.

REITER, E.; DALE, R. **Building Natural Language Generation Systems**. New York, NY, USA: Cambridge University Press, 2000. Citations on pages 21, 22, 33, and 34.

RIBEIRO, L. F. R.; GARDENT, C.; GUREVYCH, I. Enhancing AMR-to-text generation with dual graph representations. In: **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural**

Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, 2019. p. 3183–3194. Citations on pages 26, 28, 56, and 57.

RIBEIRO, L. F. R.; PFEIFFER, J.; ZHANG, Y.; GUREVYCH, I. Smelting gold and silver for improved multilingual AMR-to-Text generation. In: **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021. p. 742–750. Citation on page 58.

RIBEIRO, L. F. R.; SCHMITT, M.; SCHÜTZE, H.; GUREVYCH, I. Investigating pretrained language models for graph-to-text generation. In: **Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI**. Online: Association for Computational Linguistics, 2021. p. 211–227. Citations on pages 26, 28, 56, 57, and 59.

RIBEIRO, L. F. R.; ZHANG, Y.; GUREVYCH, I. Structural adapters in pretrained language models for AMR-to-Text generation. In: **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021. p. 4269–4282. Citations on pages 28, 57, and 59.

SELLAM, T.; DAS, D.; PARIKH, A. BLEURT: Learning robust metrics for text generation. In: **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**. Online: Association for Computational Linguistics, 2020. p. 7881–7892. Citation on page 35.

SENNRICH, R.; HADDOW, B.; BIRCH, A. Neural machine translation of rare words with subword units. In: **Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. Berlin, Germany: Association for Computational Linguistics, 2016. p. 1715–1725. Citations on pages 28 and 39.

SENO, E.; CASELI, H.; INáCIO, M.; ANCHIêTA, R.; RAMISCH, R. Xpta: um parser amr para o português baseado em uma abordagem entre línguas. **Linguamática**, v. 14, n. 1, p. 49–68, 2022. Citation on page 43.

SILVA, J.; BRANCO, A.; GONÇALVES, P. Top-performing robust constituency parsing of Portuguese: Freely available in as many ways as you can get it. In: **Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)**. Valletta, Malta: European Language Resources Association (ELRA), 2010. Citation on page 45.

SNOVER, M.; DORR, B.; SCHWARTZ, R.; MICCIULLA, L.; MAKHOUL, J. A study of translation edit rate with targeted human annotation. In: **In Proceedings of Association for Machine Translation in the Americas**. [S.l.: s.n.], 2006. p. 223–231. Citation on page 35.

SONG, L.; GILDEA, D.; ZHANG, Y.; WANG, Z.; SU, J. Semantic neural machine translation using amr. **Transactions of the Association for Computational Linguistics**, v. 7, p. 19–31, 2019. Citations on pages 24 and 29.

SONG, L.; PENG, X.; ZHANG, Y.; WANG, Z.; GILDEA, D. AMR-to-text generation with synchronous node replacement grammar. In: **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**. Vancouver, Canada: Association for Computational Linguistics, 2017. p. 7–13. Citation on page 26.

SONG, L.; ZHANG, Y.; WANG, Z.; GILDEA, D. A graph-to-sequence model for AMR-to-text generation. In: **Proceedings of the 56th Annual Meeting of the Association for Computational**

Linguistics (Volume 1: Long Papers). Melbourne, Australia: Association for Computational Linguistics, 2018. p. 1616–1626. Citations on pages 26, 28, and 57.

SOTO, X.; SHTERIONOV, D.; PONCELAS, A.; WAY, A. Selecting backtranslated data from multiple sources for improved neural machine translation. In: **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**. Online: Association for Computational Linguistics, 2020. p. 3898–3908. Citation on page 29.

STRAKA, M.; HAJIČ, J.; STRAKOVÁ, J. UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In: **Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)**. Portorož, Slovenia: European Language Resources Association (ELRA), 2016. p. 4290–4297. Citation on page 45.

TAHER, E.; HOSEINI, S. A.; SHAMSFARD, M. Beheshti-NER: Persian named entity recognition using BERT. In: **Proceedings of The First International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2019) co-located with ICNLSP 2019 - Short Papers**. Trento, Italy: Association for Computational Linguistics, 2019. p. 37–42. Citation on page 41.

TIEDEMANN, J. Parallel data, tools and interfaces in OPUS. In: **Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)**. Istanbul, Turkey: European Language Resources Association (ELRA), 2012. p. 2214–2218. Citation on page 40.

UCHIDA, H.; ZHU, M.; SENTA, T. D. UNL: Universal networking language–an electronic language for communication, understanding, and collaboration. **Tokyo: UNU/IAS/UNL Center**, 1996. Citation on page 23.

UREšOVá, Z.; HAJIč, J.; BOJAR, O. Comparing czech and english amrs. In: **Proceedings of Workshop on Lexical and Grammatical Resources for Language Processing (LG-LP 2014, at Coling 2014)**. Dublin, Ireland: Association for Computational Linguistics, 2014. p. 55–64. ISBN 978-1-873769-44-7. Citation on page 38.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, L. u.; POLOSUKHIN, I. Attention is all you need. In: GUYON, I.; LUXBURG, U. V.; BENGIO, S.; WALLACH, H.; FERGUS, R.; VISHWANATHAN, S.; GARNETT, R. (Ed.). **Advances in Neural Information Processing Systems**. [S.l.]: Curran Associates, Inc., 2017. v. 30. Citations on pages 40 and 56.

VICENTE, M. E.; BARROS, C.; AGULLÓ, F.; PEREGRINO, F. S.; LLORET, E. La generacion de lenguaje natural: análisis del estado actual. **Computación y Sistemas**, v. 19, n. 4, p. 721–756, 2015. Citations on pages 22 and 33.

