LARI Dataset: A Native Portuguese Question Answering Dataset from Brasileiras em PLN

Júlia da Rocha Junqueira[1], Larissa A. de Freitas[2] and Ulisses Brisolara Corrêa[2]
[1]Institute of Informatics, Federal University of Rio Grande do Sul (UFRGS), Porto Alegre, Brazil
[2]Center for Technological Advancement, Federal University of Pelotas (UFPel), Pelotas, Brazil
julia.junqueira@inf.ufrgs.br, {larissa, ulisses}@inf.ufpel.edu.br

Abstract

Recent advances in the field have revolutionized Question and Answering (QA). However, for languages like Portuguese, progress is often hindered by the lack of native training resources. To address this gap, this paper introduces LARI, a new dataset designed to benchmark and enhance QA in Portuguese. Our methodology combines the capabilities of the Sabiá-7B model, fine-tuned via QLoRA on a domain-specific corpus, with human validation. We utilized the book Natural Language Processing – Concepts, Techniques, and Applications in Portuguese (2nd Edition), as a case study for content extraction. The generated instances underwent expert human evaluation, achieving an average quality score of 4.47 out of 5. The final dataset, comprising 464 context-question-answer triples, is made publicly available to the community, offering a valuable resource for future research in low-resource settings.

1 Introduction

Questions and answers play an essential role both in everyday language and in the educational process. In the context of Natural Language Processing (NLP), reproducing these human interactions in an efficient and automatic manner is a challenge that has received growing attention, such as automatic Question Answering (QA). While the task aims to reproduce human interaction by creating meaningful questions and contextually relevant answers (Rus et al., 2010; Ushio et al., 2023; Allam and Haggag, 2012), its success is intrinsically linked to the quality and availability of training data.

To formulate questions that reflect deep grammatical understanding and cultural nuances (Pan et al., 2019), language models require exposure to vast amounts of high-quality linguistic examples. However, current research is predominantly skewed towards high-resource languages like English. For languages such as Portuguese, most available Portuguese QA data derives from translated English sources, which may not reflect linguistic and cultural patterns. This lack of resources limits the ability of models to handle coherence, contextual relevance, and linguistic variety in real-world scenarios.

Addressing the lack of native, educationally-grounded benchmarks is therefore essential for advancing generative QA in Portuguese. To this end, this work proposes a methodology that leverages a large language model (Sabiá-7B) as a tool to accelerate the generation of QA pairs based on paragraphs of textual information, and introduces **LARI (Lightweight Adaptation for Response and Inquiry)**, a publicly available Portuguese QA dataset comprising 464 expert-validated triples derived from the book Natural Language Processing – Concepts, Techniques, and Applications in Portuguese (2nd Edition) (Caseli and Nunes, 2023), designed to benchmark and enhance automatic question and answer generation. Although answering questions based on a single paragraph might seem like a fundamental task, it represents the cornerstone of modern NLP applications such as Machine Reading Comprehension (MRC) and Retrieval-Augmented Generation (RAG). Therefore, LARI provides a highly representative evaluation scenario to test whether models can extract accurate facts from contexts in Portuguese without suffering from hallucinations.

The LARI dataset is available at: https://huggingface.co/datasets/jjuliar/plndataset.

2 Related Works

Large-scale QA datasets originally developed for English, such as SQuAD (Rajpurkar et al., 2016), have frequently been adapted to other languages through automatic translation in order to bootstrap

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 1055–1061

resources for low- and medium-resource languages. Its Portuguese version was automatically translated by the Deep Learning Brasil group[1], using the Google Cloud API. Similar approaches have been adopted in other languages, including Spanish, German and French (Carrino et al., 2020; Möller et al., 2021; d'Hoffschmidt et al., 2020). While automatic translation enables rapid dataset construction, prior work has shown that machine-translated corpora may fail to preserve cultural nuances and language-specific phenomena, potentially affecting fidelity and evaluation reliability (Singh et al., 2025; Kreutzer et al., 2022).

