DORE: A Dataset For Portuguese Definition Generation

Anna Beatriz Dimas Furtado[1], Tharindu Ranasinghe[2], Frédéric Blain[3], Ruslan Mitkov[4]
[1]University of Galway, IE, [2]Aston University, UK,
[3]Tilburg University, NL, [4]Lancaster University, UK
annabeatriz.dimasfurtado@universityofgalway.ie, t.ranasinghe@aston.ac.uk,
f.l.g.blain@tilburguniversity.edu, r.mitkov@lancaster.ac.uk

Abstract

Definition modelling (DM) is the task of automatically generating a dictionary definition for a specific word. Computational systems that are capable of DM can have numerous applications benefiting a wide range of audiences. As DM is considered a supervised natural language generation problem, these systems require large annotated datasets to train the machine learning (ML) models. Several DM datasets have been released for English and other high-resource languages. While Portuguese is considered a mid/high-resource language in most natural language processing tasks and is spoken by more than 200 million native speakers, there is no DM dataset available for Portuguese. In this research, we fill this gap by introducing DORE; the first dataset for **D**efinition M**O**delling for Po**R**tugu**E**se containing more than 100,000 definitions. We also evaluate several deep learning based DM models on DORE and report the results. The dataset and the findings of this paper will facilitate research and study of Portuguese in wider contexts.

Keywords: Portuguese dataset, automatic generation of definitions, definition modelling, transfer learning, pretrained models.

1. Introduction

Definitions play a key role in the globalised world; they are useful for a wide range of audiences, from professionals to students (Dziemianko, 2020). They are also the building blocks of effective communication and understanding within the Information Society; it is imperative to have domain experts to ensure their accuracy, clarity, and coherence, which can be expensive (San Martín, 2021). Yet, crafting high-quality definitions demands time and effort due to their intricate and complex nature (Domínguez Vázquez and Gouws, 2023). Given the array of challenges involved, the manual creation of definitions proves to be a difficult, expensive, and arduous task (San Martín, 2021).

Introduced by Noraset et al. (2017) as a supervised machine learning (ML) task (Ni and Wang, 2017), definition modelling (DM) addresses these challenges by designing systems capable of automatically generating definitions for a specific word. Beyond its immediate application of generating definitions for dictionaries, DM can be useful for completing the WordNet, providing resources for language learners, language preservation and language description. Furthermore, it has been used as a window in explainable AI to shed the light on the quality of embeddings (Mickus et al., 2022), besides to provide results for studying LLM hallucinations and overgeneration mistakes [1]. It has also been suggested as a way to detach word-sense disambiguation from word inventories (Bevilacqua et al., 2020).

Most studies consider DM as a natural language generation (NLG) task, such as machine translation (Dabre et al., 2020), in which models are trained on annotated datasets consisting of words and their corresponding explanations (Mickus et al., 2019; Gadetsky et al., 2018). Furthermore, as DM was born in the deep learning (DL) era, many DM approaches followed DL models, such as sequence-to-sequence architectures that require a myriad of annotated instances to train their weights properly. Hence, DM algorithms depend on the availability of large annotated datasets.

Considering the importance of annotated data, numerous datasets have been established for the English (Gadetsky et al., 2018; Mickus et al., 2019; Li et al., 2020). Recently, DM datasets have also been released for other languages, including Chinese (Chang and Chen, 2019), French, German, Greek, Italian (Kabiri and Cook, 2020) and Spanish (Mickus et al., 2022). The recent shared task, Semeval-2022 Task 1: CODWOE – Comparing Dictionaries and Word Embeddings (Mickus et al., 2022), has also contributed to the creation of other datasets. However, to the best of our knowledge, no DM dataset currently exists for Portuguese. In this research, we fill this gap by releasing DORE; the first dataset for **D**efinition M**O**delling in Po**R**tugu**E**se.

