LDK 2023

GPT3 as a Lexical Knowledge Base for Portuguese?

Hugo Gonçalo Oliveira
University of Coimbra DEI,
CISUC Coimbra, Portugal
hroliv@dei.uc.pt

Ricardo Rodrigues
Polytechnic Institute of Coimbra, ESEC
CISUC
Coimbra, Portugal
rmanuel@dei.uc.pt

Abstract

We test the GPT3 language model in zero- and few-shot acquisition of lexico-semantic knowledge in Portuguese, with simple instruction prompts, and compare it with a BERT-based approach. Results are assessed in two test sets: TALES and the Portuguese translation of BATS. GPT3 outperforms BERT in all relations, with the few-shot approach being the best overall and for the majority of relations. Scores in both datasets further suggest that, despite their different creation approaches, they are equally suitable for this kind of evaluation.

1 Introduction

Large Language Models (LLMs) (Petroni et al., 2019) have been exploited in the acquisition of semantic relations, and as potential knowledge bases. When considering lexico-semantic relations, such models could be seen as alternatives to wordnets (Fellbaum, 1998).

BERT (Devlin et al., 2019), a bidirectional LLM pretrained on the masked language modelling task, is the most explored model in previous works, with fewer having explored GPT models (Radford et al., 2019; Brown et al., 2020). However, GPT3 is known for its adaptation to many tasks, often without requiring additional training, in zero- or few-shot approaches.

We take steps on the exploration of GPT3 for acquiring lexico-semantic knowledge in Portuguese, which contributes to better understanding this black-box model and to conclusions on its potential as a lexical knowledge base. Lexico-semantic relations are obtained through instruction-like prompts, in both zero- and few-shot learning scenarios. Performance is compared with previously used methods based on lexical patterns and masked language modelling with BERT (Gonçalo Oliveira, 2023). Experiments are performed in two analogy test sets, TALES (Gonçalo Oliveira et al., 2020) and a recent translation of the Bigger Analogy Test Set (BATS) (Gladkova et al., 2016) to Portuguese (hereafter, BATS-PT). Reported experiments are the first using the latter dataset, so we also look at differences between BATS-PT, resulting from manual translation, and TALES, created automatically from lexical resources in Portuguese.

Despite the simple and direct prompts used in GPT3, the BERT-based approach was outperformed overall and for every relation, with the best performance achieved by the few-shot approach. Moreover, scores in BATS-PT and TALES were not much different, which suggests that, despite their different creation approaches, they are equally suitable for this kind of evaluation.

The remainder of the paper is structured as follows: Section 2 overviews work on relation acquisition and analogy solving; Section 3 describes the experimentation setup; Section 4 reports and discusses the results; Section 5 concludes it.

2 Related Work

Semantic relations have been obtained from pretrained word embeddings, with simple analogy solving methods, such as: the vector offset with a single example (Mikolov et al., 2013); the average offset or a classifier of related words learned from a set of examples (Gladkova et al., 2016). These were assessed in the then proposed BATS, a test set that covers several relations types, including lexico-semantic relations.

More recently, semantic relations were obtained from Transformer-based LLMs, by prompting models with handcrafted (Petroni et al., 2019; Ushio et al., 2021) or induced lexical patterns (Bouraoui et al., 2020), in some cases (Bouraoui et al., 2020; Ushio et al., 2021) also assessed in BATS.

Pretrained models are generally used, as knowledge tends to be forgotten during the fine-tuning process (Wallat et al., 2020). Much work exploits BERT, by taking advantage of masked language modelling for acquiring relations with clozestyle prompts (e.g., `Paris is the capital of [MASK]`). GPT, another popular model, has not been so explored, also due to access limitations. Yet, there are examples using models of this family: an approach based in GPT2 (Radford et al., 2019) outperformed BERT and other LLMs in BATS (Ushio et al., 2021); a method (Liu et al., 2021) was proposed for searching for the best prompts when acquiring semantic relations with GPT2; and, among many tasks, GPT3 (Brown et al., 2020) was originally tested on a dataset of 374 analogies in English, in zero- and few-shot scenarios.

