Imaginary Numbers! Evaluating Numerical Referring Expressions by Neural End-to-End Surface Realization Systems

Rossana Cunha, Osuji Cynthia Chinonso, João Gabriel Moura Campos, Brian Timoney, Brian Davis, Fabio Cozman, Adriana Pagano and Thiago Castro Ferreira
Arts Faculty, Federal University Minas Gerais, Brazil
Adapt Research Centre, Dublin City University, Ireland
Escola Politécnica, University of São Paulo, Brazil
rossanacunha@ufmg.br chinonso.osuji@adaptcentre.ie joaogcampos@usp.br
brian.timoney3@mail.dcu.ie brian.davis@adaptcentre.ie fgcozman@usp.br
apagano@ufmg.br thiagocf05@ufmg.br

Abstract

Neural end-to-end surface realizers output more fluent texts than classical architectures. However, they tend to suffer from adequacy problems, in particular *hallucinations* in numerical referring expression generation. This poses a problem to language generation in sensitive domains, as is the case of robot journalism covering COVID-19 and Amazon deforestation. We propose an approach whereby numerical referring expressions are converted from digits to plain word form descriptions prior to being fed to state-of-the-art Large Language Models. We conduct automatic and human evaluations to report the best strategy to numerical superficial realization. Code and data are publicly available[1].

1 Introduction

The significant advances in deep learning for NLP and its enormous success in other text generation tasks, such as machine translation (Akhbardeh et al., 2021). As a result, approaches to surface realization of *data-to-text* systems have moved from traditional modular pipeline architectures (Reiter and Dale, 2000) to end-to-end ones. These systems transform a simple meaning representation into text without any explicit intermediate representations (Wen et al., 2015; Dušek and Jurčíček, 2016; Lebret et al., 2016; Gehrmann et al., 2018). While early neural data-to-text systems required a high amount of parallel training data, current state-of-the-art (SOTA) architectures, known as Large Language Models (LLMs) (Radford et al., 2019; Lewis et al., 2020; Raffel et al., 2020a), can deliver impressive results with less training, even excelling in zero-short or few-shot settings.

With respect to linguistic output, neural end-to-end surface realizers appear to generate more fluent text than classical pipeline architectures but are more likely to suffer from (semantic) adequacy problems, in particular, *hallucinations* (Ji et al., 2023), whereby the system produces text that contains information which is not present in the input representation. A particular hallucination problem that modern approaches seem to struggle with, unlike classical architectures, is numerical referring expression generation (Puduppully and Lapata, 2021; Wallace et al., 2019; Ji et al., 2023). For instance, let's hypothesize the case where a surface realizer produces the outcome: *"The country registered **458,098** cases of COVID-19"*, whereas the gold-standard reference points to *"The country registered **408,098** cases of COVID-19"*. Albeit there is only a single-digit difference between both texts (which can be overlooked by popular automatic quality metrics), the difference represents an arithmetic change of 50,000 and may lead readers to make drastic errors given the sensitivity of the context.

To the best of our knowledge, this problem has never been investigated in surface realization systems, despite having been addressed in other generation tasks such as text normalization (Zhang et al., 2019; Sproat, 2022), question-answering (Chen et al., 2021; Kim et al., 2022), and *text-to-speech* (Nikulásdóttir and Guðnason, 2019); tasks which also struggle to synthesize texts with numerical referring expressions represented by digits. One approach to circumvent the problem in *text-to-speech* systems is to normalise the input texts by converting numerical referring expressions from digits to plain word form descriptions prior to being fed into the system (Nikulásdóttir and Guðnason, 2019). Another technique used in Referring Expression Generation (REG) systems is slot-filling or *delexicalisation* where values like date, number, or constants are represented as a literal (Castro Ferreira et al., 2018; Cunha et al., 2020).

In the context of end-to-end surface realizers, this study raises two questions:

[1] https://github.com/BotsDoBem/LargeLM

Proceedings of the Fifth Workshop on Insights from Negative Results in NLP, pages 73–81
June 20, 2024 ©2024 Association for Computational Linguistics

	B. Portuguese			English		
	Train	Dev	Test	Train	Dev	Test
Daily Deforestation	4,062	504	484	3,874	452	462
Month Deforestation	324	20	22	456	36	26
Daily Fire	942	108	108	–	–	–
COVID-19	1,064	122	108	–	–	–
Total	6,392	754	722	4,330	488	488

Table 1: Data Statistics.

