Zero-shot Definition Modelling for Portuguese

Edison Marrese-Taylor[1, 2], Erica Kido Shimomoto[1], Enrique Reid[2]
National Institute of Advanced Industrial Science and Technology[1]
Graduate School of Engineering, The University of Tokyo[2]
{edison.marrese,kidoshimomoto.e}@aist.go.jp, erarvin2007@gmail.com

Abstract

In this paper, we propose to study the task of definition modeling for a new language: Portuguese. To that end, we collect monolingual dictionary data and perform an in-depth empirical study to test the multilingual abilities of Large Language Models (LLMs), utilizing zero-shot and few-shot approaches. We analyze the performance of Llama-2 and Mistral LLMs on monosemic terms, as our collected data does not contain context information for polysemic words. To address this limitation, we further propose to utilize LLMs to generate usage examples that can assist definition modeling of polysemic terms in the future. To validate our findings, we performed a pilot human study to evaluate the quality of generated definitions and usage examples. Our results are encouraging and suggest that LLMs can generate plausible definitions of words in Portuguese and that the COMET metric aligns well with the human evaluation. Finally, our human study indicates that using LLMs could be an alternative to obtain context information of polysemic words in Portuguese.

1 Introduction

Definition modeling is the task of estimating the probability of a textual definition given a word [1]. This task has been shown to give an arguably more transparent view of the extent to which a model captures syntax and semantics.

This task can be framed as a conditional generation, in which the definition of the word or phrase is generated given a conditioning variable. So far, existing works have followed the traditional approach, where models are trained on a corpus of word-definition pairs, to be later tested on how well they generate definitions. These approaches [2, 3, 4, 5] are mainly encoder-decoder based, encoding contextual representation for a word/phrase and using the contextual representation to generate the definition.

Despite the progress, previous work has predominantly focused on the English language. Concretely, we find just a handful of papers that work on languages other than Eng lish, namely VCDM from [5] for French and the work of [6] for Chinese. We also find that dictionary data for French and Italian have been recently released, but there are so far no approaches that leverage them [7, 8].

In light of these issues, we present a study on definition modeling from the perspective of a new language, Portuguese, introducing a dataset total of approximately 191, 499 new terms entries and 279, 985 definitions, extending language coverage for the task.

Moreover, we note that the recent success of Large Language Models (LLMs) showed that such models can achieve excellent performance on a wide variety of downstream tasks, utilizing zero-shot or few-shot approaches [9]. Despite these results, we find limited work in assessing their multilingual capabilities. To shed light on this issue, and motivated by the original ideas of [1], we test a selection of such models, namely Llama-2 [10] and Mistral 7B [11], on our collected data in Portuguese, effectively proposing to use definition modeling tasks as a probe to test the multilingual abilities of LLMs.

Our results show that prompting techniques on LLMs, either in the zero-shot or few-shot scenarios, can be used successfully for definition modeling in Portuguese. However, we also observe that the output language can be challenging to control. Human evaluation of the generated definitions by a native Brazilian Portuguese speaker shows that while not tailored for this task, COMET scores can be a valid guide in the quality evaluation of such definitions.

Our experiments focused on monosemic terms, as our collected data does not contain context data; To address this limitation, we further propose to utilize LLMs to generate usage examples that can assist definition modeling of polysemic words. We prompted GPT-3.5 [9] to generate sentences using polysemic terms in different contexts and evaluated their quality through a human evaluation. Our results indicate that using LLMs could be an alternative to obtain such example sentences for Portuguese.

2 Related Work

Our paper is primarily related to the seminal work by Noraset et al. [1] and Hill et al. [12], in which a model is tasked with generating a definition for a word given its respective embedding, or with mapping dictionary definitions to lexical representations of words, respectively. Later work has proposed improvements and extensions, introducing techniques and datasets to address shortcomings. For example, Gatedesky et al. [2] address polysemy and present a dataset from Oxford Dictionaries, where each definition is also supplemented with context sentences. Ni et al. [3] proposed an approach for automatically explaining slang English terms in a sentence and introduced another dataset from Wikipedia. Ishiwatari et al. [4] proposed to further rely on local and global contexts to disambiguate and generate better definitions, also introducing a dataset based on Cambridge Dictionaries and a dataset for French.

More recently, Huang et al. [13] studied the problem of definition specificity, tuning a model to account for hyper-focused or highly general definitions. Finally, Chen et al. [14] recently proposed to unify the seminal ideas of reverse dictionary and definition modeling in a single model to help better understand word sense and embeddings.