Portuguese is largely spoken officially on five

[1]As proposed by Zosa et al. in https://helsinki-nlp.github.io/shroom/

LREC-COLING 2024, pages 5315–5322
20-25 May, 2024.

continents, in seven countries, including Brazil and Portugal, and as a second language by more than 25 million people worldwide. Therefore, research in DM for Portuguese will be highly beneficial for millions of people, for which we lay the foundation through this paper by creating the first-ever Portuguese DM dataset, DORE. We also experiment with several DM methods on DORE. First, DM is performed as a sequence-to-sequence task using recent neural architectures. Then, we evaluate several popular large language models (LLMs), such as LLAMA2 and Falcon on Portuguese DM, using prompting. As they follow a zero-shot approach and do not need a training set, our findings can benefit a multitude of low-resource languages in definition modelling. As far as we know, this is the first time that LLMs are evaluated on low-resource DM.

Our **main contributions** can be summarised as follows:

(1) We introduce DORE, the first dataset for Portuguese definition modelling, which comprises 103,019 definitions, and we describe the steps taken to compile it.

(2) We evaluate several neural DM methods on DORE and report the results.

(3) For the first time, we evaluate several popular LLMs on DM. We use prompting to generate definitions and compare the results.

(4) We released DORE[2], as an open-access dataset alongside the trained machine-learning models.

2. Related Work

Datasets Definition Modelling (DM) has gained prominence as a deep learning problem, primarily due to its challenging nature. As mentioned before, most DM approaches have relied on supervised ML algorithms in which models are trained on annotated datasets. As a result, the NLP community has a growing interest in creating and collating datasets for DM. Noraset et al. (2017) made available the first English dataset for the DM task, composed of definitions extracted from the Oxford Dictionary. Several English datasets were released in the following years. Gadetsky et al. (2018) and Zhang et al. (2020) improved the dataset by Noraset et al. (2017) by adding more instances. Ishiwatari et al. (2019) released a DM dataset based on Wikipedia and Wikidata,

DM datasets have been proposed to other languages as well. Kabiri and Cook (2020) released the first multilingual DM dataset, including Dutch, English, French, German, Greek, Italian, Japanese, Russian and Spanish. They utilised Wiktionary, OmegaWiki, and WordNet to extract the definitions. Furthermore, Yang et al. (2020) created the CDM dataset for the Chinese definition modelling task, where the definitions were extracted from the Chinese Concept Dictionary. As mentioned before, Semeval-2022 Task 1 (Mickus et al., 2022) also contributed to creating several DM datasets in several languages, including English, Spanish, French, Italian and Russian. Huang et al. (2022) further advances DM with a dataset for Japanese. However, as far as we know, there is no DM dataset available for Portuguese.

Methods In DM's introductory paper, Noraset et al. (2017) presented an RNN-based model with an update function inspired by GRU gates to tackle word-to-sequence DM. The absence of relevant local contexts, however, hindered the production of definitions for polysemous words. To tackle this problem, Gadetsky et al. (2018) put forth two models that include contextual information for the first time. Later, Ishiwatari et al. (2019) use local context (co-text) and global context (external information) to generate unknown definitions. They employ an LSTM-based encoder-decoder model and confirm that the generation task becomes harder when the words become more ambiguous and polysemous. Contrastively, Mickus et al. (2019) recast DM as a sequence-to-sequence task rather than a word-to-sequence task; that is, context should be given as an input instead of the lemma.

Lately, transformer models revolutionised the NLP tasks (Devlin et al., 2019), and also DM. Bevilacqua et al. (2020) leverage BART (Lewis et al., 2020) for tackling DM with Word-Sense Disambiguation and Word-in-Context tasks. Similarly, Huang et al. (2021) use a T5 (Raffel et al., 2020) model to improve DM results significantly in several benchmarks. Finally, Zhang et al. (2023) explore generating bilingual definitions in English-Chinese by fine-tuning a pretrained multilingual machine translation model coupled with the exploitation of prompt combination and contrastive prompt learning. The model generates readable definitions but still produces hallucinations.

CODWOE Shared Task (Mickus et al., 2022) CODWOE focuses on generating glosses from vectors (DM track) and reconstructing embeddings from glosses (Reverse Dictionary track). They provided a second multilingual DM dataset, including English, French, Spanish, Italian, German, and Russian. Participants were encouraged to explore the potential benefits of multilingual and cross-lingual learning.

