Enhancing Writing Skills of Chilean Adolescents: Assisted Story Creation with LLMs

Hernan Lira, Luis Martí, Nayat Sanchez-Pi

Enhancing Writing Skills of Chilean Adolescents: Assisted Story Creation with LLMs

Hernan Lira
Inria Chile Research Center
hernan.lira@inria.cl

Luis Martí
Inria Chile Research Center
luis.marti@inria.cl

Nayat Sanchez-Pi
Inria Chile Research Center
nayat.sanchez-pi@inria.cl

Abstract

This study presents an automatic story generation model in Chilean Spanish, designed to assist students in the writing process and help improve their writing skills. The methodology employed includes the creation of a corpus of stories in Spanish and Chilean Spanish, as well as data processing and extraction of relevant information from the stories. The model is trained using fine-tuning and prompt-engineering techniques to adapt it to story generation. The results obtained indicate that the stories generated by the model outperform other text generation models in terms of relevant natural language processing metrics.

1 Introduction

The development of writing skills plays a crucial role in the educational formation of individuals. However, numerous challenges arise when implementing pedagogical activities related to writing. Primarily, there are limited resources, including educators and educational tools, to adequately accommodate the diverse population of students. Consequently, providing a tailored educational experience for each student becomes an overwhelming task, particularly concerning creative endeavors. Secondly, students encounter various obstacles when engaging in writing practices. Among these challenges, "writer's block" or "blank page syndrome" stands out as a pivotal hindrance, impeding the effective translation of conceived ideas into coherent sentences and narrative elements.

The solution to this problem involves the development of a machine-learning tool aimed at assisting students in creating stories. The goal is then to engage students by simplifying the process of generating stories rooted in their personal interests and ideas. This entails using a set of questions designed by educators, along with the corresponding responses provided by students, to generate a natural language story. Subsequently, students can revise and continue writing based on this initial narrative foundation.

While a mechanistic transposition could be devised to convert the response elements into a predetermined narrative template, such an approach conflicts with the fundamental principles of encouraging creativity and spontaneity. These principles are central to the objectives of our study. Leveraging recent advancements in this field, we propose a solution based on text generation using a pre-trained large language model. This approach effectively addresses the multifaceted aspects of the problem while remaining computationally viable.

Section 2 of the paper outlines the context, background, and technical challenges of the project. In Section 3 we discuss preliminary research. Section 4 details our solution, including the model and

NeurIPS'23 Workshop on Generative AI for Education (GAIED).

the prototype. Section 5 presents and discusses results and, finally, Section 6 outlines future work and conclusions.

2 Background

Pre-trained Language Models for text generation Autoregressive language models are trained to predict the next word, denoted as x_i, conditioned on all previous words in the sequence $\langle x_1, x_2, \ldots, x_{i-1} \rangle$. Without loss of generality, the goal is to maximize the log-likelihood $\ell = \sum \log(P(x_i|x_1, x_2, ..., x_{i-1}; \theta_T))$, where θ_T represents the model parameters (Min et al., 2021).

Transformer-based models (Vaswani et al., 2017), like generalized pretrained transformer (GPT) models (Radford et al., 2018, 2019; Brown et al., 2020; OpenAI, 2023), undergo a pre-training process on an extensive unlabeled corpus. They exclusively utilize the autoregressive decoder component of the transformer architecture, characterized by the stacking of multiple decoder layers. This architectural design enables comprehensive attention to all preceding words within the sequence when predicting the next word (Radford et al., 2018).

These models are currently referred to as *pre-trained large language models* due to their extensive parameter count, demanding substantial computational resources for training, including high-end hardware, time, and large datasets. The cost associated with training these models reaches hundreds of thousands or even millions of dollars, limiting their development to major tech companies. Once a pre-trained model is available, it can be fine-tuned for specific tasks (Li et al., 2021; Qiu et al., 2020; Serrano et al., 2022).