Previous work discussed so far has mostly focused on definitions in the English language. One main exception in this context is the work of Reid et al. [5], who presented the first study on definition modelling for the French language with the release of a dataset collected from Le Petit Robert.

Finally, we note that over the past few years, dictionary datasets in several languages derived from Wikictionary have been released, including English (ENGLAWI) [15], French [7], and Italian [8]. We find, however, that these datasets are not accompanied by models that leverage them. Moreover, an important distinction in this regard is that these dictionaries are built on the base of crowdsourcing, where quality could be a concern.

3 Data

Our interest in the Portuguese language derives from its importance in terms of the number of native speakers, as Portuguese is listed among the top-5 most spoken native language in the world and has been previously regarded as one of the ten most influential languages in the world [16]. Portuguese also shows in the top three Indo-European languages with the largest number of speakers according to Wikipedia, with 236 M native speakers.

We choose a readily available dictionary to serve as a source for our dataset, Dicio[1]. We collect the term and the available definitions. Table 1 summarizes the main characteristics of the collected data, compared against existing resources in English and French. We can see that our data is substantially richer.

Table 1: Summary of our collected dataset, compared to prior relevant corpora. In the table, Mono. stands for Monosemic, i.e. terms with a single definition.

Dataset	Terms	Defs.	Mono.
Oxford - en	36,767	122,319	44.07%
Le Petit Robert - fr	33,507	-	-
Dicio - pt (ours)	191,499	279,985	77.41%

An important distinction between our corpus and recent prior work is that our collected data does not contain examples of word usage. Previous research has shown that this information is critical in allowing models to disambiguate a specific meaning for a given term in the case of polysemy. With this in mind, for our experiments in this paper, we select the subset of terms that exhibit only a single meaning and thus experiment in a *monosemic* scenario.

We also note that our data contains multiple entries for the same term, as inflections of verbs (e.g., tenses) and adjectives (e.g., gender) are present in the dictionary. This problem leads to an artificially inflated amount of available data. To alleviate this issue, we rely on spacy[2] to identify word lemmas, utilizing the "*core_news_sm*" model for Portuguese, keeping only the original term matched the lemmatized word. The resulting dataset is split into the 80/10/10 ratio. Table 2 below shows the exact details of the split sizes.

4 Experiments

Since our interest is to utilize our task as a proxy to better understand the multilingual abilities of Large Language

[1] https://www.dicio.com.br/

[2] https://spacy.io/

Table 2: Details of the size of each split for our collected data, compared against the Oxford dataset.

Dataset	Train	Valid	Test
Oxford (en)	15,770	6,884	6,834
Dicio (pt)	118,591	14,824	14,824

models, our approach for definition modeling is based on prompting and in-context learning. Regarding models, we consider Llama2 [10], specifically the chat versions, and the recently-released Mistral models [11]. For the former, we utilize the 13B-parameter models. The latter we quantize to 4-bits using QLoRA [17] in order to fit into our GPU memory. We test two settings, as detailed below:

1. A zero-shot approach, where the model is directly asked to generate the definition of the word
2. A 5-shot setting, where we incorporate term-definition examples in the prompt before requesting the definition for the target term. These shots are randomly sampled from the training data and kept constant across examples.

In all cases, the input to the model is *"Define the {language} word '{term}'. Use only {language} to reply."*, where {language} and {term} are variables denoting the target language and the term to define.

Regarding evaluation, previous work in definition modeling has mainly used n-gram overall metrics such as BLEU [18] and METEOR [19]. As the latter is language-specific, here we report BLEU, relying on the sacrebleu[3] implementation [20]. As metrics based on n-gram overlap do not capture nuance in the generations [5, 13], we follow works that adopted machine learning-based metrics and experiment with COMET [21][4].

Furthermore, to evaluate the ability of the LLMs to generate definitions in the target language, we rely on a fasttext-based language classification model [22], a linear model based on character n-grams. It can recognize 176 languages and was trained on 400 million tokens from Wikipedia and sentences from the Tatoeba website.

Finally, we perform human evaluation on a subset of the definitions generated by each model and setting. We recruit a native speaker of Brazilian Portuguese to evaluate a set of 100 generated definitions by each of the two chosen models in both settings, where we picked the 50 best and 50 worst generations based on the COMET metric, resulting in a total of 400 definitions. The speaker has to evaluate the quality of the generated definitions based on the following Likert scale: 1 - Very poor: A definition of a completely different word / Wrong definition; 2 - Poor: A definition of a related term; 3 - Acceptable: A vague definition of the term; 4 - Good: A definition of the term, but with few mistakes; 5 - Very Good: A correct definition of the term.