[2] https://huggingface.co/datasets/multidefmod/dore

Dictionary	N. of Senses	Scraping	Context	Research use
Dicionário Michaelis	350,000	No	Partially	No
Dicionário Houaiss	376,500	No	Partially	No
Dicionário Aulete	818,000	No	Partially	No
Dicionário Priberam	100,000	No	Unordered examples	Yes
Oxford Português	146,000	Unknown	Partially	Yes
Dicio	400,000	Yes (Request)	Unordered examples	Yes
Portuguese Wiktionary	> 270,501	Yes	For some entries	Yes

Table 1: Summary of potential data sources and their features

3. Dataset Construction

3.1. Data Collection

To collect data, we began by conducting extensive research into potential data sources, carefully evaluating their copyright status and quality. While DM typically relies on dictionary data, practical challenges in accessing these resources, as detailed in the following subsection, often make it necessary to leverage existing other resources. Subsequently, we extracted data from sources that aligned with our criteria.

3.2. DORE dataset

Definition Modelling relies on two primary resources: definitions and contexts, both typically found in dictionaries. Fortunately, recent technological advancements have made electronic dictionaries readily available, obviating the necessity for digitising printed materials. For the Portuguese language, surprisingly, e-dictionaries present unordered examples, which makes it challenging for readers (and machines) to connect them to corresponding senses.

Concerning the Portuguese, there are at least seven free monolingual e-dictionaries available for online consultation. They are: Michaelis, Houaiss, Aulete, Priberam, Portuguese Oxford (entries embedded into the Google search engine), Dicio and Portuguese Wiktionary. We survey these resources primarily because they are freely accessible and open to the public.

Table 1 summarises potential data sources and key features for this research, such as the number of senses, permission to scrape, research use permission, and the availability of the contexts. However, due to permission restrictions, we were only able to retrieve data from *Dicio* and *Portuguese Wiktionary*.

Dicio is a free e-dictionary that contains more than 400,000 senses. Entries include grammatical information (part of speech, plural form, etc.), definitions, and examples (occasionally). Dicio attempts to represent the contemporary Portuguese language and is conducive for research purposes.

Wiktionary is an online, crowdsourced dictionary aiming at becoming the universal polyglot dictionary. It covers more than 900 languages and features definitions, examples of use (occasionally), grammatical information (i.e., gender), and domain of use. For Portuguese, it contains more than 100,000 entries covering multiple varieties of the language.

To obtain data from the dictionaries, we employed a Python script to perform web scraping on each website. One notable challenge we encountered was the absence of comprehensive entry lists on these dictionary websites. Consequently, we resorted to employing word lists to generate the necessary URLs for data retrieval. The word lists used for creating the URLs in this dataset were sourced from *Wiktionary* dumps provided by Kaikki's Project[3] and word lists made available by Dicio.

In Table 3, we compare DORE with the other language resources available for definition modelling task in other languages. For the English dataset, we combined the data from the Oxford Dictionary (Gadetsky et al., 2018), the GCIDE and Wordnet dataset (Noraset et al., 2017), the Wiktionary, Omega and Wordnet collected by Kabiri and Cook (2020), and the CODWOE shared task (Mickus et al., 2022). For the other languages, we combined data proposed by Kabiri and Cook (2020) and the CODWOE shared task (Mickus et al., 2022). Although the results show that English boasts abundant resources for the DM task, featuring a vast number of instances, it is important to note that DORE has a compatible number of instances with other languages. Finallt, Table 2 shows examples of definitions from the DORE dataset with our respective translations.

4. Methods

In order to test the suitability of DORE for definition modelling, we exploit several deep learning models which are state-of-the-art in DM (Section 2).

We first divided DORE into a training and test set following a 0.8 split of the complete dataset.