Inspired by the SQuAD format, Sayama et al. (2019) proposes FaQuAD, a reading comprehension dataset in Portuguese. Unlike traditional QA datasets based solely on question–answer pairs, FaQuAD emphasizes context-dependent answer extraction and allows multiple correct answer spans per question. It is composed by 249 contexts and 900 questions, which were extracted from official documents and Wikipedia pages regarding universities from Brazil.

Aligned with DBpedia Portuguese, de Araujo et al. (2020) presented a dataset in Portuguese for evaluating semantic QA, derived from an adaptation of the QALD collection, with the aim of enabling the evaluation of QA systems that operate on linked data.Finardi et al. (2023) proposes BACEN FAQ, a dataset that comprises frequently asked questions of the central bank of Brazil. Their work also highlights the need for open resources to foster research in the Brazilian NLP community.

Färber et al. (2025) discuss the limitations of using automatically translated data and the difficulty of demonstrating local contexts, and proposes MedPT, a public large-scale corpus for the Brazilian Portuguese medical domain. Ehlert et al. (2025); Rikters (2018); Niu et al. (2021) also discusses about how using translated datasets results in data that, while syntactically correct, may not reflect authentic language use.

Pirozelli et al. (2024) proposes a dataset for the Portuguese language (also bilingual in English) designed for reading comprehension and QA tasks on topics related to the ocean, the Brazilian coast, and climate change. The data in the dataset were extracted from scientific abstracts and reports . The dataset is composed of 2.26 thousand sets of contexts, questions, and answers, and also includes an extension with multiple-choice questions.

Despite these contributions, there remains a scarcity of publicly available datasets for Portuguese. In particular, existing resources rarely provide long, information-dense questions and answers aligned with realistic educational contexts, which limits their applicability to modern generative QA systems.

3 Methodology

This section discusses the methodology used for the development of the dataset proposed in this work. The methodology can be divided into four main parts: data collection, preprocessing and generation of questions and answers, followed by a human evaluation process of the generated data. Figure 1 presents the workflow of this methodology.

3.1 Data Acquisition

As the first step, the data was obtained by performing web scraping of the digital version of the book Natural Language Processing – Concepts, Techniques, and Applications in Portuguese (2nd Edition) (Caseli and Nunes, 2023). The book balances the presentation of historical knowledge and essential foundations of NLP, ensuring that even the oldest yet still relevant concepts are preserved and studied. At the same time, the book stays up to date by including recent techniques and applications, reflecting the state of the art in the field.

From the obtained HTML, each paragraph (tag <p>) was added to the dataset as a context. Other elements such as images, figures, equations, tables, footnotes, examples, and section introductions were removed, as well as paragraphs containing references to any of the aforementioned items.

3.2 Preprocessing

The dataset obtained via web scraping was cleaned using RegEX[2], removing contexts with fewer than 150 tokens, since examples below this threshold do not provide enough content to generate a relevant question or answer. Additionally, contexts with more than 950 tokens were removed to avoid model generation overflow. Finally, the prompt was formatted for the model as follows:

- *Given the context, generate a question and an answer. The answer to each question is a*

[1]Available at: http://www.deeplearningbrasil.com.br/

[2]Documentation available at: https://docs.python.org/3/library/re.html#module-re

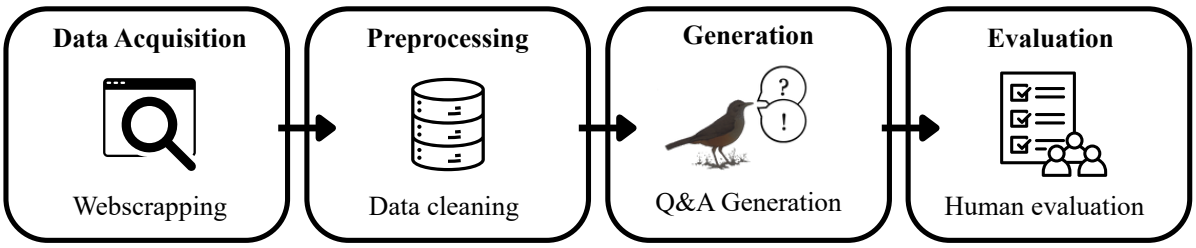

Figure 1: The four-stage pipeline used to construct the dataset: Data Acquisition, Preprocessing, Generation, and Evaluation.