INPUT

[DEFORESTATION_MONTH][INTENTS] TOTAL_DEFORESTATION (area="322.91"", location="deter-amz", month="4", year="2021") [HISTORY] [PARAGRAPH]

PORTUGUESE OUTPUT

O Instituto Nacional de Pesquisa Espaciais (INPE) informou que foram desmatados 322.91 km² na Amazônia Legal, em abril de 2021.

ENGLISH OUTPUT

The National Institute for Space Research (INPE) detected 322.91 sq km of deforestation in the Legal Amazon in April 2021.

Figure 1: Example of Portuguese and English Meaning Representation inputs and their corresponding outputs.

(RQ1) *How well do state-of-the-art end-to-end surface realizers generate numerical referring expressions?*
(RQ2) *Are numerical referring expressions better verbalized when represented by digits or text (spell-out form)?*

To answer these questions, we conducted automatic and human evaluations with three SOTA LLMs: GPT-2 (Radford et al., 2019), BART (Lewis et al., 2020), T5 (Raffel et al., 2020b), and their multilingual counterparts. These models were used to verbalize English and Brazilian Portuguese news about Amazon Deforestation, Fire Alerts, and COVID-19 cases using **four** different strategies, which we discuss in Section 3. Code and data will be publicly available.

2 Data

For training and evaluating the models, we used automatic-generated reports by *BotsDoBem*, a group of Twitter robot-journalists such as CoronaReporter[2] and DaMata[3], which publish news in Brazilian Portuguese and English. For Brazilian Portuguese, the dataset comprises of i) both daily and monthly reports on deforestation in the Legal Amazon area of Brazil (Rosa Teixeira et al., 2020), ii) daily reports about Fires in the Brazilian Biomes, as well as iii) COVID-19 cases in the country (Campos et al., 2020). For English, the dataset comprises of daily and monthly reports on deforestation in the Legal Amazon. Although automatically generated, these texts contain a high number of numerical referring expressions, making them suitable for our goal of evaluating how well neural end-to-end surface realizers generate numerical referring expressions. Table 1 introduces the number of instances per language and domain, split into training, development, and test sets. Each instance in the corpus consists of a meaning representation and a corresponding gold-standard verbalization in Brazilian Portuguese or English representing the sentence of a report. For both languages, the verbalizations were automatically generated by the pipeline system described in Rosa Teixeira et al. (2020) and Campos et al. (2020).

Figure 1 illustrates the structure of instances in both the English and Portuguese datasets, which consist of meaning representations starting with a tag representing the report domain, followed by a tag that marks the beginning of the sentence intents (e.g., `INTENTS`). Each intent in the meaning representation follows the *intent-attribute-value* schema. Finally, the tag `[HISTORY]` marks where the verbalization of the previous sentences in the paragraph of the target will be depicted. In the example, the tag `[PARAGRAPH]` means that the target sentence is at the beginning of the paragraph.

3 Numerical Referring Expressions

To evaluate the effectiveness of a neural end-to-end surface realizer in generating numerical expressions, we consider **two forms** of number representation: digits and word (*spell-out*) form descriptions. These are assessed in both the meaning representations and the verbalizations, resulting in a total of **four** distinct strategies:

1. Numbers represented by digits in the meaning representation and the reference texts (*no desc*);
2. Numbers are described in the input meaning representation in spell-out form and digits in the target references (*desc src*);
3. Numbers represented by digits in the meaning representations and spell-out form descriptions in the target references (*desc trg*); and
4. Numbers are described in a spell-out form in both the input meaning representations and target references (*desc*).