[3] https://kaikki.org

Lemma	Definition
Abacaxi (Pineapple)	Planta originária do Brasil cultivada em muitas regiões quentes por causa de suas frutas de polpa açucarada e saborosa. (A plant native to Brazil, cultivated in many warm regions due to its sweet and tasty pulp.)
Abacaxi (Pineapple)	[gíria] pessoa ou coisa maçante, complicada ou desagradável. ([slang] a boring, unpleasant person or situation.)
Florescência (Flowering)	[Botânica] Situação em que uma flor está no processo de maturação; antese. ([Botany] Situation in which a flower is in the process of ripening; anthesis.)
Florescência (Taurean)	[Figurado] Forte, como um touro. ([Figurative] Strong, similar to a bull.)
Taurino (Flowering)	Ação ou efeito de florescer; florescimento. (The act or effect of blooming; blossoming.)
Desopilar (Distract)	[Figurado] Afastar da mente as preocupações ou os problemas; alegrar-se, divertir-se. ([Figurative] Keep worries or problems out of the mind; rejoice, have fun.)

Table 2: Instances of DORE dataset. English translations are in blue.

Lag.	Lemmas	Unique	Senses	Avg char.	Avg words
EN	877,001	509,994	1.72	56.04	9.46
FR	200,880	55,068	6.2	75.63	14.30
ES	75,057	33,860	2.12	80.84	14.75
IT	62,465	35,987	1.73	78.97	13.61
PT	103,019	27,978	5.43	72.38	11.38

Table 3: Dataset statistics featuring language, number of instances, number of unique instances, number of senses per lemma, average number of characters per definition, and average number of words per definition, respectively.

Model	BLEU	TER	BLEURT	BERTScore
mBERT	0.18	0.78	0.52	0.61
XLM-R Large	0.22	0.75	0.54	0.62
BERTimbau Large	0.16	0.81	0.51	0.60
mBART	0.25	0.73	0.61	0.69
mT5 Base	0.24	0.75	0.60	0.68
mT5 Large	0.27	0.73	0.63	0.70
GPT	0.37	0.68	**0.68**	**0.76**
Falcon 7B	0.31	0.71	0.62	0.74
Llama 2 7B	0.32	0.70	0.64	0.72

Table 4: The result of different ML models in DORE test set by the different ML architectures

Following machine learning models were used. We group models according to their architectures:

General Transformers - We created a Seq2Seq model from general transformers by adding a transformer decoder, which takes the encoder's output and generates the target sequences. We only used the same transformer as the encoder and decoder. We experimented with several general-purpose transformer models that support Portuguese, including mBERT (Devlin et al., 2019), XLM-Roberta (Conneau et al., 2020), and BERTimbau-large (Souza et al., 2020).

Text Generation Transformers - We also experimented with several text generation transformers as they have provided excellent results in English DM tasks. Specifically, we explored mBART (Lewis et al., 2020) and several mT5 (Xue et al., 2021) variants.

For both types of transformer models, we employed a batch size of 16, Adam optimiser with learning rate $1e{-4}$, and a linear learning rate warm-up over 10% of the training data. During the training process, the parameters of the transformer model were updated. The models were trained only using the training data and evaluated while training using an evaluation set that had one-fifth of the rows in training data. We performed early stopping if the evaluation loss did not improve over three evaluation steps. All the models were trained for three epochs.

LLMs Finally, we evaluate how LLMs perform in DORE, a recent trend as we discussed before. We used two prompts to get a response from LLMs. For the instances where the context was available, we used the following prompt: "Provide the definition of *{WORD}* appearing in this context *{CONTEXT}* in Portuguese".

For the instances where the context was not available, we used the following prompt: "Provide the definition of *{WORD}* in Portuguese."

We used several LLMs for prompting. We first use Davinci-003 through OpenAI API (Brown et al., 2020). Additionally, we used Falcon-7B-Instruct (Almazrouei et al., 2023) and Llama-2-7B-32K-Instruct (Touvron et al., 2023). All of these models are available in HuggingFace (Wolf et al., 2020), and we use the LangChain implementation. As we followed a zero-shot prompting approach, we did not use any instances for the training set for LLMs.