segment from the corresponding context. The answer should be short and direct. ### Context: <context> ### Question:<completion>

3.3 Generation

For this step, we proposed a model based on Sabiá-7B (Pires et al., 2023). The model was quantized to reduce its size and complexity for deployment in resource-constrained environments. It was loaded in 4-bit format using the NF4 quantization type to better preserve information. To further reduce memory usage during fine-tuning, the LoRA method is applied alongside quantization. LoRA reduces the number of trainable parameters by introducing low-rank adaptations. Only 3.89% of the model's parameters (262 million out of 6.7 billion) were fine-tuned.

The fine-tuning was configured with a batch size of 1 for both training and evaluation to accommodate memory constraints. Gradient accumulation was not employed. We used the 32-bit AdamW optimizer with a learning rate of 2e-4 and a cosine learning rate scheduler. Instruction Fine-Tuning (IFT) was performed on the Pirá dataset (Pirozelli et al., 2024). In total, 1,624 question-answer pairs were automatically generated from the book's paragraphs.

It is worth noting that the Sabiá-7B model was primarily used in this research as a bootstrapping tool to mitigate human effort in generating candidate questions and answers. The main contribution of this work is not to attest to the already known high performance of models of this scale, but rather to present LARI as a open dataset validated by human experts.

3.4 Evaluation

To ensure the quality and factual accuracy of the dataset, we opted for expert human evaluation. Specifically, the evaluation was conducted by 39 domain experts, all of whom are contributing authors of the book.

The generated dataset was divided into chapters. Subsets corresponding to each chapter were evaluated by the respective annotator expert of that chapter, with a maximum of 25 questions and answers per annotator. Each annotator received a brief tutorial and a form for each subset. The form used a 1-to-5 scale (DeVellis, 2003) to assess the context, question, and answer, as well as a final eliminatory question regarding the overall quality of the question and answer. In total, the 39 participating annotators produced a final dataset of 688 entries.

4 Results

The results are organized according to the quality and relevance of the questions and answers, considering the Likert scale (DeVellis, 2003) used to measure contextual coherence, question formulation, and answer adequacy. Below, we detail the findings of this evaluation, focusing on the model's effectiveness in generating content aligned with the book's context and the quality of the entries that make up the final dataset.

First, 39 responses were obtained from the invited annotators. It is worth noting that annotators received at most 25 trios (context, question, and answer); therefore, not all of them received the exact amount described, as some chapters contained less usable text than others. In total, the resulting dataset comprised 688 trios. As part of the evaluation, an analysis was carried out regarding the overall adequacy and coherence of the generated questions and answers. Of the 688 evaluated trios, 464 (approximately 67%) were considered adequate and coherent, while 224 (33%) did not meet this expectation.

Regarding the context, most evaluations (316) assigned a score of 5, suggesting that the contexts

were considered excellent by the majority of evaluators. The second highest concentration was score 4 (113), indicating that many contexts were also evaluated as good. Only a small fraction of the contexts received low scores (2 and 1), suggesting that there were few cases in which the context did not meet the evaluators' expectations. The same trend was visible in the rest of the trios.

Overall, the results indicates that both the contexts and the questions and answers were well evaluated, with a clear majority of scores three, four and five. However, the questions and answers category showed a slightly broader distribution of lower scores (one and two), suggesting that in some cases, the questions and answers were not as consistent or satisfactory as the contexts.

The majority of the questions, representing 95.7% (or 444 questions), received a score equal to or higher than three, indicating that most questions were considered acceptable or better in quality by the evaluators. Only 4.3% of the questions, corresponding to twenty questions, received scores below three. The same pattern is visible in the equivalent sample for the answers. Overall, the performance in formulating questions and answers was predominantly positive, with only a small fraction failing to meet the expected quality criteria.