4.1 Results

Table 3: Results of our experiments using LLMs for zero-shot and few-shot definition modelling in Portuguese. In the table, Comp. is short for compliance, the % of cases where the answer is in the correct language.

Model	Performance		
	BLEU	COMET	Comp.
Llama-2-13b-chat	0.162	0.513	**0.968**
+ 5 shots	**3.460**	**0.527**	0.962
Mistral-7B-Instruct	0.140	0.475	0.625
+ 5 shots	3.034	0.488	0.898

Table 3 summarizes the results of our experiments. We see that providing the model with examples tends to lead to better performance, as expected. However, an important issue here is that LLMs are sometimes unable to follow instructions, which in our case often leads to the model generating outputs in English. We also observed that models often hallucinate definitions which may "feel" correct, but in fact are not. Better mechanisms to control these kinds of behaviors are required.

Table 4 shows the human evaluation results of the sampled generated definitions. First, we can observe that the Likert score on the Best sets is higher than on the Worst sets, which shows that the COMET metric is well aligned with the human evaluation. The overall positive Pearson correlation between these values supports this point. We also observe that the correlation in the Worst sets is overall smaller. This result is likely because while COMET varied between values of 0.2 to 0.4, these definitions mostly were tagged as "Very poor". Furthermore, we observe that the correlation in the Best set in the zero-shot scenario is almost 0 for Mistral and negative for Llama2. In this scenario, we

[3] https://github.com/mjpost/sacrebleu
[4] We use the "*wmt22-comet-da*" model, which has multilingual support.

Table 4: Results of the human evaluation of the generated definitions. Corr. is short for the Pearson correlation between Likert score and COMET values.

Model	Set	Likert	COMET	Corr.
Llama-2-13b-chat	Best	3.860	0.639	-0.224
	Worst	1.240	0.385	0.118
	All	2.550	**0.512**	0.752
+ 5 shots	Best	4.000	0.623	0.623
	Worst	1.460	0.339	0.102
	All	**2.730**	0.481	0.721
Mistral-7B-Instruct	Best	3.260	0.577	0.089
	Worst	1.140	0.320	0.181
	All	2.200	0.448	0.633
+ 5 shots	Best	1.780	0.530	0.186
	Worst	1.180	0.306	-0.053
	All	1.480	0.418	0.897

noticed that models generate very long definitions. For Mistral, these generations were rather “Very good”, with rich details, or “Very Poor”, containing wrong information; however, COMET could not detect such differences. Llama2 generations were more precise and correct, but we hypothesize that COMET could not correctly evaluate them due to the length difference between reference and generated definitions.

4.2 Tackling Polysemy

Our results are limited to the set of terms for which we only have one definition. Though our results suggest that LLMs can, to some extent, generate plausible definitions for words in Portuguese, our empirical study offers no insight into the more challenging scenario of polysemy.

Therefore, we propose to utilize LLMs to obtain example sentences for polysemic terms in our dictionary. We run a pilot study using GPT-3.5 [9] to generate such example sentences. Concretely, we prompt the model as follows: *In the Portuguese Language, the word ’{term}’ can mean “{definition}”. Please give me a sentence in that language where this word is used.*, where {term} and {definition} are placeholders for variables denoting a given pair of a term and its corresponding definition.

For our study, we sample a subset of 59 terms from our data, with a total of 156 different meanings. We utilize the official OpenAI API to feed this data into the model. To assess the viability of this approach, we subject these sentences to a thorough human evaluation. Specifically, we ask a native Brazilian Portuguese speaker to evaluate the generated sentences according to the following simplified Likert scale: 1 - Poor: The sentence uses the term with a different meaning; 2 - Acceptable: The sentence uses the term with the specified meaning, but in the wrong context/sounds unnatural; 3 - Good: The sentence uses the term with the specified meaning.

Table 5: Results of the human evaluation of the generated example sentences. The numbers indicate the percentage of sentences evaluated in each category

Poor	Acceptable	Good
7.69	16.67	75.64

The results of this evaluation are in Table 5. We can see that most of the generated examples received the “Good” score, with an average of 2.68, which indicates that using this specific LLM could be an alternative to obtain such example sentences for Brazilian Portuguese.