5. Results

The results of the aforementioned models are shown in Table 4. All the models were evaluated using the test set. We used several evaluation metrics to compare the models, BLEU (Papineni et al., 2002), and TER (Snover et al., 2006). However, both of these metrics lack semantic understanding. Therefore, we also used two recent NLG evaluation metrics, BLEURT (Sellam et al., 2020) and BERTScore (Zhang et al., 2019).

BLEU and TER, commonly known as NLG metrics, report low results for all groups, which resonates with previous NLG and DM investigations in that current NLG metrics are not satisfactory (Mickus et al., 2022).

Unsurprisingly, the best performing group is the LLMs, probably due to their incomparable parameter size and pre-embedded encyclopedic knowledge. GPT outperforms other LLMs slightly. It is worth noting that text generation transformers closely trail the LLM results, with the mT5 Large model surpassing its base variant and mBART. These numbers are compatible with those obtained in experiments with smaller datasets in other romance languages, such as French, Spanish and Italian (Mickus et al., 2022). However, BERTimbau Large model, which is explicitly trained on Portuguese text, provides the worst results from the experimented models. Overall, the results demonstrate that language models designed explicitly for text generation excel in the DM task in Portuguese, even when they are multilingual.

6. Conclusion

We introduced DORE, the first dataset for automatic generation of definitions in Portuguese. We demonstrate DORE's usefulness by performing Definition Modelling for Portuguese for the first time with several pretrained models together with popular LLMs. The results show that LLMs perform better in Portuguese DM. We released DORE and our code publicly with a view to fostering more research on various tasks in Portuguese.

As future work, we intend to expand DORE with more instances of definitions. We also plan to include in-context examples of lemmas, which can be useful for future experiments and other NLP tasks, such as word sense disambiguation and word in context. Besides that, we also plan to harness other datasets to perform cross-lingual learning.

Acknowledgments

The computational experiments in this paper were conducted on the Aston EPS Machine Learning Server, funded by the EPSRC Core Equipment Fund, Grant EP/V036106/1.

Ethics Statement

As mentioned in 3, the Data section, DORE was collected from publicly available resources, and none of the definitions were edited. We sought permission from Dicio to use definitions for this research. Similar to previous research, we shared the definitions and their lemmas. Also, we released DORE and corresponding models under the Creative Commons Attribution-Non Commercial-ShareAlike (CC-BY-NC-SA) 4.0 International Public License, which prevents users from editing any instances of the dataset. While DORE and related models are publicly available, we released it as a gated dataset so that users need to comply with the license to request access. We reinforce that models and dataset should be used for research only.

7. Bibliographical References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. 2023. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*.

Michele Bevilacqua, Marco Maru, and Roberto Navigli. 2020. Generationary or "how we went beyond word sense inventories and learned to gloss". In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7207–7221.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Ting-Yun Chang and Yun-Nung Chen. 2019. What does this word mean? explaining contextualized embeddings with natural language definition. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6064–6070.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020.

Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A survey of multilingual neural machine translation. *ACM Comput. Surv.*, 53(5).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

María José Domínguez Vázquez and Rufus H Gouws. 2023. The Definition, Presentation and Automatic Generation of Contextual Data in Lexicography. *International Journal of Lexicography*, page ecac020.

Anna Dziemianko. 2020. Smart advertising and online dictionary usefulness. *International Journal of Lexicography*, 33(4):377–403.

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

Han Huang, Tomoyuki Kajiwara, and Yuki Arase. 2021. Definition modelling for appropriate specificity. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2499–2509, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shonosuke Ishiwatari, Hiroaki Hayashi, Naoki Yoshinaga, Graham Neubig, Shoetsu Sato, Masashi Toyoda, and Masaru Kitsuregawa. 2019. 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)*, pages 3467–3476.

Arman Kabiri and Paul Cook. 2020. Evaluating a multi-sense definition generation model for multiple languages. In *International Conference on Text, Speech, and Dialogue*, pages 153–161. Springer.

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.