Lexico-semantic relations are especially challenging to acquire and to assess because, in opposition to morphological and to several encyclopedic relations (e.g., `capitalOf`, `hasCurrency`), they are not functions (e.g., a concept often has many hyponyms or parts). For Portuguese, related work has focused on these relations: word embeddings were exploited for enriching OpenWordNet-PT (Gonçalo Oliveira et al., 2021); BERT was used for the detection of hyponymy pairs (Paes, 2021), and for completing a range of lexico-semantic relations (Gonçalo Oliveira, 2023). The latter was assessed in TALES, similar to BATS, but for Portuguese. Previous work for Portuguese (Gonçalo Oliveira, 2022) has also suggested that GPT2 was not a good option for validating instances of lexico-semantic relations, and BERT would be better suited.

3 Experimentation Setup

This section describes the datasets and models used in this work, the approach for testing GPT3, and the adopted evaluation metrics.

3.1 Datasets

BATS comprises 40 files, each targeting a different linguistic relation. Each file has 50 entries, with two columns: a source word and a list of target words, related to the former by the relation specified in the filename.

Relations are organised in four groups: grammatical inflections, word-formation, lexicographic and encyclopedic relations. BATS was originally created for English, but the files of the ten lexicographic relations have recently been translated into several languages, in the scope of a use case in the NexusLinguarum COST Action[1]. These files comprise: *hypernyms (animals, miscellaneous)*; *hyponyms*, *meronyms (whole-substance, member-group, whole-part)*; *synonyms (intensity, exact)*; *antonyms (gradable, binary)*. We use the Portuguese translation of BATS. Table 1 illustrates this dataset with one line for each covered relation, its original BATS identifier, and an example entry.

TALES is a similar test set, but created automatically, based on the most frequent relations and their instances in ten Portuguese lexical resources. It adopts the same format as BATS, but covering 14 lexico-semantic relations, which are not exactly the same: *has-hypernym* and *hypernym-of*, each between abstract nouns, concrete nouns, and verbs; *part-of*, *has-part*; *purpose-of*, *has-purpose*; *synonym* (nouns, verbs, and adjectives); *antonym* (adjectives).

Both BATS and TALES can be used for assessing language models in the acquisition of lexico-semantic knowledge, based on predicting the target words for a given source.

3.2 Models

Two transformer models were used for acquiring lexico-semantic relations in Portuguese. GPT3 is an auto-regressive LLM with 175B parameters, 96 attention layers and a 3.2M batch size. We have used the *text-davinci-003* engine, available through the OpenAI API[2]. GPT3 is known to be multilingual, and may thus be prompted in Portuguese for generating text in this language. Temperature was set to 0.1, to force the model to produce the most probable sequences, and to avoid a non-deterministic behaviour. The results of GPT3 are compared to those by BERTimbau-large (Souza et al., 2020), a BERT model pretrained for Brazilian Portuguese, with 24 layers and 335M parameters, which can be seen as a baseline.

3.3 Approach

GPT3 was used in two scenarios in which it is known to perform well: zero-shot, where the model was prompted with an instruction that included the source word; and few-shot, where a similar prompt was concatenated to the same instruction instantiated for five examples of the same type, each followed by the respective list of target words. We used simple generic instructions asking for ten related words and changed the relation name accordingly (see Table 2). Since GPT3 is very flexible with its prompts, we did not put much effort on