Technical Challenges: Data Availability and the Spanish Language The effectiveness of these models is impeded by the scarcity of extensively annotated datasets. Many text generation methodologies heavily depend on large volumes of manually annotated data, restricting their usability in domains lacking a sufficient number of annotated examples (Li et al., 2021). Deep neural networks, often struggle to adapt to these datasets, resulting in suboptimal performance in real-world applications. Consequently, these models face challenges in capturing intricate relationships between contextual cues and word meanings, as well as in creating contextual word representations that enhance the quality of the generated text. This challenge is further compounded when considering the need for text generation in Chilean Spanish. Although Spanish is one of the world's most widely spoken languages, the availability of language models specifically tailored to the Spanish language remains limited.

User-Centric Application Overview Before delving into the technical details of the story generation model, it is essential to describe the process that a student (user) undergoes while utilizing the application. The application offers a user-friendly interface to guide students through the story creation process. Users begin by responding to a set of designed questions aimed at gathering information and ideas related to the story's theme. These questions are curated by educators to ensure meaningful content and context. Once the user submits their responses, the application generates an initial text based on the provided information. At this point, the user has three options: they can edit the generated text and continue working with the model, proceed without changes, or modify their initial responses to the questions. This iterative process empowers users to refine and shape the story according to their preferences and creative choices. They have the flexibility to make adjustments to the generated text, ensuring the final result aligns with their vision. The application, in conjunction with pedagogical activities, aims to inspire and support students in crafting unique narratives that delve into their thoughts, experiences, and interests.

3 Related Work

Among the technical challenges mentioned, the adaptation of these language models to the Spanish language is one of the primary hurdles. Currently, in comparison to the English language, there are few robust language models dedicated to Spanish. Most of these models are dedicated to the encoder part of the architecture, i.e., pattern recognition in the language to develop multipurpose models. The following studies stand out in this regard.

The study RigoBERTa (Serrano et al., 2022) presents a language model based on the BERT architecture(Devlin et al., 2018). This model is trained using a large Spanish corpus and has demonstrated its effectiveness in various applications. Another model named BETO (José et al., 2020) focuses on creating a pre-trained BERT model specifically for the Spanish language, along with an evaluation dataset. This model has been trained on a Spanish corpus and evaluated on various natural language processing tasks, yielding promising results. Finally, BERTIN (De la Rosa et al., 2022) proposes an efficient methodology for pre-training a language model in Spanish using perplexity sampling. This approach enhances the quality of the Spanish language model and its ability to generate coherent and relevant text.

While the aforementioned studies represent a significant advancement in adapting language models to the Spanish language, none of these studies specifically focus on text generation. Although there are some promising results, it is important to note that there are still pending challenges because the quality and fluency of generated text can vary. The ability to capture the nuances of the language and the diversity of language styles is not yet robust. For these reasons, in this study, generic language models like the ones mentioned earlier cannot be simply employed. Instead, it is necessary to create a model specifically tailored to the task of text generation in the domain of storytelling in Chilean Spanish.

4 Methods

Our text generation model is built upon the foundation of a large pre-trained language model, GPT-3. These models, trained on extensive corpora, have demonstrated their capabilities in various NLP tasks, including translation, summarization, and question-answering (Li et al., 2021). However, adapting them to specific domains and tasks, such as text generation, requires specialized techniques (Raffel et al., 2020). In our case, the task is text generation, and the domain is centered around storytelling in Spanish. To adapt a model trained in one domain (domain A) for a different domain (domain B), specific strategies can be employed. This approach involves leveraging the deep layers of the architecture from domain A and fine-tuning only the outer layer(s) specific to domain B, which is specialized for the particular task. This method conserves computational resources while optimizing performance. To achieve domain adaptation, we first acquire or create a corpus containing Chilean Spanish stories. Subsequently, we have two options for adaptation: Fine-tuning (Wolf et al., 2020) and Prompt Engineering (Yuan et al., 2021). Once the model is successfully adapted to our domain, we establish the complete processing and training pipeline. Finally, we develop the application that hosts the adapted model.