5 Conclusions

In this paper, we present a study on definition modeling for the Portuguese language, which we propose as a proxy to assess the multilingual abilities of Large Language Models. Our encouraging results suggest that LLMs can generate plausible definitions of words in Portuguese. Furthermore, our human evaluation showed that the COMET metric could be used as a guide for generation quality and that prompting LLMs could be an alternative to obtain context information on polysemic terms in Portuguese.

For future work, we would like to incorporate more languages into our study and broaden our human evaluation scope to understand better how these models behave. We also plan to run LLM finetuning experiments on our data to test if this leads to better performance and behavior.

References

[1] Thanapon Noraset, Chen Liang, Larry Birnbaum, and Doug Downey. Definition modeling: Learning to define word embeddings in natural language. In **Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence**, AAAI’17, pp. 3259–3266, San Francisco, California, USA, February 2017. AAAI Press.

[2] Artyom Gadetsky, Ilya Yakubovskiy, and Dmitry Vetrov. Conditional generators of words definitions. In **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**, pp. 266–271, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[3] Ke Ni and William Yang Wang. Learning to Explain Non-Standard English Words and Phrases. In **Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)**, pp. 413–417, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing.

[4] Shonosuke Ishiwatari, Hiroaki Hayashi, Naoki Yoshinaga, Graham Neubig, Shoetsu Sato, Masashi Toyoda, and Masaru Kitsuregawa. Learning to Describe Unknown Phrases with Local and Global Contexts. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 3467–3476, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[5] Machel Reid, Edison Marrese-Taylor, and Yutaka Matsuo. VCDM: Leveraging Variational Bi-encoding and Deep Contextualized Word Representations for Improved Definition Modeling. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 6331–6344, Online, November 2020. Association for Computational Linguistics.

[6] Cunliang Kong, Liner Yang, Tianzuo Zhang, Qinan Fan, Zhenghao Liu, Yun Chen, and Erhong Yang. Toward Cross-Lingual Definition Generation for Language Learners, October 2020.

[7] Nabil Hathout and Franck Sajous. Wiktionnaire's Wikicode GLAWIfied: A Workable French Machine-Readable Dictionary. In **Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)**, pp. 1369–1376, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA).

[8] Basilio Calderone, Matteo Pascoli, Franck Sajous, and Nabil Hathout. Hybrid Method for Stress Prediction Applied to GLAFF-IT, a Large-Scale Italian Lexicon. In **Language, Data, and Knowledge**, Lecture Notes in Computer Science, pp. 26–41, Cham, 2017. Springer International Publishing.

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. In **Advances in Neural Information Processing Systems**, Vol. 33, pp. 1877–1901. Curran Associates, Inc., 2020.

[10] Hugo Touvron et al. Llama 2: Open foundation and fine-tuned chat models, 2023.

[11] Albert Q. Jiang et al. Mistral 7b, 2023.

[12] Felix Hill, Kyunghyun Cho, Anna Korhonen, and Yoshua Bengio. Learning to understand phrases by embedding the dictionary. **Transactions of the Association for Computational Linguistics**, Vol. 4, pp. 17–30, 2016.

[13] Han Huang, Tomoyuki Kajiwara, and Yuki Arase. Definition Modelling for Appropriate Specificity. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 2499–2509, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[14] Pinzhen Chen and Zheng Zhao. A Unified Model for Reverse Dictionary and Definition Modelling. In **Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)**, pp. 8–13, Online only, November 2022. Association for Computational Linguistics.

[15] Franck Sajous, Basilio Calderone, and Nabil Hathout. ENGLAWI: From Human- to Machine-Readable Wiktionary. In **12th International Conference on Language Resources and Evaluation (LREC 2020)**, p. 3016, May 2020.

[16] George Weber. The world's 10 most influential languages, May 2011.

[17] Tim Dettmers et al. Qlora: Efficient finetuning of quantized llms, 2023.

[18] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[19] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In **Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization**, pp. 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[20] Matt Post. A call for clarity in reporting BLEU scores. In **Proceedings of the Third Conference on Machine Translation: Research Papers**, pp. 186–191, Brussels, Belgium, October 2018. Association for Computational Linguistics.

[21] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A Neural Framework for MT Evaluation. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 2685–2702, Online, November 2020. Association for Computational Linguistics.

[22] Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. Learning word vectors for 157 languages. In **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).