[1] https://nexuslinguarum.eu/

[2] https://openai.com/api

ID	Relation	Entries	
L01	Hypernyms (animals)	anaconda *(anaconda*	cobra/réptil/boa/serpente/ofídio *snake/reptile/boa/serpent/ophidian)*
L02	Hypernyms (misc)	banheira *(tub*	contentor/artefacto/unidade/objeto/... *container/artefact/unit/object)*
L03	Hyponyms	igreja *(church*	capela/abadia/basílica/catedral *chapel/abbey/basilica/cathedral)*
L04	Meronyms (whole-substance)	atmosfera *(atmosphere*	gás/oxigénio/hidrogénio/nitrogénio/... *gas/oxygen/hydrogen/nitrogen)*
L05	Meronyms (member-group)	pássaro *(bird*	bando *flock)*
L06	Meronyms (whole-part)	academia *(academia*	faculdade/universidade/instituto *college/university/institute)*
L07	Synonyms (intensity)	choro *(cry*	grito/chio/guincho/berro/pranto *scream/shriek/screech)*
L08	Synonyms (exact)	fazenda *(cloth*	tecido/têxtil/pano *fabric/material/textile)*
L09	Antonyms (gradable)	capaz *(able*	cobra/réptil/boa/serpente/ofídio *unable/incompetent/unequal)*
L10	Antonyms (binary)	anterior *(anterior*	posterior *posterior)*

Table 1: Example entries in the Portuguese BATS files and their English translation (original).

their tuning, and leave this for future work. Still, we empirically discovered that prompts should specifically ask for Portuguese words, otherwise we would risk that, for some entries, GPT3 generates words in other languages, often Spanish. Moreover, including the number of required answers, in this case, 10, conditions the model to generate a numbered list of this size, in any case, easy to parse. Since the number of target words in the dataset is variable and it would be incoherent to give examples asking for ten but followed by a different number, we drop the `10` from the instructions in the few-shot approach.

The BERT approach followed Gonçalo Oliveira (2023) closely. BERT was prompted with a set of masked lexical patterns indicative of the target relations — e.g., `a [MASK] é um tipo de <s>` (in English, *`[MASK] is a type of <s>`*) for hyponyms. For TALES, we relied on the same patterns[3], also used for relations in BATS-PT. We only had to make a few additions to the *part-of* patterns, to better cover the *whole-substance* and *member-group* sub-types.

Differently from previous work, instead of looking at individual performances for each pattern, we add a "training" step where the best patterns are selected for each relation. The final top-10 predictions result from ranking the top-10 predictions of each of the top-5 patterns, considering their overall scores, given by the model — if there were patterns *ex aequo*, more than 5 patterns could be selected, which happened in some cases.

In order to select the best patterns, datasets were split into training and test. This had been done in BATS, for instance, by Bouraoui et al. (2020), who opted for 90%–10%, and by Rezaee and Camacho-Collados (2022), 50%–50%. The latter was our option: one half of the entries was assigned to the train portion, and the other to the test.[4] A 90%–10% split was not seen as an option because testing in only five examples (10%×50) of each relation would be too narrow for any conclusions.

Splitting the dataset was not necessary for GPT3 but, for comparison over the same data, we also run GPT3 in the test portion only. Moreover, in the few-shot scenario, the five given examples were randomly selected from the training portion, which introduced some variability in the prompts.

3.4 Metrics

Accuracy (Acc) is a common metric for assessing analogy solving in datasets like BATS. It computes the proportion of source words for which the first prediction is one of the targets. Since this is too restrictive for most lexico-semantic relations, we also compute the more relaxed Accuracy@10 (Acc@10) — i.e., the proportion of source words for which one of the targets is among the top-10 predictions; and the Mean Average Precision@10 (MAP@10), which, considering that there may be

[3] BERTimbau patterns for TALES are available from https://github.com/NLP-CISUC/PT-LexicalSemantics/blob/master/Patterns/BERT_patterns_for_TALES_v2.txt

[4] For reproducibility, we make the TALES splits available at https://github.com/NLP-CISUC/PT-LexicalSemantics/tree/master/TALESv1.1_splits.

ID	Prompt
L01 / L02	lista 10 hiperónimos, em português, da palavra <s>:
L03	lista 10 hipónimos, em português, da palavra <s>:
L04	lista 10 substâncias, em português, da palavra <s>:
L05	lista 10 conjuntos ou grupos, em português, da palavra <s>:
L06	lista 10 partes, em português, da palavra <s>:
L07 / L08	lista 10 sinónimos, em português, da palavra <s>:
L09 / L10	lista 10 antónimos, em português, da palavra <s>:

Table 2: Prompts used for relation acquisition from GPT3. Each translates to *list 10 <r>, in Portuguese, of the word <s>:*, where *r* is a name typically given to the related words, and *s* is the source word.

more than one correct answer in the top-10, accounts for the number of predicted target words and their ranking.