During the experiments, an issue observed was that the model struggled to generate questions and answers when the context lacked quality, whether in terms of length or content. Despite the web-scraping process having attempted to avoid paragraphs without sufficient standalone context (as its shown in the Section 3.2), some cases still compromised the quality of the remaining trio. For example:

> "There is still a type of morpheme (or phoneme, depending on the approach) that has not been explored here because it is not relevant to NLP studies: the linking vowels and consonants. They have no meaning, but are sometimes inserted between a root and a suffix or an affix for phonological reasons."

The context example above received a score of two, with the following observation from one of the annotators: *"The context refers to what is missing in the chapter."* The model's difficulty in generating appropriate questions and answers was reinforced by the author's observation, which pointed out that the context used referred to gaps in the chapter. This means the model was attempting to generate content based on incomplete or insufficient information, which compromised the quality of the generated question and answer.

During the experiments, it was observed that answer quality fluctuates significantly depending on the quality of the question. The results illustrates the positive relationship between the average scores assigned to questions and their corresponding answers, with a trend line indicating that as the perceived quality of the questions increases (higher average score), the answer quality also tends to be better evaluated. The linear relationship suggested by the trend line reveals a consistent positive correlation, where higher-rated questions are generally associated with equally well-rated answers. The absence of significant dispersion or notable outliers reinforces the hypothesis of a direct and predictable connection between the two variables.

4.1 Dataset Characteristics

The final dataset was post-processed so that questions and answers with low scores and flagged as invalid were removed. The distribution of questions and answers by chapter is presented in the chart in Figure 2.

The chapters "Information Extraction" *(Extração de Informação)* and "Textual Complexity and Related Tasks" *(Complexidade Textual e Suas Tarefas Relacionadas)* stand out with the highest number of valid questions and answers. In contrast, chapters such as "Ethical Issues in AI and NLP" *(Questões Éticas em IA e PLN)*, "Dialogue and Interactivity" *(Dialogo e Interatividade)* and "Evaluation of Language Technologies" *(Avaliação de Tecnologias de Linguagem)* show a low number of valid questions and answers. This outcome is related to the number of questions and answers generated for each chapter, where shorter chapters or those with less raw text received less content.

To further detail the linguistic properties of the resulting dataset, we computed length and vocabulary metrics across the final 464 context-question-answer triples. The extracted contexts have an average length of 125.34 words (ranging from 48 to 457). The formulated questions average 10.82 words, while the validated answers average 20.62 words. LARI demonstrates rich lexical diversity, encompassing a unique vocabulary of 6,795 tokens out of a total of 60,367 words.

Finally, the evaluation results presents that, although the model demonstrated competence in gen-