VILCA, G. C. V.; CABEZUDO, M. A. S. A study of abstractive summarization using semantic representations and discourse level information. In: SPRINGER. **Proceedings of the 20th International Conference on Text, Speech, and Dialogue**. Prague, Czech Republic, 2017. p. 482–490. Citation on page 33.

WANG, C.; XUE, N.; PRADHAN, S. A transition-based algorithm for amr parsing. In: **Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**. Denver, Colorado: Association for Computational Linguistics, 2015. p. 366–375. Citation on page 43.

WANG, T.; WAN, X.; JIN, H. AMR-To-Text Generation with Graph Transformer. **Transactions of the Association for Computational Linguistics**, v. 8, p. 19–33, 01 2020. ISSN 2307-387X. Citations on pages 56 and 57.

WANG, T.; WAN, X.; YAO, S. Better amr-to-text generation with graph structure reconstruction. In: BESSIERE, C. (Ed.). **Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20**. [S.l.]: International Joint Conferences on Artificial Intelligence Organization, 2020. p. 3919–3925. Citations on pages 56, 57, and 59.

WANNER, L.; BOSCH, H.; BOUAYAD-AGHA, N.; CASAMAYOR, G.; ERTL, T.; HILBRING, D.; JOHANSSON, L.; KARATZAS, K.; KARPPINEN, A.; KOMPATSIARIS, I. *et al.* Getting the environmental information across: from the web to the user. **Expert Systems**, Wiley Online Library, v. 32, n. 3, p. 405–432, 2015. Citation on page 33.

WEIZENBAUM, J. Eliza—a computer program for the study of natural language communication between man and machine. **Communications ACM**, Association for Computing Machinery, New York, NY, USA, v. 9, n. 1, p. 36–45, Jan. 1966. ISSN 0001-0782. Citations on pages 15 and 21.

XU, D.; LI, J.; ZHU, M.; ZHANG, M.; ZHOU, G. XLPT-AMR: Cross-lingual pre-training via multi-task learning for zero-shot AMR parsing and text generation. In: **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**. Online: Association for Computational Linguistics, 2021. p. 896–907. Citation on page 58.

XUE, N.; BOJAR, O.; HAJIČ, J.; PALMER, M.; UREŠOVÁ, Z.; ZHANG, X. Not an interlingua, but close: Comparison of English AMRs to Chinese and Czech. In: **Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)**. Reykjavik, Iceland: European Language Resources Association (ELRA), 2014. p. 1765–1772. Citations on pages 26 and 38.

ZEMAN, D.; POPEL, M.; STRAKA, M.; HAJIČ, J.; NIVRE, J.; GINTER, F.; LUOTOLAHTI, J.; PYYSALO, S.; PETROV, S.; POTTHAST, M.; TYERS, F.; BADMAEVA, E.; GOKIRMAK, M.; NEDOLUZHKO, A.; CINKOVÁ, S.; JR., J. H.; HLAVÁČOVÁ, J.; KETTNEROVÁ, V.; UREŠOVÁ, Z.; KANERVA, J.; OJALA, S.; MISSILÄ, A.; MANNING, C. D.; SCHUSTER, S.; REDDY, S.; TAJI, D.; HABASH, N.; LEUNG, H.; MARNEFFE, M.-C. de; SANGUINETTI, M.; SIMI, M.; KANAYAMA, H.; PAIVA, V. de; DROGANOVA, K.; ALONSO, H. M.; ÇÖLTEKIN, Ç.; SULUBACAK, U.; USZKOREIT, H.; MACKETANZ, V.; BURCHARDT, A.; HARRIS, K.; MARHEINECKE, K.; REHM, G.; KAYADELEN, T.; ATTIA, M.; ELKAHKY, A.; YU, Z.; PITLER, E.; LERTPRADIT, S.; MANDL, M.; KIRCHNER, J.; ALCALDE, H. F.; STRNADOVÁ, J.; BANERJEE, E.; MANURUNG, R.; STELLA, A.; SHIMADA, A.; KWAK, S.; MENDONÇA, G.; LANDO, T.; NITISAROJ, R.; LI, J. CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies. In: **Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies**. Vancouver, Canada: Association for Computational Linguistics, 2017. p. 1–19. Citation on page 45.

ZHANG, Y.; GUO, Z.; TENG, Z.; LU, W.; COHEN, S. B.; LIU, Z.; BING, L. Lightweight, dynamic graph convolutional networks for AMR-to-text generation. In: **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. Online: Association for Computational Linguistics, 2020. p. 2162–2172. Citations on pages 26, 35, 56, and 57.

ZHAO, Y.; CHEN, L.; CHEN, Z.; CAO, R.; ZHU, S.; YU, K. Line graph enhanced AMR-to-text generation with mix-order graph attention networks. In: **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**. Online: Association for Computational Linguistics, 2020. p. 732–741. Citations on pages 56 and 57.

ZHU, J.; LI, J.; ZHU, M.; QIAN, L.; ZHANG, M.; ZHOU, G. Modeling graph structure in transformer for better AMR-to-text generation. In: **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**. Hong Kong, China: Association for Computational Linguistics, 2019. p. 5459–5468. Citations on pages 26, 28, 56, and 57.

ZHU, Y.; HEINZERLING, B.; VULIĆ, I.; STRUBE, M.; REICHART, R.; KORHONEN, A. On the importance of subword information for morphological tasks in truly low-resource languages. In: **Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)**. Hong Kong, China: Association for Computational Linguistics, 2019. p. 216–226. Citation on page 40.


ICMC
USP
SÃO CARLOS