4 Results and Discussion

Tables 3 and 4 report on the scores of the three tested approaches, respectively in BATS-PT and in TALES. Scores are presented for each relation and as an average of all. In addition to zero- and few-shot with GPT3, we tested three variations of the BERT approach, with the best patterns optimised for each metric. However, since differences were minimal, we present only the scores of the patterns optimised for accuracy.

The few-shot approach is clearly the best in both datasets. In BATS-PT, it achieves the best performance in every relation in each of the three metrics, except for meronyms (member-group), with the best scores in two metrics, and for synonyms (intensity) and antonyms (gradable), with the best score in only one (in *ex aequo*). In TALES, the results are quite similar. Only in a handful of cases few-shot is outperformed by zero-shot (or has the same score), and fewer yet by BERT. Surprisingly, despite no training nor prompt tuning, zero-shot GPT3 is better than BERT for almost every relation and metric.

Performance is variable across relation types. In BATS-PT, *hypernyms (animals)* is one of the best relations for all approaches, whereas zero- and few-shot perform equally well for *antonyms (gradable)*. Lowest performances by few-shot are for *meronyms (member-group)* and *synonyms (intensity)*, the same as for the zero-shot. Specifically in the *member-group* relation, we observe some confusion with hypernymy and co-hyponymy (e.g., parlamentar [*parliamentarian*] and legislador [*legislator*] for senador [*senator*]) and, for zero-shot, answers that are groups of other things (e.g., rebanho [*herd*] or matilha [*pack*], for pássaro [*bird*]). In few-shot, however, shorter lists are generated, often with less or no incorrect answers.

In TALES, all approaches perform especially well for *antonyms*, and zero-shot achieves top-performance in *synonyms (verbs)*. The other synonymy relations are among the top-performing in few-shot, whereas the best performance of BERT is for *has-hypernym (abstract)*.

We highlight that the average scores in BATS-PT are not substantially different from those in TALES. Overall, few-shot performs slightly better in BATS-PT, and zero-shot in TALES. BERT is very similar in both test sets. Moreover, there is a similar trend for equivalent relations: models generally perform better for *antonymy* and *hypernymy*, and worse for *meronymy*. BATS-PT was not originally created for Portuguese, but it is the result of thorough manual translation, whereas TALES was created specifically for Portuguese, but automatically. To some extent, this validates the approach adopted for creating TALES. But it does not mean that any of the datasets cannot be improved. In fact, low scores in TALES' *has-part* and *part-of* relations can be partially explained by limitations of the dataset. TALES is based on redundancy across lexical resources and the following reasons may result in less consensual and incomplete entries: (*i*) to reach the 50 entries, *has-part* and *part-of* are the relations for which required redundancy was the lowest (Gonçalo Oliveira et al., 2020); (*ii*) there are several sub-types of meronymy, defined differently across resources.

Reference scores for TALES (Gonçalo Oliveira, 2023) use the same BERT-based approach, but in the full dataset, without combining patterns. Though not comparable, differences suggest that the combination of patterns is not always beneficial. Yet, the best patterns have to be selected from part of the data. Moreover, we should add that, with only 50 entries, the train-test split has a noticeable impact on the selection of patterns and on the