Yinqiao Li, Chi Hu, Yuhao Zhang, Nuo Xu, Yufan Jiang, Tong Xiao, Jingbo Zhu, Tongran Liu, and Changliang Li. 2020. Learning architectures from an extended search space for language modeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6629–6639, Online. Association for Computational Linguistics.

Timothee Mickus, Denis Paperno, and Matthieu Constant. 2019. Mark my word: A sequence-to-sequence approach to definition modeling. In *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*, pages 1–11, Turku, Finland. Linköping University Electronic Press.

Timothee Mickus, Kees Van Deemter, Mathieu Constant, and Denis Paperno. 2022. Semeval-2022 task 1: CODWOE – comparing dictionaries and word embeddings. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1–14, Seattle, United States. Association for Computational Linguistics.

Ke Ni and William Yang Wang. 2017. 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)*, pages 413–417, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Thanapon Noraset, Chen Liang, Larry Birnbaum, and Doug Downey. 2017. Definition modeling: Learning to define word embeddings in natural language. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 3259–3266. AAAI Press.

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*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

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

Antonio San Martín. 2021. A Flexible Approach to Terminological Definitions: Representing Thematic Variation. *International Journal of Lexicography*, 35(1):53–74.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. Bertimbau: pretrained bert models for brazilian portuguese. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9*, pages 403–417. Springer.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 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, Online. 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.

Liner Yang, Cunliang Kong, Yun Chen, Yang Liu, Qinan Fan, and Erhong Yang. 2020. Incorporating sememes into chinese definition modeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1669–1677.

Haitong Zhang, Yongping Du, Jiaxin Sun, and Qingxiao Li. 2020. Improving interpretability of word embeddings by generating definition and usage. *Expert Systems with Applications*, 160:113633.

Hengyuan Zhang, Dawei Li, Yanran Li, Chenming Shang, Chufan Shi, and Yong Jiang. 2023. Assisting language learners: Automated trans-lingual definition generation via contrastive prompt learning. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 260–274, Toronto, Canada. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

8. Language Resource References

Ting-Yun Chang and Yun-Nung Chen. 2019. What does this word mean? explaining contextualized embeddings with natural language definition. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6064–6070.

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

Han Huang, Tomoyuki Kajiwara, and Yuki Arase. 2022. JADE: Corpus for Japanese definition modelling. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6884–6888, Marseille, France. European Language Resources Association.

Shonosuke Ishiwatari, Hiroaki Hayashi, Naoki Yoshinaga, Graham Neubig, Shoetsu Sato, Masashi Toyoda, and Masaru Kitsuregawa. 2019. 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)*, pages 3467–3476.

Arman Kabiri and Paul Cook. 2020. Evaluating a multi-sense definition generation model for multiple languages. In *International Conference on Text, Speech, and Dialogue*, pages 153–161. Springer.

Yinqiao Li, Chi Hu, Yuhao Zhang, Nuo Xu, Yufan Jiang, Tong Xiao, Jingbo Zhu, Tongran Liu, and Changliang Li. 2020. Learning architectures from an extended search space for language modeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6629–6639, Online. Association for Computational Linguistics.

Timothee Mickus, Denis Paperno, and Matthieu Constant. 2019. Mark my word: A sequence-to-sequence approach to definition modeling. In *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*, pages 1–11, Turku, Finland. Linköping University Electronic Press.

Timothee Mickus, Kees Van Deemter, Mathieu Constant, and Denis Paperno. 2022. Semeval-2022 task 1: CODWOE – comparing dictionaries and word embeddings. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1–14, Seattle, United States. Association for Computational Linguistics.

Thanapon Noraset, Chen Liang, Larry Birnbaum, and Doug Downey. 2017. Definition modeling: Learning to define word embeddings in natural language. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 3259–3266. AAAI Press.

Liner Yang, Cunliang Kong, Yun Chen, Yang Liu, Qinan Fan, and Erhong Yang. 2020. Incorporating sememes into chinese definition modeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1669–1677.

Haitong Zhang, Yongping Du, Jiaxin Sun, and Qingxiao Li. 2020. Improving interpretability of word embeddings by generating definition and usage. *Expert Systems with Applications*, 160:113633.