To exemplify, Table 2 depicts the **four** strategies of a pair of meaning representations and their corresponding English verbalizations. We utilized

[2] `https://twitter.com/CoronaReporter`
[3] `https://twitter.com/DaMataReporter`

	Numeric Referring Expressions			
Strategies	**Area**	**Month**	**Year**	**Input MR**
no desc	322.91	4	2021	In April 2021, 322.91 sq km of the Legal Amazon were deforested, according to data from the National Institute for Space Research (INPE)
desc src	three hundred and twenty-two point nine one	four	two thousand and twenty-one	In April 2021, 322.91 sq km of the Legal Amazon were deforested, according to data from the National Institute for Space Research (INPE).
desc trg	322.91	4	2021	In April *two thousand and twenty-one*, *three hundred and twenty-two point nine one* sq km of the Legal Amazon were deforested, according to data from the National Institute for Space Research (INPE).
desc	three hundred and twenty-two point nine one	four	two thousand and twenty-one	In April *two thousand and twenty-one*, *three hundred and twenty-two point nine one* sq km of the Legal Amazon were deforested, according to data from the National Institute for Space Research (INPE).

Table 2: The strategies and representations of the numeric referring expressions. Strategies are *highlighted*.

the Python library[4], *num2words*, to transform numerical digits into their textual counterparts. This library is effective for both English and Brazilian Portuguese languages.

4 Experiments

To address our first research question **(RQ1)**, we evaluate the performance of three LLMs in generating numerical references: I) GPT-2, ii) BART, and iii) T5 for English domains. Additionally, for Portuguese, we fine-tuned GPorTuguese-2 (Guillou, 2020), a Brazilian Portuguese version of GPT-2, as well as mBART-50 (Tang et al., 2020) and mT5 (Xue et al., 2021), which are the multilingual versions of BART and T5, respectively. These models were selected due to a more sustainable perspective of LLMs (Rillig et al., 2023) and the environmental implications of the new LLMs, such as ChatGPT (OpenAI, 2023) and BARD[5]. The model training process involved 30 epochs, a learning rate of 1e-5, a batch size of 1, 5 early stops, and a maximum token length of 300.

4.1 Automatic Evaluation

We computed the BLEU score (Papineni et al., 2002) of the system to analyze the generated texts' fluency automatically and whether errors in numerical referring expressions are reflected in its result.

4.2 Human Evaluation

To answer our research questions **(RQ1)** and **(RQ2)**, we performed a human evaluation against the outcomes of our evaluated approaches.

Method We perform the human evaluation following the methodology of Thomson and Reiter (2020), which aims to quantify the quality of automatically generated texts according to the following taxonomy of errors: *Incorrect Number*, *Incorrect Named Entity*, *Incorrect Word*, *Context*, *Not Checkable* and *Other*. Besides these categories, a *Fluency* error category was incorporated into the evaluation, which allowed raters to assess the output for issues related to text flow acceptability. We are primarily interested in the dimensions concerning the number errors i.e., *Incorrect Number* and *Incorrect Word*. We also drew on best practices concerning error analysis and reporting as described in van Miltenburg et al. (2021).

Data preparation and Annotation process Overall, we selected 20% of a stratified sample, comprising 852 instances of Brazilian Portuguese output (per strategy and model). Three linguistically proficient annotators assessed these instances. To ensure reliability, a duplicate batch was evaluated by the same three raters. For English, all 240 outputs (per strategy per model) were independently annotated by two linguistically proficient raters. This process followed a pilot annotation of 50 instances for each language to clarify any ambiguities in the annotation guidelines before the full annotation task. Brazilian and English annotators and/or raters are members of the research team.

It is worth noting that for the Portuguese dataset annotators evaluated different entries in the first and second batches, allowing for inter-rater agreement assessment. To reduce bias during double annotation, access to corresponding entries in different batches was not allowed. For both datasets, in line with Thomson and Reiter (2020) methodology, we removed any disagreement as a result of raters not following annotations guidelines.