Relation	**BERT**			**GPT3** (zero-shot)			**GPT3** (five-shot)		
	Acc	Acc@10	MAP@10	Acc	Acc@10	MAP@10	Acc	Acc@10	MAP@10
L01	0.76	0.92	0.60	0.84	0.96	0.79	**1.00**	**1.00**	**0.93**
L02	**0.42**	0.88	0.35	0.29	0.71	0.43	**0.42**	**0.96**	**0.62**
L03	0.21	0.50	0.24	0.46	0.58	0.44	**0.50**	**0.67**	**0.50**
L04	0.24	0.60	0.31	0.20	0.36	0.24	**0.52**	**0.68**	**0.53**
L05	0.08	**0.44**	0.19	0.04	0.08	0.06	**0.20**	0.28	**0.24**
L06	0.00	0.20	0.04	0.20	0.40	0.22	**0.32**	**0.64**	**0.38**
L07	0.08	0.12	0.09	**0.36**	**0.72**	**0.44**	0.20	**0.72**	0.43
L08	0.12	0.32	0.14	0.48	0.68	0.50	**0.60**	**0.92**	**0.73**
L09	0.32	0.48	0.35	**0.76**	**0.88**	**0.71**	0.72	0.80	**0.71**
L10	0.48	0.78	0.48	0.57	0.61	0.57	**0.74**	**0.83**	**0.73**
Average	0.27	0.52	0.28	0.42	0.60	0.44	**0.52**	**0.75**	**0.58**

Table 3: Performance in BATS-PT.

Relation	**BERT**			**GPT3** (zero-shot)			**GPT3** (five-shot)		
	Acc	Acc@10	MAP@10	Acc	Acc@10	MAP@10	Acc	Acc@10	MAP@10
Antonyms (adjectives)	0.48	0.52	0.42	0.76	0.88	0.80	**0.96**	**0.96**	**0.94**
Purpose-of	0.16	0.20	0.18	0.20	0.32	0.23	**0.40**	**0.40**	**0.35**
Has-Purpose	0.16	0.36	0.23	0.32	**0.64**	0.45	**0.60**	0.60	**0.60**
Has-Hypernym (abstract)	0.44	**0.80**	0.35	0.44	**0.80**	**0.50**	**0.80**	**0.80**	0.45
Has-Hypernym (concrete)	0.24	0.64	0.24	0.40	0.68	0.43	**0.80**	**0.80**	**0.49**
Has-Hypernym (verbs)	0.08	0.48	0.20	0.60	0.88	**0.62**	**0.92**	**0.92**	0.57
Hypernym-Of (abstract)	0.20	**0.68**	0.29	0.36	**0.68**	0.39	**0.68**	**0.68**	**0.40**
Hypernym-of (concrete)	0.48	0.88	0.50	0.48	0.72	0.52	**0.96**	**0.96**	**0.74**
Hypernym-Of (verbs)	0.04	0.48	0.17	0.52	**0.84**	**0.56**	**0.76**	0.76	0.47
Has-Part	0.16	**0.36**	0.22	0.08	0.12	0.09	**0.36**	**0.36**	**0.26**
Part-Of	0.08	0.40	0.14	0.16	0.20	0.17	**0.48**	**0.48**	**0.34**
Synonyms (nouns)	0.24	0.72	0.33	0.60	0.92	0.65	**1.00**	**1.00**	**0.74**
Synonyms (verbs)	0.32	0.76	0.33	**0.56**	**0.96**	**0.56**	**0.56**	**0.96**	0.54
Synonyms (adjectives)	0.20	0.76	0.28	0.48	0.72	0.47	**0.96**	**0.96**	**0.67**
Average	0.23	0.57	0.28	0.43	0.67	0.46	**0.73**	**0.76**	**0.54**

Table 4: Performance in TALES.

performance achieved for some relations.

There are no reference scores for BATS-PT, but there are for the English version, where the accuracy reported by Ushio et al. (2021) is 81% with GPT2, substantially higher than few-shot's 56% in BATS-PT. Despite GPT3 being more powerful, the lower performance is a consequence of a simpler approach, and suggests that there is room for improvement, for instance, if we invest in prompt tuning. Yet, languages are different, and BATS-PT may have resulted in a more challenging dataset, for a less-resourced language.

5 Conclusions

We have seen that, to some extent, GPT3 can be used as a lexical knowledge base for Portuguese. When compared to handcrafted knowledge bases, the coverage of GPT3 is difficult to meet. Moreover, performance is variable across relations, but this also happens for automatically created knowledge bases. GPT3 clearly outperformed a BERT-based approach, which had shown improvements against approaches based on static word embeddings (Gonçalo Oliveira, 2023). The best performance is achieved with a few-shot approach with simple direct prompts, without previous tuning, which suggests that there is still room for improvement.