4.1 Annotation, Data Collection, and Corpus Creation

The corpus comprises the entirety of information and knowledge that our model will utilize to adapt to the specific domain from the pre-trained language model. As discussed in Section 2, several Spanish corpora are available across various domains, primarily unlabeled. These corpora serve as suitable candidates for developing a general-purpose Spanish language model (Serrano et al., 2022). However, they fall short when it comes to constructing a domain-specific model. To address this limitation, acquiring a relevant dataset for storytelling becomes imperative. The initial step involves the gathering of stories in Spanish. Additionally, we need to establish a connection between question-answer pairs and the generated text, considering that the model generates stories based on students' responses to specific questions. Ultimately, by appending data from both sources, we can construct the corpus that will be employed for our final model.

To enhance the educational value of our application, educators have formulated a series of questionnaires to steer students in crafting stories and to stimulate introspection and exploration of their initial concepts. In our application, we select a subset of these questionnaires to acquire the story's framework and to have students clarify the underlying motivation for their narrative. Table 1 presents an exhaustive compilation of the questionnaires, along with their respective relevance to specific facets of the story's framework. These questionnaires aim to collect information about the narrative's context, progression, and resolution, with a particular focus on elucidating the emotional states arising from the unfolding events.

In conjunction with the set of questions, our internal data sources encompass 769 narratives crafted by students in response to prior assignments. To ensure the accessibility of pertinent data for our

Table 1: Subset of questions considered in the final application.

	Question	Structure
1	Who is the character in the story?	Context - characters
2	What characteristics of the character would you highlight?	Context - characters
3	How does the story begin?	Context
4	Where does the story take place?	Context - setting
5	What is the situation or conflict that drives the narrative?	Conflict
6	How do the characters react to the conflict?	Conflict
7	What emotions arise in response to the conflict?	Conflict
8	Where does the story take place?	Optional

model, these narratives have been meticulously annotated with the corresponding questions, employing the Label Studio[1] tool. The objective is to leverage these annotations to facilitate the model's comprehension of the connections between the questions and the textual content generated by the students.

To expand our collection of Spanish narratives, we employ web scraping techniques and perform PDF-to-text conversion. Initially, we gather narratives from the Santiago en 100 palabras portal[2], as they closely align with the context of our research problem. These narratives are authored by a diverse group of Spanish speakers, primarily of Chilean origin, and encompass a wide range of plots and writing styles, with only the highest-quality narratives being selected and published. Additionally, a subset of narratives by Spanish-language authors is incorporated to enhance the initial dataset.

We establish associations between the collected narratives and a set of questions, utilizing a subset of questions to reduce computational complexity. To achieve this, we initially employ language models for information extraction. Our approach is inspired by the work of (Veyseh et al., 2021a), who demonstrated that fine-tuned GPT-2 language models could generate labels from original texts. However, this technique has a drawback as it may introduce noise into the dataset, including errors in grammar, nonsensical sentences, or inaccurate labels. Expanding on this, (Veyseh et al., 2021b) introduced a neural network architecture known as "student-teacher." In this framework, the "teacher" network is trained with previously labeled narratives to establish foundational knowledge, while the "student" network is trained with a combination of labeled and unlabeled narratives, guided by specific constraints to ensure consistency in the generated labels. This methodology notably enhances the quality of the final labeling. Another approach under consideration involves generating these labels through a Question Answering methodology. A language model of the BERT type (Devlin et al., 2018), focusing on the encoder component of the transformer architecture, is trained to formulate questions based on text (Alberti et al., 2019). Labels are generated when there is a high level of statistical confidence.

The corpus utilized by the Spanish story generation model is curated. It comprises a total of 30,151 stories, each subjected to extensive processing to extract the required structural information. Conceptually, if represented as a table, each row represents an individual story. The columns encompass the story's textual content, labels denoting their corresponding start and end positions, and pertinent metadata that necessitates inclusion. The undertaken efforts have effectively addressed several challenges inherent in data usage within language models (Min et al., 2021), including encoding structural details within the input and preserving fidelity between the generated and original text.