5 Results

The error rates and BLEU scores for each numerical strategy and model for both English and Portuguese are presented in Table 3. Numerical errors were found to be the most common type across both languages. However, the numerical error rates

[4] https://pypi.org/project/num2words/
[5] https://bard.google.com/

Error Rate Full Results for English (EN) and Brazilian Portuguese (PT)																		
S	LM		Number		Named Entity		Word		Context		Uncheckable		Other		Fluency		BLEU	
	EN	PT	EN	PT	EN	PT	EN	PT	EN	PT	EN	PT	EN	PT	EN	PT	EN	PT
No Desc	T5	mT5	0.48	0.45	0.05	0.02	0.08	0.08	0.03	0.03	0.03	0.03	0.08	0.07	0.05	0.03	0.69	0.58
	GPT2	GPT2-pt	0.65	0.24	0.28	0.04	0.03	0.03	0.53	0.03	0.08	0.01	0.60	0.14	0.95	0.09	0.14	0.60
	BART	mBART	0.50	0.34	0.00	0.04	0.00†	0.00†	0.08	0.09	0.00†	0.01	0.00	0.07	0.15	0.10	0.61	0.51
	Avg.	**Avg.**	**0.54***	**0.34**	**0.11**	**0.04**	**0.03**	**0.04***	**0.21**	**0.05***	**0.03**	**0.02**	**0.23**	**0.09**	**0.38**	**0.07**	**0.48**	**0.56**
D. Source	T5	mT5	0.45†	0.37	0.00†	0.01	0.00†	0.08	0.00†	0.05	0.05	0.01	0.00†	0.04	0.08	0.02	0.69	0.59
	GPT2	GPT2-pt	1.00	0.19†	0.05	0.01	0.00†	0.02	0.03	0.12	0.00†	0.01	0.00†	0.08	0.23	0.03	0.41	0.61
	BART	mBART	0.48	0.28	0.00†	0.03	0.00†	0.01	0.05	0.01	0.00†	0.01	0.03	0.13	0.23	0.05	0.62	0.59
	Avg.	**Avg.**	**0.64**	**0.28***	**0.02***	**0.01**	**0.00***	**0.04**	**0.03***	**0.06**	**0.02**	**0.01**	**0.01**	**0.09**	**0.18**	**0.03***	**0.57**	**0.60**
D. Target	T5	mT5	0.95	0.87	0.00†	0.00†	0.00†	0.04	0.03	0.00†	0.00†	0.00†	0.00†	0.01†	0.00†	0.02†	0.87†	0.65
	GPT2	GPT2-pt	0.95	0.90	0.03	0.01	0.00†	0.04	0.13	0.11	0.13	0.00†	0.00†	0.08	0.18	0.05	0.35	0.64
	BART	mBART	0.90	0.79	0.05	0.06	0.08	0.09	0.05	0.09	0.03	0.00†	0.00†	0.10	0.18	0.09	0.60	0.61
	Avg.	**Avg.**	**0.93**	**0.85**	**0.03**	**0.02**	**0.03**	**0.06**	**0.07**	**0.07**	**0.05**	**0.00***	**0.00***	**0.06**	**0.12**	**0.05**	**0.60***	**0.64**
Desc	T5	mT5	0.93	0.90	0.00†	0.00†	0.03	0.05	0.03	0.02	0.00†	0.00†	0.00†	0.01	0.00†	0.06	0.66	0.68†
	GPT2	GPT2-pt	0.90	0.80	0.13	0.01	0.03	0.12	0.23	0.14	0.03	0.00†	0.05	0.03	0.25	0.15	0.28	0.67
	BART	mBART	1.00	0.89	0.00†	0.00	0.03	0.07	0.03	0.03	0.00†	0.00†	0.00†	0.03	0.05	0.15	0.58	0.65
	Avg.	**Avg.**	**0.94**	**0.87**	**0.04**	**0.00***	**0.03**	**0.08**	**0.09**	**0.06**	**0.01***	**0.00***	**0.02**	**0.02**	**0.10***	**0.12**	**0.50**	**0.67***

Table 3: Error rates and BLEU score for the 4 numerical strategies and 3 language models – Higher error rates denote more errors. Higher BLEU scores denote greater Fluency. ∗(Lowest error rate among strategies averages); †(Lowest error rate among model and strategy combinations); S (Strategies); and D (Desc).