This was also the first time that BATS-PT was used as a benchmark. The fact that the scores achieved were comparable to those in TALES, despite its automatic creation, contributes to validating the utility of both datasets.

Future directions would be to test alternative prompts and to experiment with more recent LLMs, such as the recently release GPT4 (OpenAI, 2023). However, we should not forget that GPT is a black-box architecture, which prevents a deeper analysis and a direct fix of its errors. This adds to the fact that we know that GPT3 and GPT4 were trained in much data, but not exactly on which data, which may raise relevant questions for evaluation — e.g., did it learn from the test examples? While it cannot have learned from BATS-PT, because the dataset has not been released yet, we may question whether it learned from the original dataset, which, through deep inference, may help with other languages.

Acknowledgements: We would like to thank the remaining translators of BATS to Portuguese, Purificação Silvano and Sara Carvalho. This work was based upon work in COST Action CA18209 Nexus Linguarum, supported by COST (European Cooperation in Science and Technology). http://www.cost.eu/; and partially supported by the Portuguese Recovery and Resilience Plan (PRR) through project C645008882-00000055, Center for Responsible AI.

References

Zied Bouraoui, Jose Camacho-Collados, and Steven Schockaert. 2020. Inducing relational knowledge from BERT. In *Proc of AAAI Conference on Artificial Intelligence*, pages 7456–7463. AAAI Press.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

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

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Procs of NAACL 2016 Student Research Workshop*, pages 8–15. ACL.

Hugo Gonçalo Oliveira. 2023. On the acquisition of WordNet relations in Portuguese from pretrained masked language models. In *Procs of 12th Global WordNet Conference*, GWC. Global WordNet Association.

Hugo Gonçalo Oliveira, Tiago Sousa, and Ana Alves. 2020. TALES: Test set of Portuguese lexical-semantic relations for assessing word embeddings. In *Procs of ECAI 2020 Workshop on Hybrid Intelligence for Natural Language Processing Tasks (HI4NLP 2020)*, volume 2693 of *CEUR Workshop Proceedings*, pages 41–47. CEUR-WS.org.

Hugo Gonçalo Oliveira. 2022. Exploring transformers for ranking Portuguese semantic relations. In *Procs of the 13th Language Resources and Evaluation Conference*, LREC 2022, pages 2573–2582, Marseille, France. ELRA.

Hugo Gonçalo Oliveira, Fredson Silva de Souza Aguiar, and Alexandre Rademaker. 2021. On the Utility of Word Embeddings for Enriching OpenWordNet-PT. In *Procs of 3rd Conference on Language, Data and Knowledge (LDK 2021)*, volume 93 of *OASIcs*, pages 21:1–21:13, Dagstuhl, Germany. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021. GPT understands, too. *arXiv preprint arXiv:2103.10385*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Procs of Workshop track of ICLR*.

OpenAI. 2023. GPT-4 Technical Report. ArXiv:2303.08774 [cs].

Gabriel Escobar Paes. 2021. Detecção de hiperônimos com bert e padrões de hearst. Master's thesis, Universidade Federal de Mato Grosso do Sul.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proc 2019 Conf on Empirical Methods in Natural Language Processing and 9th Intl Joint Conf on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473. ACL.

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

Kiamehr Rezaee and Jose Camacho-Collados. 2022. Probing relational knowledge in language models via word analogies. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3930–3936.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: Pretrained BERT models for Brazilian Portuguese. In *Procs of Brazilian Conf on Intelligent Systems (BRACIS 2020)*, volume 12319 of *LNCS*, pages 403–417. Springer.

Asahi Ushio, Luis Espinosa Anke, Steven Schockaert, and Jose Camacho-Collados. 2021. BERT is to NLP what AlexNet is to CV: Can pre-trained language models identify analogies? In *Procs 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 3609–3624, Online. Association for Computational Linguistics.

Jonas Wallat, Jaspreet Singh, and Avishek Anand. 2020. BERTnesia: Investigating the capture and forgetting of knowledge in BERT. In *Procs of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 174–183, Online. Association for Computational Linguistics.