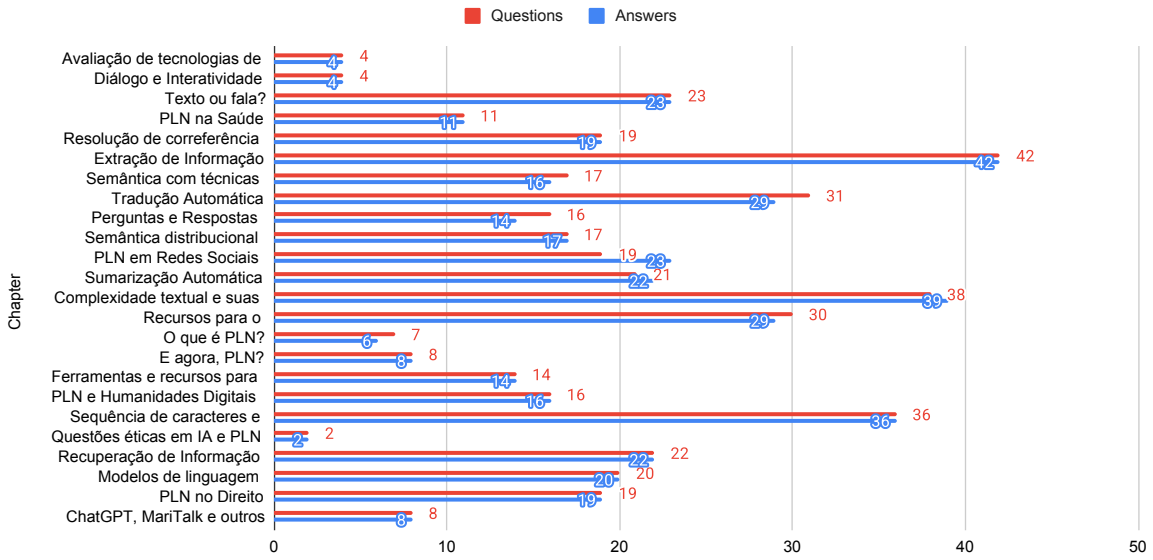

Figure 2: Distribution of valid (with the score given by the reviewers bigger than/equal to 3) questions and answers by chapter.

erating questions and answers, contextual adequacy and formulation quality still present significant difficulties. The analysis revealed that failures in contextualization compromised the clarity and relevance of the answers, suggesting that consistency and depth in the provided contexts are crucial for the model's effectiveness. Furthermore, the observations indicate that improvements can be made in the presentation format and accuracy of the generated questions and answers. Thus, the study underscores the need for continuous refinements in both the model and the initial dataset to ensure that the generated content is not only coherent but also aligned with expectations and demands.

Additionally, the final result for each item in the trio was given by the weighted average of the scores, where: the contexts achieved an average of 4.58, the questions reached an average of 4.48, and the answers reached an average of 4.37. Finally, the set achieved a score based on the simple average of these three items, 4.47.

5 Final Considerations

The automatic generation of questions and answers plays an important role in advancing human–computer interaction, particularly in educational settings, yet it remains challenging for low-resource languages such as Portuguese. This work addresses this gap by proposing a methodology for constructing QA-oriented datasets and by publicly releasing the LARI dataset, contributing a new resource to support research in Portuguese.

Human evaluation indicates that most context–question–answer trios are coherent, while also revealing a strong dependency between context quality and the quality of the generated questions and answers. Insufficient contexts lead to substantial degradation in the remaining components, whereas well-formulated questions tend to yield higher-quality answers.

We expect that both the proposed methodology and the LARI dataset will support future studies on QA in Portuguese. Future work includes exploring additional fine-tuning strategies, expanding the training corpus to increase content diversity, and investigating the applicability of this approach to domains beyond the current case study.

Acknowledgments

Below, we present a complete list of the respondents who participated in the evaluation process, whose collaboration was essential for the development of this study.

Aline Aver Vanin, Aline Macohin, Aline Paes, Ana Paula Banza, Arnaldo Candido Junior, Brenda Salenave Santana, Brielen Madureira, Cláudia Freitas, Crysttian Arantes Paixão, Daniela Barreiro Claro, Daniela Vianna, Edresson Casanova, Eduardo Cortes, Elisa Terumi Rubel Schneider, Eloize Seno, Evandro Fonseca, Fernanda Olival, Fla-

viane R. Fernandes Svartman, Helena de Medeiros Caseli, Helena Freire Cameron, Joaquim Santos, Larissa Freitas, Lucelene Lopes, Maria das Graças Volpe Nunes Maria José Bocorny Finatto, Marlo Souza, Paula Christina Figueira Cardoso, Priscila Osório Côrtes, Renata Vieira, Ricardo Marcacini, Sandra Maria Aluísio, Sheila Castilho, Sidney Evaldo Leal, Solange Rezende, Tayane Soares, Valéria de Paiva, Viviane P. Moreira, Vládia Pinheiro, Yohan Bonescki Gumiel.

We would like to express our sincere gratitude to the evaluators who kindly contributed their time and expertise to the completion of this research.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001, CNPQ, and Cenpes/Petrobras. We would like to thank the FAPERGS - Brasil for Financial Support, Award Agreement 22/2551-0000598-5. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research.