Incorrect Number Error Rate							
Strategies	LM	English (EN)			B. Portuguese (PT)		
		DM	DD	Overall	DM	DD	Overall
No Desc	T5/mT5	0.50	0.45†	0.48	0.55	0.18	0.36
	GPT2/GPT2-pt	0.65	0.65	0.65	0.00†	0.00†	0.00†
	BART/mBART	0.50	0.50	0.50	0.18	0.09	0.14
	Avg.	**0.55***	**0.53***	**0.54***	**0.24***	**0.09**	**0.17***
Desc Source	T5/mT5	0.45†	0.45†	0.45†	0.73	0.07	0.40
	GPT2/GPT2-pt	1.00	1.00	1.00	0.27	0.00†	0.14
	BART/mBART	0.50	0.45†	0.48	0.45	0.00†	0.23
	Avg.	**0.65**	**0.63**	**0.64**	**0.48**	**0.02***	**0.25**
Desc Target	T5/mT5	0.95	0.95	0.95	1.00	0.68	0.84
	GPT2/GPT2-pt	0.95	0.95	0.95	0.82	0.74	0.78
	BART/mBART	0.95	0.85	0.90	0.82	0.55	0.69
	Avg.	**0.95**	**0.92**	**0.93**	**0.88**	**0.66**	**0.77**
Desc	T5/mT5	1.00	0.85	0.93	1.00	0.68	0.84
	GPT2/GPT2-pt	0.90	0.90	0.90	0.82	0.52	0.67
	BART/mBART	1.00	1.00	1.00	1.00	0.69	0.84
	Avg.	**0.97**	**0.92**	**0.94**	**0.94**	**0.63**	**0.78**
Kappa Statistic		**0.94**	**0.92**	**0.93**	**1.00**	**0.99**	**0.97**

Table 4: Results displaying the "Incorrect Number" error rates in English and Portuguese, categorized by strategies, with higher values indicating more errors. To facilitate comparison, we present results solely for the Monthly (DM) and Daily Deforestation (DD) domains, which are common to both languages. ∗(Lowest error rate among strategies averages) and †(Lowest error rate among model and strategy combinations).

varied depending on the language, strategy, and models used.

In English, the average results per strategy indicated that using text to represent numerical references did not yield a positive impact. This is evidenced by the `No Desc` strategy, which resulted in the lowest error rate. However, when examining the results per model, T5(`Desc Source`) strategy presented the lowest error rate, followed by BART(`Desc Source`) and T5(`No Desc`) strategies. In terms of automatic evaluation, the `Desc Target` strategy yielded the highest BLEU score with T5 being the best model in this strategy. The **Kappa** coefficient for inter-rater agreement regarding *Incorrect Number* error for both languages reached up to **0.90** according to Table 4, indicating a reasonable consensus between human evaluations.

Contrary to English, describing Portuguese numerical referring expressions in the `Desc Source` strategy resulted in the lowest error rate. The model with the fewest errors was GPT2-pt(`Desc Source`) strategy. Regarding the automatic evaluation, the `Desc` strategy yielded the highest BLEU score (0.68) with mT5, being the best model in this strategy for Portuguese.

It is important to note that Brazilian Portuguese approaches were evaluated across more domains than their English counterparts due to differences in both datasets. To compare the numerical error rate of models across languages, Table 4 presents the numerical error rate of approaches in daily and monthly Amazon deforestation domains, which share identical meaning representations in English and Portuguese. Based on the `Incorrect Number Error Rate` results, the `No Desc` was the best strategy in both languages. While error rates between daily and monthly deforestation were similar in English, Portuguese utterances in daily report format introduced fewer numerical errors than monthly reports, likely due to the higher amount of daily deforestation training sentences for Portuguese models.

6 Conclusion and Limitations

Finally, we revisit the research questions outlined in Section 1: **(RQ1)** A human evaluation was performed to annotate different error categories, such as numerical, named entities, context, word, uncheckable, other, and fluency errors. Results depicted across languages, models, and numerical strategies show the numerical error rate as the highest among the errors. Hence, concerning this research question, there is clear evidence that pure state-of-the-art large language models struggle to generate adequate and faithful numerical referring expressions. **(RQ2)** Results demonstrated that the Brazilian Portuguese approach `Desc Source` performs better. However, for English, representing numerical references in spell-out form did not help regardless of whether it was present in the source meaning representation (`Desc Source`), in the target text (`Desc Target`) or both (`Desc`). As depicted in Table 1, we report lower results for English when compared with Portuguese. This may result from the smaller size of the English dataset compared to Brazilian Portuguese. Moreover, surprisingly, for English, fine-tuning LLMs with smaller amounts of training data did not appear to produce higher results than originally hoped. More experiments will be needed however to verify this.