4.2 Domain Adaptation and Training

Fine-tuning is used to inject task-specific information into a language model by making precise adjustments to its parameters using a task-specific dataset (Radford et al., 2019; Kalyan et al., 2021). The majority of computational effort during fine-tuning is directed toward ensuring that the model's output layers generate the desired representation of the input. Additionally, these output layers must

[1] `https://labelstud.io/`

[2] `https://www.santiagoen100palabras.cl/web/%23libros`

Algorithm 1: Selection of the Best Prompt and Fine-Tuned Model

```
Input  : Base models, prompt designs, and fine-tuned language models
Output : Best prompt and best fine-tuned model
foreach base model do
    foreach prompt design do
        Test model with the prompt design;
        Measure model's result;
    Select the best prompt based on average results;
foreach base language model do
    Fine-tune the model on the corpus;
foreach hyperparameter combination do
    Create a model with hyperparameter combination;
    Optimize parameters using Bayesian optimization;
    if result is better than previous best result then
        Update best result and best model;
Validate the best model;
foreach candidate model do
    foreach portion of the corpus do
        Perform an experiment with the candidate model and measure performance;
    Select the best candidate model based on average performance;
Use the best prompt and best candidate model for the application.
```

effectively distill the essential information from the embeddings of each token into the requisite number of classes, which corresponds to the number of future words to be generated.

In our model, we specifically employ the *Intermediate Fine-Tuning* technique, which is based on the concept of incorporating a corpus containing a substantial number of labeled examples (Phang et al., 2018). Furthermore, intermediate fine-tuning can be categorized into domain adaptation and task adaptation techniques, with our model utilizing the domain adaptation approach. The advantage of this strategy lies in its flexibility, as the corpus need not be constrained solely to specific labels or the exact domain of the given task. It may include examples without labels or with slightly different domains, thereby increasing the corpus's size. Furthermore, a larger corpus helps mitigate the risk of overfitting the model to a very limited dataset, particularly given the inherent constraints of the problem. Consequently, this technique is employed to accommodate the fact that not all collected stories in our dataset are labeled.

A commonly employed approach for prompt construction is the template-based approach (Min et al., 2021). This method transforms NLP tasks into formats that closely align with the objectives of pre-trained models, effectively harnessing the knowledge stored within these models while requiring fewer examples in the corpus and considering the task to solve. Various strategies for template design have been proposed, including pattern-exploiting (Schick and Schütze, 2020), prefix-prompts (Li and Liang, 2021), and demonstration learning (Gao et al., 2020), among others. In constructing and designing our templates, we draw inspiration from the work of (Zhao et al., 2021), which illustrates the impact of prompt variation on the final model's performance. This approach allows us to fine-tune the template to optimize the model's story generation output. We create multiple prompt designs based on templates to explore their effectiveness.

After domain adaptation of the base language model through fine-tuning and prompt-tuning, a variety of base model versions are available. To optimize their performance, these models are trained with diverse hyperparameter combinations using Bayesian optimization (Garnett, 2023). The base models under consideration include GPT-2 (Radford et al., 2019), GPT-3 (Brown et al., 2020), and BLOOM (Scao et al., 2022), each of which offers different versions based on their parameter counts. The model training and hyperparameter tuning process utilizes domain-adapted base language models as input. These models are trained on the corpus, and their parameters are optimized through Bayesian optimization. Model validation entails comparing results across various hyperparameter combinations to select the combination yielding the best performance. This iterative process en-

Table 2: CBL, MSJ, BES, and BRT metrics as an indicator of fluency and coherence. BBL as an indicator of divergence. ⇑: higher is better; ⇓: lower is better. BRT values are negated here.

	Model	Fluency/Coherence				Divergence
		CBL ⇑	MSJ ⇑	BES ⇑	BRT ⇓	BBL ⇑
Corpus	GPT3	**39.6**	**15.0**	**87.0**	**8.6**	**25.6**
	BLOOM	33.5	14.6	86.8	9.5	24.8
	BLOOM-7b1	26.6	14.6	86.7	9.7	25.0
	GPT-2 Spanish	27.2	11.6	86.6	8.6	24.6
	BERTIN GPT-J	27.5	14.7	86.8	9.5	25.1
Etiqueta	GPT3	**39.3**	**14.9**	**87.1**	**8.8**	**25.3**
	BLOOM	32.3	14.5	86.1	9.8	24.2
	BLOOM-7b1	26.2	14.1	86.0	9.9	25.5
	GPT-2 Spanish	26.8	11.3	85.7	9.0	23.9
	BERTIN GPT-J	26.9	14.2	86.1	9.7	25.0