References

Ali Mohamed Nabil Allam and Mohamed Hassan Haggag. 2012. The question answering systems: A survey. *International Journal of Research and Reviews in Information Sciences (IJRRIS)*, 2(3).

Casimiro Pio Carrino, Marta R. Costa-jussà, and José A. R. Fonollosa. 2020. Automatic Spanish translation of SQuAD dataset for multi-lingual question answering. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5515–5523, Marseille, France. European Language Resources Association.

H. M. Caseli and M. G. V. Nunes, editors. 2023. *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*. BPLN. https://brasileiraspln.com/livro-pln.

Denis Andrei de Araujo, Sandro José Rigo, Paulo Quaresma, and João Henrique Muniz. 2020. A portuguese dataset for evaluation of semantic question answering. In *International Conference on Computational Processing of the Portuguese Language*, pages 217–227. Springer.

Robert F. DeVellis. 2003. Scale development: Theory and applications. *Sage London*, 13:0–176.

Martin d'Hoffschmidt, Wacim Belblidia, Quentin Heinrich, Tom Brendlé, and Maxime Vidal. 2020. Fquad: French question answering dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1193–1208.

Allan Duarte Ehlert, Júlia da Rocha Junqueira, Larissa Astrogildo de Freitas, and Ulisses Brisolara Corrêa. 2025. A review of recent advances on automatic question generation for the portuguese language. In *The International FLAIRS Conference Proceedings*, volume 38.

Paulo Finardi, Wanderley M. Melo, Edgard D. Medeiros Neto, Alex F. Mansano, Pablo B. Costa, and Vinicius F. Caridá. 2023. Portuguese faq for financial services. *Preprint*, arXiv:2311.11331.

Fernanda Bufon Färber, Iago Alves Brito, Julia Soares Dollis, Pedro Schindler Freire Brasil Ribeiro, Rafael Teixeira Sousa, and Arlindo Rodrigues Galvão Filho. 2025. Medpt: A massive medical question answering dataset for brazilian-portuguese speakers. *Preprint*, arXiv:2511.11878.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan Van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, and 1 others. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Timo Möller, Julian Risch, and Malte Pietsch. 2021. Germanquad and germandpr: Improving non-english question answering and passage retrieval. *arXiv preprint arXiv:2104.12741*.

Zhaoyang Niu, Guoqiang Zhong, and Hui Yu. 2021. A review on the attention mechanism of deep learning. *Neurocomputing*, 452:48–62.

Liangming Pan, Wenqiang Lei, Tat-Seng Chua, and Min-Yen Kan. 2019. Recent advances in neural question generation. *arXiv preprint arXiv:1905.08949*.

Ramon Pires, Hugo Abonizio, Thales Sales Almeida, and Rodrigo Nogueira. 2023. Sabiá: Portuguese large language models. In *Intelligent Systems*, pages 226–240, Cham. Springer Nature Switzerland.

Paulo Pirozelli, Marcos M José, Igor Silveira, Flávio Nakasato, Sarajane M Peres, Anarosa AF Brandão, Anna HR Costa, and Fabio G Cozman. 2024. Benchmarks for pirá 2.0, a reading comprehension dataset about the ocean, the brazilian coast, and climate change. *Data Intelligence*, 6(1):29–63.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. *arXiv e-prints*, arXiv:1606.05250.

Matīss Rikters. 2018. Impact of corpora quality on neural machine translation. In *Human language technologies–The Baltic perspective*, pages 126–133. IOS Press.

Vasile Rus, Brendan Wyse, Paul Piwek, Mihai Lintean, Svetlana Stoyanchev, and Cristian Moldovan. 2010. The first question generation shared task evaluation challenge.

Hélio Fonseca Sayama, Anderson Viçoso Araujo, and Eraldo Rezende Fernandes. 2019. Faquad: Reading comprehension dataset in the domain of brazilian higher education. In *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 443–448.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I. Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre F. T. Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2025. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 18761–18799.

Asahi Ushio, Fernando Alva-Manchego, and Jose Camacho-Collados. 2023. A practical toolkit for multilingual question and answer generation. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.