As evidenced in the results, this study confirms that Large Language Models struggle to generate numerical referring expressions, although T5 has performed better. The proposed strategy to solve the problem did not affect English, although it decreased numerical errors when describing the numbers on the source of Portuguese trials. Hence this strategy for describing numbers may help in low-resource scenarios.

For future work, we plan to extend our experiments to GPT3 and GPT4[6]. However, since these models are neither free, nor reproducible due to limited or no information concerning model size, architecture, training parameters, and data set creation, we will investigate related open-source variations such as BLOOM[7] and GPT-J[8].

7 Ethics Statement

As highlighted in the Human Evaluation Subsection 4.2, all annotators are members of the research group and were responsible for evaluating with an equal amount of occurrences; hence ethical approval for conducting research with human subjects was not required. All data is publicly available (see Data Subsection 2 for more information). No consent from data subjects was required as this data is purely factual, containing no personal data, and hence compliant with the EU's General Data Protection Regulation (GDPR)[9].

8 Acknowledgements

This publication has emanated from research conducted with the financial support of the National Council for Scientific and Technological Development (CNPQ) under grants 313103/2021-6 and 305753/2022-3; the Foundation for the Coordination and Improvement of Higher Education Personnel (CAPES) under grants 88887.488096/2020-00 and 88887.508597/2020-00; the State Funding Agency of Minas Gerais (FAPEMIG) under Grant No APQ-01.461-14; the Science Foundation Ireland under CRT-AI Grant No 18/CRT/622; and ADAPT, the SFI Research Centre for AI-Driven Digital Content Technology at Dublin City University under Grant No 13/RC/2106_P2. Furthermore, we thank the Center for Artificial Intelligence (C4AI-USP) and the support of the São Paulo Research Foundation (FAPESP Grant No 2019/07665-4) and IBM Corporation.

References

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

João Campos, André Teixeira, Thiago Ferreira, Fábio Cozman, and Adriana Pagano. 2020. Towards fully automated news reporting in brazilian portuguese. In *Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional*, pages 543–554. SBC.

Thiago Castro Ferreira, Diego Moussallem, Ákos Kádár, Sander Wubben, and Emiel Krahmer. 2018. NeuralREG: An end-to-end approach to referring expression generation. In *Proceedings of the 56th Annual*

[6] `https://openai.com/blog/chatgpt`

[7] BLOOM: BigScience Large Open-science Open-access Multilingual Language Model – `https://huggingface.co/bigscience/bloom`

[8] `https://huggingface.co/docs/transformers/model_doc/gptj`

[9] `https://gdpr-info.eu/recitals/no-159/`

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1959–1969, Melbourne, Australia. Association for Computational Linguistics.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, et al. 2021. Finqa: A dataset of numerical reasoning over financial data. *arXiv preprint arXiv:2109.00122*.

Rossana Cunha, Thiago Castro Ferreira, Adriana Pagano, and Fabio Alves. 2020. Referring to what you know and do not know: Making referring expression generation models generalize to unseen entities. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2261–2272, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ondřej Dušek and Filip Jurčíček. 2016. Sequence-to-sequence generation for spoken dialogue via deep syntax trees and strings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 45–51, Berlin, Germany. Association for Computational Linguistics.

Sebastian Gehrmann, Falcon Dai, Henry Elder, and Alexander Rush. 2018. End-to-end content and plan selection for data-to-text generation. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 46–56, Tilburg University, The Netherlands. Association for Computational Linguistics.

Pierre Guillou. 2020. Gportuguese-2 (portuguese gpt-2 small): a language model for portuguese text generation (and more nlp tasks...).

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Jeonghwan Kim, Junmo Kang, Kyung-min Kim, Giwon Hong, and Sung-Hyon Myaeng. 2022. Exploiting numerical-contextual knowledge to improve numerical reasoning in question answering. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1811–1821.

Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1203–1213, Austin, Texas. 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.

Anna Björk Nikulásdóttir and Jón Guðnason. 2019. Bootstrapping a Text Normalization System for an Inflected Language. Numbers as a Test Case. In *Proc. Interspeech 2019*, pages 4455–4459.

OpenAI. 2023. Gpt-4 technical report.

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 on association for computational linguistics*, pages 311–318. 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.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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

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

Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems*. Studies in Natural Language Processing. Cambridge University Press, Casmbridge, U.K.

Matthias C Rillig, Marlene Ågerstrand, Mohan Bi, Kenneth A Gould, and Uli Sauerland. 2023. Risks and benefits of large language models for the environment. *Environmental Science & Technology*, 57(9):3464–3466.

André Luiz Rosa Teixeira, João Campos, Rossana Cunha, Thiago Castro Ferreira, Adriana Pagano, and Fabio Cozman. 2020. DaMata: A robot-journalist covering the Brazilian Amazon deforestation. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 103–106, Dublin, Ireland. Association for Computational Linguistics.

Richard Sproat. 2022. Boring problems are sometimes the most interesting. *Computational Linguistics*, 48(2):483–490.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Craig Thomson and Ehud Reiter. 2020. A gold standard methodology for evaluating accuracy in data-to-text systems. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 158–168, Dublin, Ireland. Association for Computational Linguistics.

Emiel van Miltenburg, Miruna Clinciu, Ondřej Dušek, Dimitra Gkatzia, Stephanie Inglis, Leo Leppänen, Saad Mahamood, Emma Manning, Stephanie Schoch, Craig Thomson, and Luou Wen. 2021. Underreporting of errors in NLG output, and what to do about it. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 140–153, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. 2019. Do NLP models know numbers? Probing numeracy in embeddings. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 5307–5315.

Tsung-Hsien Wen, Milica Gašić, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1711–1721, Lisbon, Portugal. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Hao Zhang, Richard Sproat, Axel H Ng, Felix Stahlberg, Xiaochang Peng, Kyle Gorman, and Brian Roark. 2019. Neural models of text normalization for speech applications. *Computational Linguistics*, 45(2):293–337.

A Appendix

A.1 Annotation Guidelines

After the most common error cases were identified and the treatment for the most difficult cases was agreed upon, annotators followed common guidelines for the rest of the evaluation process, as described in the following list:

- **Incorrect Number:** Has incorrect numerical values (e.g., model verbalizes an area value of “354” as “345”); Numerical values not verbalized in numerical form in the final texts were considered incorrect (e.g., “three hundred fifty-four” instead of “354”);
- **Incorrect Named Entity:** verbalizes entities incorrectly or verbalizes entities that do not exist;
- **Incorrect Word:** occurrence of spelling errors;
- **Context Error:** verbalizes some communicative intent incorrectly (e.g., verbalizes last month’s deforestation variation instead of total area deforestation);
- **Not checkable:** adds information that is not present in the input semantic representation in the verbalized text;
- **Other:** other types of verbalization errors;
- **Fluency:** the hypothesis verbalizes a not fluent text.

The annotation guidelines are summarised below:

- Entries were distributed in a collaborative spreadsheet.
- Each row consisted of the original Meaning Representation (MR), the generated hypothesis, and the rating categories.
- LLMs used to generate the entries were omitted in the spreadsheet.
- The spreadsheet was formatted to highlight the options (y - red; n - green) aiming to aid/ease the process with visual cues.
- Difficult cases were commented on to be further discussed within the group of raters, fostering improvements in the guidelines.
- Once, the annotation was finished, the spreadsheets were exported in .csv files for result computation.

Expected output The most affected state and municipality were respectively Pará (177.84 sq km) and Altamira, in the state of Pará (51.07 sq km). **Deforestation monthly intents** TOTAL_DEFORESTATION(area="177.84", location="deter-amz", month="4", state="PA", year="2021") [SEP] TOTAL_DEFORESTATION(area="51.07", city="Altamira", location="deter-amz", month="4", state="PA", year="2021")
T5 *nodesc* The most affected state and municipality were respectively Pará (**177.84 sq km**) and Altamira, in the state of Pará (**51.07 sq km**).
T5 *desctrg* The state with the most deforestation in the month was Pará (**one hundred and seventy-seven point eight four sq km**), and the most devastated municipality was Altamira / Pará, where deforestation amounted to **fifty-one point zero seven sq km**.
T5 *descsrc* The state with the most deforestation in the month was Pará (**177.84 sq km**), and the most devastated municipality was Altamira / Pará, where deforestation amounted to **51.07 sq km**.
T5 *desc* The state with the most deforestation in the month was Pará (**one hundred and seventy-seven point eight four sq km**), and the most devastated municipality was Altamira / Pará, where deforestation amounted to **fifty-one point zero seven sq km**.

Table 5: Sample from T5 outputs for English considering all 4 strategies. T5 performed as the best model for English. The numeric referring expressions are **bolded**.

A.2 Expected Output

A sample from the expected output is presented in Table 5 considering the meaning representation and each strategy in English. Furthermore, Tables 6 and 7 show Human Evaluation results for Portuguese and English languages and highlight problems regarding generating numerical referring expressions.

Language	Incorrect Number	Incorrect Named Entity	Incorrect Word	Context Error
Input	area="322.91"	city="Novo Progresso, Itaituba"	-	-
English	The National Institute for Space Research (INPE) estimated that deforestation of the Legal Amazon amounted to **2,322.91 sq km** in April two thousand and twenty, which is a one hundred and twenty-six percent increase from the previous month.	The National Institute for Space Research (INPE) reported that deforestation amounted to twenty-one point seven five sq km in the state of **Pará**, in the state of **Pará**, in February two thousand and twenty.	The main **class** of deforestation was clear-cut deforestation, which removes all **vegetetation** of the soil, responsible for 317.93 sq km of deforested area."	The most affected state and municipality were respectively **Pará** (177.84 sq km) and Altamira / Pará, in the state of Pará.
Input	cases="4091801" deaths="125584"	uc="PARQUE NACIONAL DO JAMANXIM"	-	-
Portuguese	São registrados, no total, **135.584** mortes e **4.093.801** casos de #COVID19 no Brasil.	O INPE gerou alerta para devastação (0,19 km²) causada pelo desmatamento com solo exposto, que remove totalmente a vegetação da floresta, no dia 10 de agosto de 2020 na **PARQUE NACIO**	A cidade mais atingida foi SANTAQUITÉRIA, **em CEARÁ**, que registrou 22 focos de incêndio.	O Instituto Nacional de Pesquisas Espaciais(INPE) registrou um total de quinhentos e sessenta e nove focos de queimadas no território brasileiro, no dia onze de outubro de dois mil e vinte, **o território brasileiro foi atingido**.

Table 6: Examples of categories of error in human evaluation for English and Brazilian Portuguese.

Language	Not Checkable	Other	Fluency Problem
Input	-	-	
English	The main cause of deforestation was the destruction of the soil, which leaves the soil clear of vegetation.	The National Institute for Space Research (INPE) in Pará, where the most affected municipality was **Novo Pro**	The National Institute for Space Research (INPE) reported that deforestation amounted to 21.75 sq km in the state of Pará, **in the state**
Input	area="0.32" day="22" month="8"	-	-
Portuguese	O INPE gerou alerta para devastação **(0,22 km2)** causada pelo desmatamento com solo exposto, que remove totalmente a vegetação da floresta, no dia 22 de agosto de 2020 na RESERVA EXTRATIVISTA VERDE PARA SEMPRE / Pará - no mês já são 2 dias com alertas e 0,32 km2 desmatar.	**A A A A A A A A BIOLÓGICA NASCENTES DA SERRA DO CACHIMBO** somou dois vírgula sete três km2 de área desmatada no mês de novembro de dois mil e vinte.	Com um total de mil quinhentos e setenta e oito vírgula oito sete km2, **o desmatamento com solo exposto**, deixando a terra sem vegetação, a principal causa de destruição da Amazônia Legal no mês foi o **desmatamento com solo exposto, deixando a terra sem vegetação**.

Table 7: Examples of categories of error in human evaluation for English and Brazilian Portuguese.