sures the final model is finely tailored to the specific storytelling task. The routine for training the model, establishing its hyperparameters, and selecting the best model is outlined in Algorithm 1. The outcome is a collection of candidate models, each with its fine-tuned hyperparameters. For every candidate model, a series of experiments are carried out across different segments of the corpus. The performance achieved in these experiments is then averaged, and the candidate model that attains the best overall result is singled out. This chosen model will serve as the foundation for the application.

5 Results

For each of these base models, there exist various versions, which differ in terms of their parameter count. In the case of the BLOOM base model, the distinct versions are associated with the quantity of pre-trained parameters. This holds true for both GPT-2 and GPT-J as well. Notably, these latter two models have been specifically optimized for the Spanish language, with the caveat that they are smaller in scale compared to GPT-3. Table 2 presents the outcomes of the base models following their complete traversal through the pipeline detailed in the preceding section. The results encompass the CBL, MSJ, BES, and BRT metrics across two scenarios: firstly, utilizing the entire corpus, and secondly, employing solely the portion of the corpus that contains labeled stories.

It is evident that the model trained based on GPT-3 outperforms other models across all metrics, both when assessing the entire corpus and when focusing solely on the labeled stories. As anticipated, the results for all models exhibit a decline when considering only the labeled stories from the corpus, underscoring the necessity of expanding the story collection and information extraction methodology. Moreover, the text generated by GPT-3 exhibits remarkable qualitative robustness. Responses to the initial questions, as specified in the prompt, seamlessly integrate into the text. Coherence prevails among the characters, their attributes, and the unfolding events. Furthermore, the text length remains appropriate. Notably, in the conducted tests following hyperparameter optimization, the base GPT-3 model exhibits the least uncertainty in terms of text quality.

5.1 Human Evaluation

A qualitative preliminary assessment was conducted by educators who meticulously examined the narratives generated by the Spanish story generation model while it was utilized by a group of 5 high-school students and 5 adults from Chile. The evaluation encompassed a range of metrics, including coherence, fluency, grammar, relevance, and human-likeness (Jin et al., 2023) to provide a comprehensive appraisal of the model's performance. Coherence analysis focused on the logical flow of the stories, assessing the model's ability to maintain a consistent and meaningful narrative structure. Fluency evaluation considered the linguistic fluency and naturalness of the generated text, with a keen eye on vocabulary, syntax, and discourse markers. Grammar assessment entailed scrutinizing the syntactic correctness of the text, detecting and categorizing grammatical errors or deviations from standard language conventions. Relevance was a paramount metric, ensuring that

the generated stories aligned with the contextual prompts and the intended storytelling objectives captured by the initial questions asked to the users. Moreover, human-likeness evaluation sought to gauge the extent to which the narratives resembled those crafted by human authors in terms of style, tone, and narrative voice. Notably, using a Likert scale, we found that the average scores for coherence, fluency, grammar, and relevance emerged as the highest, signifying that the generated Spanish narratives effectively fulfilled the fundamental linguistic and storytelling requisites. These metrics underscored the model's proficiency in sustaining logical narrative structures, linguistic fluency, syntactic correctness, and contextual relevance. Conversely, human-likeness, the most subjective metric, revealed potential areas for enhancement, which was not entirely unexpected, given the inherently intricate nature of assessing the model's ability to emulate the stylistic nuances, tone, and narrative voice characteristic of human-authored stories. This preliminary evaluation by educators across diverse participant groups furnishes valuable insights into the model's performance, highlighting its strengths and elucidating avenues for refinement within the domain of narrative generation.

6 Conclusion

In this research, we have developed a language model tailored for the automatic generation of stories in Spanish, aimed at aiding students throughout the writing process and enhancing their writing skills. Our contributions have encompassed various pivotal domains. Firstly, we have curated a corpus of Spanish stories, encompassing both annotated and unannotated content. For the latter, we have implemented data processing and information extraction techniques, amplifying our model's capacity to capture the fundamental traits of Spanish stories and yielding more precise and coherent outcomes in text generation. Concerning the model's construction, we harnessed the potential of the pre-trained transformer-based language model, GPT-3. Through fine-tuning and prompt engineering methodologies, we have customized this model to our specific domain of Spanish story generation, achieving elevated levels of precision and coherence in text generation. In terms of performance, our results substantiate that the stories generated by our model surpass other state-of-the-art text generation models, as evidenced by significant improvements in relevant natural language processing metrics for Spanish stories.

Acknowledgments and Disclosure of Funding

This work was made in the framework -and funded in part- under a collaboration of the Inria Chile Research Center with forEach S.L. and Fundación Ciudad Literaria for Project Ficzine (`https://www.ficzine.com/`). We would like to thank their contribution, and, in particular, Franco Margozzini (forEach) and Conrado Soto (FCL).

This work is also funded by ANID Strengthening R&D capabilities Program CTI220002 Inria Chile, Inria Challenge OcéanIA, and Inria associated team SusAIn.

References

Alberti, C., Andor, D., Pitler, E., Devlin, J., and Collins, M. (2019). Synthetic qa corpora generation with roundtrip consistency. *arXiv preprint arXiv:1906.05416*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

De la Rosa, J., Ponferrada, E. G., Villegas, P., Salas, P. G. d. P., Romero, M., and Grandury, M. (2022). Bertin: Efficient pre-training of a spanish language model using perplexity sampling. *arXiv preprint arXiv:2207.06814*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Gao, T., Fisch, A., and Chen, D. (2020). Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*.

Garnett, R. (2023). *Bayesian optimization*. Cambridge University Press.

Jin, Y., Bhatia, A., Wanvarie, D., and Le, P. T. (2023). Towards improving coherence and diversity of slogan generation. *Natural Language Engineering*, 29(2):254–286.

José, C., Gabriel, C., Rodrigo, F., and Jorge, P. (2020). Spanish pre-trained bert model and evaluation data. *PML4DC at ICLR*, 2020.

Kalyan, K. S., Rajasekharan, A., and Sangeetha, S. (2021). Ammus: A survey of transformer-based pretrained models in natural language processing. *arXiv preprint arXiv:2108.05542*.

Li, J., Tang, T., Zhao, W. X., and Wen, J.-R. (2021). Pretrained language models for text generation: A survey. *arXiv preprint arXiv:2105.10311*.

Li, X. L. and Liang, P. (2021). Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Min, B., Ross, H., Sulem, E., Veyseh, A. P. B., Nguyen, T. H., Sainz, O., Agirre, E., Heinz, I., and Roth, D. (2021). Recent advances in natural language processing via large pre-trained language models: A survey. *arXiv preprint arXiv:2111.01243*.

OpenAI (2023). Gpt-4 technical report.

Phang, J., Févry, T., and Bowman, S. R. (2018). Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*.

Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., and Huang, X. (2020). Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10):1872–1897.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving language understanding by generative pre-training.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., et al. (2022). Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Schick, T. and Schütze, H. (2020). Exploiting cloze questions for few shot text classification and natural language inference. *arXiv preprint arXiv:2001.07676*.

Serrano, A. V., Subies, G. G., Zamorano, H. M., Garcia, N. A., Samy, D., Sanchez, D. B., Sandoval, A. M., Nieto, M. G., and Jimenez, A. B. (2022). Rigoberta: A state-of-the-art language model for spanish. *arXiv preprint arXiv:2205.10233*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Veyseh, A. P. B., Lai, V., Dernoncourt, F., and Nguyen, T. H. (2021a). Unleash gpt-2 power for event detection. 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)*, pages 6271–6282.

Veyseh, A. P. B., Van Nguyen, M., Min, B., and Nguyen, T. H. (2021b). Augmenting open-domain event detection with synthetic data from gpt-2. In *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part III 21*, pages 644–660. Springer.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Yuan, W., Neubig, G., and Liu, P. (2021). Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

Zhao, Z., Wallace, E., Feng, S., Klein, D., and Singh, S. (2021). Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR.