On the Acquisition of WordNet Relations in Portuguese from Pretrained Masked Language Models

Hugo Gonçalo Oliveira
CISUC, Department of Informatics Engineering
University of Coimbra, Portugal,
hroliv@dei.uc.pt

Abstract

This paper studies the application of pretrained BERT in the acquisition of synonyms, antonyms, hypernyms and hyponyms in Portuguese. Masked patterns indicating those relations were compiled with the help of a service for validating semantic relations, and then used for prompting three pretrained BERT models, one multilingual and two for Portuguese (base and large). Predictions for the masks were evaluated in two different test sets. Results achieved by the monolingual models are interesting enough for considering these models as a source for enriching wordnets, especially when predicting hypernyms of nouns. Previously reported performances on prediction were improved with new patterns and with the large model. When it comes to selecting the related word from a set of four options, performance is even better, but not enough for outperforming the selection of the most similar word, as computed with static word embeddings.

1 Introduction

As it happens for many other tasks in the domain of Natural Language Processing (NLP), transformer-based language models have been explored in the acquisition of semantic relations, towards their application in the creation or enrichment of knowledge bases, or on their direct usage as knowledge bases (AlKhamissi et al., 2022). More precisely, having in mind that a typical application of language models is text completion, transformer-based models have been used for completing lexical patterns, in what can be seen as a shortcut to earlier research on the acquisition of relations from textual corpora (e.g., Hearst (1992)). If the focus are lexico-semantic relations, such an approach can be useful for enriching wordnets (Fellbaum, 1998).

In this study, we build on previous efforts, specifically those targeting the Portuguese language (Gonçalo Oliveira, 2022), and evaluate the acquisition of synonymy, antonymy, and hypernymy-hyponymy from BERT models, namely the base and large versions of BERT pretrained exclusively for Portuguese (Souza et al., 2020), and the multilingual BERT. Evaluation is made on two test sets, both covering different variations of the target relations, and starting with source words, but with different goals: in B^{2}SG (Wilkens et al., 2016), a related word has to be selected from four options; in TALES (Gonçalo Oliveira et al., 2020), one related word has to be predicted. Since the approach is not just dependent on the models, several patterns were handcrafted for each target relation, building on previous work, but also on the adaptation of patterns used in the scope of VARRA (Freitas et al., 2015), a service for searching for and validating instances of lexico-semantic relations by resorting to Portuguese corpora.

After fixing the first argument of each instance as the source word, patterns were used to prompt the BERT models, results were evaluated in the test sets, and conclusions were drawn. Performance with the multilingual model was poor, and the large model is generally the best option. When selecting the correct candidate in B^{2}SG, results are positive, but end up being outperformed by simply selecting the option that maximises similarity, computed in a model fine-tuned for computing semantic similarity or in static word embeddings. Predicting the related words is more challenging. Nevertheless, top performances are achieved when predicting hypernyms and results can still be useful for suggesting new relation instances to wordnets. Moreover, using the large version of the model and including the VARRA patterns contributed to improvements in previously reported performance in TALES.

In the remainder of the paper, Section 2 overviews related work on the automatic acquisition of semantic relations from text and language models; Section 3 describes the adopted approach in more detail, focusing on the patterns, the test sets and the models; Section 4 reports on the best patterns for each relation and test set, together with their performance; Section 5 summarises the main conclusions and future directions of this work.

2 Related Work

The enrichment of wordnets with relations extracted automatically from corpora has a long tradition, following the work of Hearst (1992), where a set of lexico-syntactic patterns denoting hyponymy was presented and applied to the acquisition of relation instances. To minimise human intervention, hyponymy patterns were learned automatically with distant supervision (Snow et al., 2005), and patterns for other relations were learned and ranked with weak (Pantel and Pennacchiotti, 2006), in both cases using seed examples from Princeton WordNet (Fellbaum, 1998). On relation extraction from Portuguese text (de Abreu et al., 2013), only a minority is focused on lexico-semantic relations. These include rule-based approaches for acquiring hyponymy (de Freitas and Quental, 2007) and part-of (Markov et al., 2014) relations from corpora; as well as other relations from dictionary definitions (Gonçalo Oliveira et al., 2008).

A more recent alternative is to acquire relations from distributional models, such as word embeddings. Even if relations are not explicit, analogies (Mikolov et al., 2013) have been computed for a broad range of syntactic and semantic relations. Besides the unsupervised discovery of hypernymy instances (Chang et al., 2018), the performance of simple analogy was improved by learning to compute related words from multiple examples (Drozd et al., 2016), more specifically, from the BATS test set, which covers synonymy, antonymy and hypernymy, among other syntactic and encyclopaedic relations. The previous were also applied to Portuguese word embeddings, when used to solve lexico-semantic analogies in TALES (Gonçalo Oliveira et al., 2020), a test of with the same format as BATS (Drozd et al., 2016). Despite the low accuracy, among the predictions there are useful suggestions that may be manually added to wordnets, as it happened with OpenWordNet-PT (Gonçalo Oliveira et al., 2021).

But the current paradigm in NLP are transformer-based models, like BERT (Devlin et al., 2019) or GPT (Radford et al., 2019), and there has also been work on using them as knowledge bases (AlKhamissi et al., 2022). Even if they are not ready for explicitly retrieving semantic relations, using the right prompts can result in the acquisition of related words, in what can be seen as a shortcut for earlier corpora-based approaches, i.e., these models are pre-trained in large collections of text and are good at filling blanks (Petroni et al., 2019; Ettinger, 2020), completing sentences (Radford et al., 2019), or computing their likelihood (Goldberg, 2019; Paes, 2021).

Among other efforts, pretrained BERT has been assessed for the presence of relational knowledge using discrete prompts (Petroni et al., 2019); for relation induction (Bouraoui et al., 2020), starting with a small number of patterns and seeds; or for classifying semantic relations based on attention weights (Chizhikova et al., 2022). Some researchers conclude that the prompting approach suits better some relations (e.g., hypernymy) than others (Ettinger, 2020), while others have shown that BERT is not very good at predicting hyponymy relations inherited through transitivity (Lin and Ng, 2022). For Portuguese, recent work exploited BERT for detecting hyponymy pairs (Paes, 2021), ranking automatically extracted relation instances (Gonçalo Oliveira, 2022), or acquiring new instances (Gonçalo Oliveira, 2022).

3 Approach

Gonçalo Oliveira (2022) proposed the acquisition of lexico-semantic relations from BERTimbau (Souza et al., 2020), a BERT model pre-trained for Portuguese, using prompts that indicated the target relations. Since BERT is pretrained on masked language modelling in a large corpus, the pretrained version should be enough for acquiring lexico-semantic relations. Some considerations were made on setting the prompts and results were evaluated in the TALES (Gonçalo Oliveira et al., 2020) test of lexico-semantic analogies. However, results were limited to using BERTimbau-base and to an initial set of handcrafted patterns. Here, we augment the previous work by considering a second dataset, B^{2}SG, other BERT models, and additional patterns adapted from VARRA (Freitas et al., 2015), which lead to improvements on performance. Moreover, we discuss synonymy in more detail.

3.1 Prompts

Our approach consists of acquiring triples $\langle x_1, r, x_2 \rangle$, where r is a relation predicate and x_1 and x_2 are the relation arguments. This is performed by prompting masked language models (MLMs) with cloze-style patterns indicating the target relation (r), where one of the arguments (x_1) is fixed and the other (x_2) is masked. For instance, the lexical pattern "a x_2 is a type of x_1" typically indicates hypernym(x_1, x_2). Thus, to acquire hypernyms of *dog*, x_1 and x_2 are respectively replaced by the word `dog` and by the `[MASK]` token, resulting in the prompt "`a dog is a type of [MASK]`". Expected predictions for the `[MASK]` would be *animal* or *mammal*.

Useful patterns for acquiring the relations of interest were compiled and made available by Gonçalo Oliveira (2022). However, they did not cover several patterns handcrafted for VARRA (Freitas et al., 2015), a service for searching for and validating instances of semantic relations in Portuguese, through the corpora of the AC/DC project (Santos and Bick, 2000). So, we decided to review the original list and include adaptations of the VARRA patterns. Table 1 illustrates this adaptation with some patterns and the resulting masked prompts. Since VARRA patterns include regular expressions, with some optional and alternative tokens, some adaptations resulted in more than one masked pattern.

3.2 Test Sets

Two different datasets were used for assessing to what extent BERT could predict correctly-related words for the masks. B^{2}SG (Wilkens et al., 2016) is similar to the WordNet-Based Synonymy Test (Freitag et al., 2005), but based on the Portuguese part of BabelNet (Navigli and Ponzetto, 2012) and partially evaluated by humans[1]. It contains frequent Portuguese nouns and verbs (source words) followed by four candidates, out of which only one is related, and is organised in six relations: synonymy (1,171 entries for nouns, 435 for verbs), antonymy (145 nouns, 167 verbs), and hypernymy (758 nouns, 198 verbs), all of them used in this study. The following are examples for noun-synonymy and verb-hypernymy:

- `cataclismo desastre_noun talha_noun obesidade_noun alusão_noun`
  (cataclysm disaster carving obesity allusion)

- `danificar lesar_verb rastrear_verb divertir_verb embaraçar_verb`
  (damage harm track amuse embarrass)

[1] B^{2}SG is available from http://www.inf.ufrgs.br/pln/resource/B2SG.zip

When using source words as the fixed argument, B^{2}SG can be used for assessing whether BERT ranks the related candidate as the best fit for the mask.

TALES (Gonçalo Oliveira et al., 2020) is a test of lexico-semantic analogies, created from the contents of ten Portuguese lexical resources[2]. It covers 14 relation types, but we focus on ten: synonymy (nouns, verbs, and adjectives); antonymy (adjectives); hypernymy and hyponymy (each between abstract nouns, concrete nouns, and verbs). TALES format is similar to BATS (Drozd et al., 2016). For each relation, it includes 50 entries with two columns: a source word and a list of related words (target). The following are examples for antonymy and concrete-hyponymy:

- `novo velho/idoso/entradote`
  (young old/aged/oldish);

- `edifício construção/estrutura/artefato`
  (building construction/structured/artefact)

When using source words as the fixed argument, TALES can be used for assessing whether the predictions for the mask correspond to target words.

Since the adopted naming of the files can be confusing, we note that in the hypernymy files of B^{2}SG, the source word is a hyponym of the correct option, whereas in the hypernymy files of TALES, the source word is a hypernym of the target words.

3.3 Masked Language Models

Three BERT models were used in this study, namely, two versions of BERTimbau (Souza et al., 2020), for Portuguese, and the multilingual version of BERT. All of them are available from the HuggingFace hub and were used with the `transformers`[3] Python library. Specifically, for answering TALES, the `fill-mask` pipeline of this library was used. For B^{2}SG, we resorted to the FitBERT[4] tool, also based on the `transformers` library.

BERTimbau was pretrained in a large corpus of Brazilian Portuguese and has two versions: BERTimbau-base[5], hereafter BERT-base, with 12

[2] TALES is available from https://github.com/NLP-CISUC/PT-LexicalSemantics/tree/master/TALESv1.1

[3] https://huggingface.co/transformers/

[4] https://github.com/Qordobacode/fitbert

[5] neuralmind/bert-base-portuguese-cased

Relation	VARRA	Masked
Synonym-of	`[lema="PALAVRA1"] "," "isto" "é" "," [lema="PALAVRA2"]`	X_1, isto é, [MASK]
Antonym-of	`[word="nem|seja|quer"] [lema="PALAVRA1"] [lema=","]* [word="nem|seja|quer"] [lema="PALAVRA2"]`	nem X_1, nem [MASK] seja X_1, seja [MASK] quer X_1, quer [MASK]
Hypernym-of	`[lema="PALAVRA1"] [pos="ADJ.*"]* [lema=","]* [lema="tal"]* "como" [pos="DET.*"]* [pos="ADJ.*"]* [lema="PALAVRA2"]`	X_1, tal como [MASK]
Hypernym-of	`[lema="PALAVRA2" & pos="N.*"] "e" [lema="outro"] [lema="PALAVRA1" & pos="N.*"]`	X_1 e outro [MASK]

Table 1: VARRA patterns and their adaptation to masked patterns.

layers and 110M parameters; and BERTimbau-large[6], hereafter BERT-base, with 24 layers and 335M parameters. The multilingual BERT, hereafter BERT-ML[7], was pretrained on Wikipedia for 104 languages, has 12 layers and 110M parameters.

The multilingual model XLM-RoBERTA-large[8] was also explored, but it performed around the random chance in B^{2}SG (25% accuracy), so its results are omitted.

4 Results

This section reports on the best patterns for each test and relation, and discusses the achieved evaluation scores. For each test, scores are also compared with alternative approaches.

4.1 Performance in B^{2}SG

After fixing the source words for the prompt (X_1), BERT models were assessed in the selection of the related word for each entry in B^{2}SG, out of the four options. FitBERT was used for this – given a masked sentence and a list of options, this tool ranks the options according to their suitability for the mask, based on pre-softmax logit scores, as performed by Goldberg (2019).

From the resulting ranks, we compute two metrics: accuracy, i.e., the proportion of entries for which the related word was ranked first; and the average rank of the related word, a continuous value between 1 (top) and 4 (bottom). Table 2 summarises the achieved results. For each relation, it shows the most accurate pattern for each model, followed by its accuracy (Acc) and average rank (Rank) for the three models. When the best pattern was the same for multiple models, the table includes the best patterns overall. Patterns are translated to English, and those adapted from VARRA are marked with a V. The full list of patterns is available from a GitHub repository[9].

The first conclusion is that BERT-large is the best option for every relation but verb-antonymy, where the highest rank is achieved with this model, but not the highest accuracy, which is by BERT-base. This is not surprising because BERT-large has more layers and more parameters, used for better representations that should result in better predictions, even if this is not always the case. On the other hand, performance with BERT-ML is generally above random chance (25%), but consistently lower than for the other models. This only confirms that monolingual models are a better option for this monolingual task.

Performance is better for relations between nouns than for relations between verbs. The best performance is for noun-antonymy, followed by noun-hypernymy, and the worse is for verb-synonymy and verb-antonymy. This suggests either that relations between verbs are more difficult to capture by lexical patterns, or that the best patterns for verb relations are harder to think of.

Since the entries of B^{2}SG are limited to four options, a suitable approach for answering this test would be to simply select the candidate that maximises similarity with the source word. To analyse how the adopted pattern-based approach compares to the previous approach in this test, we resorted to embeddings for selecting the candidate word that was the most similar to the source. Different BERT models and models of static word embeddings were tested, namely: (i) `CLS` token of BERT-base and of BERT-large; (ii) mean pooling of BERT-base and BERT-large tokens; (iii) BERTimbau-large fine-tuned for Semantic Textual Similarity in Portuguese[10]; (iv) 300-sized word2vec (CBOW and Skip-gram) and GloVe embeddings, pretrained for Portuguese (Hartmann et al., 2017). Table 3 puts the accuracies of the previous side-by-side with the best accuracies of the pattern-based approach.

With BERT-large, the best performance for synonymy was slightly improved, but this was not

[6] `neuralmind/bert-large-portuguese-cased`
[7] `bert-base-multilingual-cased`
[8] `xlm-roberta-large`
[9] `https://github.com/NLP-CISUC/PT-LexicalSemantics/tree/master/Patterns`
[10] `rufimelo/bert-large-portuguese-cased-sts`

Relation	PoS	Pattern		BERT-ML		BERT-base		BERT-large	
				Acc	Rank	Acc	Rank	Acc	Rank
Synonym-of	N	X_1 é o mesmo que [MASK] (X_1 is the same as [MASK])		0.35	2.22	0.57	1.71	**0.64**	**1.58**
Synonym-of	N	X_1, isto é, [MASK] (X_1, this is, [MASK])	V	0.33	2.23	0.58	1.71	0.62	1.60
Synonym-of	N	X_1 é sinónimo de X_2 (X_1 is a synonym of [MASK])		0.37	2.20	0.50	1.88	0.52	1.82
Synonym-of	V	X_1, isto é, [MASK] (X_1, this is, [MASK])	V	0.32	2.28	0.50	1.80	**0.56**	**1.67**
Synonym-of	V	X_1, ou seja, [MASK] (X_1, i.e., [MASK])	V	0.49	1.85	0.54	1.73	0.37	2.17
Synonym-of	V	querer X_1 é o mesmo que querer [MASK] (willing to X_1 is the same as willing to [MASK])		0.38	2.14	0.47	1.86	0.44	1.86
Antonym-of	N	nem [MASK], nem X_1 (not X_1, nor [MASK])	V	0.44	2.03	0.76	1.64	**0.77**	**1.36**
Antonym-of	N	X_1 é o contrário de [MASK] (X_1 is the opposite of [MASK])		0.46	1.92	0.72	1.44	0.77	1.37
Antonym-of	N	X_1 é diferente de X_2 (X_1 is different than [MASK])		0.40	2.06	0.68	1.51	0.72	1.43
Antonym-of	V	se está a X_1 não está a [MASK] (if it is X_1, it is not [MASK])		0.46	1.95	0.60	1.69	0.62	**1.61**
Antonym-of	V	nem [MASK], nem X_1 (not X_1, nor [MASK])	V	0.29	2.31	**0.63**	1.64	0.61	**1.61**
Antonym-of	V	quer X_1, quer [MASK] (whether X_1 or [MASK])	V	0.30	2.26	0.60	1.71	0.61	1.69
Hypernym-of	N	X_1, isto é, um tipo de [MASK] (X_1, this is, a type of [MASK])	V	0.44	2.02	0.68	1.50	**0.71**	**1.43**
Hypernym-of	N	X_1, isto é, uma espécie de [MASK] (X_1, this is, a kind of [MASK])	V	0.41	2.06	0.63	1.57	0.70	1.44
Hypernym-of	N	X_1 é um tipo de [MASK] (X_1 is a type of [MASK])		0.42	2.04	0.65	1.58	0.67	1.54
Hypernym-of	V	a X_1 ou outras formas de [MASK] (X_1 or other forms of [MASK])	V	0.36	2.20	0.61	1.60	**0.66**	**1.54**
Hypernym-of	V	a X_1 ou outros modos de [MASK] (X_1 or other modes of [MASK])		0.37	2.13	0.57	1.65	0.61	1.56
Hypernym-of	V	[MASK] é hiperónimo de X_1 ([MASK] is a hypernym of X_1)		0.19	2.59	0.47	1.79	0.62	1.60

Table 2: Best performing patterns in B^{2}SG and their performance.

Relation	PoS	BERT-b (patterns)	BERT-l (patterns)	BERT-b (CLS)	BERT-l (CLS)	BERT-b (tokens)	BERT-l (tokens)	BERT-STS	CBOW	Skip	GloVe
Synonym-of	N	0.58	0.64	0.60	0.67	0.59	0.66	0.80	0.71	**0.83**	0.81
Synonym-of	V	0.54	0.56	0.55	0.51	0.54	0.54	**0.75**	0.66	0.68	0.70
Antonym-of	N	0.76	0.77	0.72	0.63	0.69	0.64	0.78	0.70	0.81	**0.83**
Antonym-of	V	0.63	0.62	0.51	0.51	0.49	0.57	0.68	0.67	0.69	**0.71**
Hypernym-of	N	0.68	0.71	0.59	0.61	0.59	0.62	0.76	0.65	0.76	**0.80**
Hypernym-of	V	0.61	0.66	0.52	0.51	0.54	0.54	**0.71**	0.64	0.66	0.70

Table 3: Accuracy of similarity methods in B^{2}SG.

the case for the other relations, suggesting that synonymy is better captured by approaches for computing semantic similarity, even if trained in longer sequences, than with fixed patterns. With BERT-STS, performance was improved for all relations. Despite being fine-tuned for computing the similarity between sentences, the model showed to adapt well-enough to single words, as in B^{2}SG, also confirming the benefits of fine-tuning. But this is was still not enough for outperforming the best static word embeddings, GloVe, in all relations. In fact, BERT-STS only achieved the best performance in two relations, both between verbs (synonymy and hypernymy). This might be related to the higher number of inflections of verbs and how each model handles them, i.e., a different entry for each inflection in static word embeddings vs word piece tokenization and contextual embeddings in BERT. Nevertheless, the fact that all target relations are connected to similarity, plus the constrain of only four candidates, make GloVe embeddings the best option overall for B^{2}SG, with the top performance in half of the relations.

4.2 Performance in TALES

With TALES, we wanted to assess how well the pattern-based approach could be used for actually predicting the related words, not restricted to a set of options. For each prompt, again, we fix the source word and use the models for predicting words for the mask. Based on the predictions, two metrics are computed, namely: accuracy, i.e., the proportion of entries for which the first prediction was correct; accuracy@10, i.e., the proportion of entries for which a correct prediction was among the top-10 predictions.

Table 4 summarises the achieved results. For each relation, it shows the most accurate pattern for each BERTimbau model, followed by its accuracy (Acc) and accuracy@10 (Acc@10) for the three models. When the best pattern is the same for both, the table includes the two best patterns. Patterns are translated to English, and those adapted from VARRA are followed by a V.

As expected, when predictions are not constrained to four options, performance is much lower. BERT-large tends to perform better than BERT-base, except for hyponymy relations. i.e., when predicting hypernyms. Curiously, top performances are achieved for these relations, between abstract and concrete nouns, which is in line with previous work for English (Ettinger, 2020). A probable cause is the smaller number of hypernyms when compared to hyponyms. On the other hand, the lowest performances are in the prediction of synonym adjectives, concrete hyponyms, and verb hypernyms.

We note that some of the top performances were achieved by VARRA patterns, including for hypernymy and hyponymy. A particularly productive pattern was "`um(a)` X_1`, isto é, um tipo de [MASK]`", which achieved the best performance in abstract and concrete hyponymy. In addition to the new patterns, BERT-large also contributed to an overall improvement of the performances reported in previous work (Gonçalo Oliveira, 2022). We highlight the improvements on the relations between abstract nouns, specifically, an increase of 0.26 points in the accuracy of abstract hyponymy and of 0.14 in abstract hypernymy.

As in previous work, we compared the performances achieved by this approach with those of analogy-solving methods in static word embeddings. Table 5 puts the best accuracies with the pattern-based approach side-by-side the best accuracies with the four analogy-solving methods used by Drozd et al. (2016) – Similarity, 3CosAdd, 3CosAvg, LRCos – in the same three models of static word embeddings used in the B^{2}SG.

There are three relations for which performance is better with static word embeddings. Two of them are noun-synonymy and adjective-synonymy, which confirms the anticipated challenge of capturing synonymy with a single lexical pattern. The third relation is verb-hypernymy, for which there were no patterns in VARRA, and we could not add many more to the used list. Using BERT-large made it possible to improve the performance for concrete-hypernymy.

5 Conclusion

This paper reports on the experimentation of BERT models for Portuguese for answering relation tests, by prompting them with patterns that indicate synonymy, antonymy, hypernymy and hyponymy relations. Our first conclusion was that monolingual models perform substantially better than a multilingual model. Second, when it comes to selecting the related word from a limited set of options, the proposed approach performs ok, even if better for relations between nouns than between verbs. However, this turns out not being so useful, because it is outperformed by simply selecting the most similar word, as computed in a fine-tuned BERT or in static word embeddings. Third, this approach can be used for predicting related words, in this case, better for noun hypernyms, as in previous work for English (Ettinger, 2020). We also note the positive impact of using BERT-large and of including the patterns of a relation validation service, which enabled the improvement of previously reported results in the same dataset.

At the same time, there is still much room for improvement, and performances achieved suggest that it might be risky to create or enrich a knowledge base in a completely automatic fashion. Yet, given that the reported evaluation ends up being limited by the contents of the test sets, in the future, it could be interesting to test how far one could go by adopting this approach for the creation of a knowledge base completely from scratch. Additional conclusions could be taken from manually evaluating a sample of extracted instances. We should, nevertheless, look at BERT as an alternative source of knowledge, capable of providing suggestions for enriching knowledge bases, even if they need to be manually-validated before actual inclusion. This would be similar to what happened in the enrichment of OpenWordNet-PT (Gonçalo Oliveira et al., 2021), with suggestions computed from static word embeddings.

Finally, given that the prompts play a key role on this approach, it is always on our mind to test more and more patterns. So far, performance could be improved with the inclusion of patterns from a relation validation service, but additional patterns, potentially better, could be discovered by processing large corpora, as others did (Jiang et al., 2020;

Relation	PoS	Pattern		BERT-ML		BERT-base		BERT-large	
				Acc	Acc@10	Acc	Acc@10	Acc	Acc@10
Synonym-of	N	`X1 é sinónimo de [MASK]` (X_1 is a synonym of [MASK])		0.02	0.20	**0.28**	0.64	0.20	**0.70**
Synonym-of	N	`X1 é o mesmo que [MASK]` (X_1 is the same as [MASK])		0.04	0.08	0.20	0.58	0.20	0.66
Synonym-of	V	`X1 é o mesmo que [MASK]` (X_1 is the same as[MASK])		0.12	0.24	0.12	0.80	**0.34**	**0.90**
Synonym-of	V	`estar a X1 é o mesmo que estar a [MASK]` (to be X_1 is the same to be [MASK])		0.18	0.44	0.20	0.68	0.26	0.82
Synonym-of	ADJ	`estar X1 é o mesmo que estar [MASK].` (being X_1 is the same as being [MASK])		0.14	0.42	0.06	0.46	**0.24**	0.54
Synonym-of	ADJ	`ser X1 é o mesmo que ser [MASK].` (being X_1 is the same as being [MASK])		0.06	0.24	0.14	0.54	0.22	**0.64**
Antonym-of	ADJ	`ser [MASK] é o contrário de ser X1` (being X_1 is the opposite of being [MASK])		0.08	0.22	0.26	0.40	**0.38**	**0.48**
Antonym-of	ADJ	`nem X1, nem [MASK]` (not X_1, nor [MASK])	V	0.02	0.06	0.34	0.40	0.34	0.46
Hypernym-of	Abstract	`a [MASK] é um tipo de X1` (the [MASK] is a type of X_1)		0.08	0.24	0.22	0.60	**0.38**	0.66
Hypernym-of	Abstract	`uma [MASK], isto é, um tipo de X1` (a [MASK], this is, a type of X_1)	V	0.04	0.32	0.32	**0.70**	0.26	0.62
Hypernym-of	Concrete	`o [MASK], que é um tipo de X1` (the [MASK], which is a type of X_1)	V	0.08	0.20	0.20	0.54	**0.24**	**0.56**
Hypernym-of	Concrete	`a [MASK] é um tipo de X1` (the [MASK] is a type of X_1)		0.04	0.12	0.14	0.38	0.22	0.36
Hypernym-of	V	`como [MASK] e outros modos de X1` (like [MASK] and other modes of X_1)		0.00	0.04	0.08	0.54	**0.20**	**0.58**
Hypernym-of	V	`como [MASK] ou outras maneiras de <r>` (like [MASK] and other manners of X_1)		0.00	0.02	0.12	0.42	0.08	0.24
Hyponym-of	Abstract	`um X1, isto é, um tipo de [MASK]` (a X_1, this is, a type of [MASK])	V	0.02	0.46	0.24	0.60	**0.40**	0.62
Hyponym-of	Abstract	`uma X1, isto é, uma espécie de [MASK]` (a X_1, this is, a kind of [MASK])	V	0.06	0.38	0.12	**0.66**	0.28	0.64
Hyponym-of	Concrete	`uma X1, isto é, um tipo de [MASK]` (a X_1, this is, a type of [MASK])	V	0.10	0.40	**0.60**	**0.88**	0.56	0.80
Hyponym-of	Concrete	`um X1, isto é, um tipo de [MASK]` (a X_1, this is, a type of [MASK])	V	0.06	0.32	0.58	**0.88**	0.58	**0.88**
Hyponym-of	V	`como X1 ou outras maneiras de [MASK]` (like X_1 and other manners of [MASK])		0.18	0.54	**0.24**	0.64	0.18	**0.70**
Hyponym-of	V	`X1 é como [MASK], mas` (X_1 is like [MASK], but)		0.08	0.10	0.08	0.24	0.12	0.50

Table 4: Best performing patterns in TALES and their performance.

Relation	PoS	BERT-base	BERT-large	Sim	3CosAdd	3CosAvg	LRCos
Synonym-of	N	0.28	0.20	0.28^{*}	0.18^{*}	$0.32^{\times}$	$\mathbf{0.38}^{+}$
Synonym-of	V	0.12	**0.34**	0.20^{+}	0.12^{+}	0.24^{+}	0.30^{+}
Synonym-of	ADJ	0.06	0.24	0.26^{*}	0.10^{*}	$\mathbf{0.28}^{+}$	0.26^{+}
Antonym-of	ADJ	0.26	**0.38**	0.20^{*}	0.14^{*}	0.24^{+}	0.28^{*}
Hypernym-of	Abstract	0.22	**0.38**	0.20^{+}	$0.06^{\times+}$	0.20^{+}	0.16^{*+}
Hypernym-of	Concrete	0.20	**0.24**	0.18^{+}	$0.10^{\times}$	0.20^{*}	0.20^{+}
Hypernym-of	V	0.08	0.20	0.14^{*}	$0.08^{\times}$	0.12^{+}	$\mathbf{0.22}^{*}$
Hyponym-of	Abstract	0.24	**0.40**	0.08^{*}	0.08^{*}	0.10^{*}	0.12^{*}
Hyponym-of	Concrete	**0.60**	0.56	0.10^{+}	$0.04^{\times*+}$	0.14^{+}	$0.28^{\times}$
Hyponym-of	V	**0.24**	0.18	0.14^{+}	0.16^{*}	$0.16^{\times+}$	0.22^{+}

Table 5: Accuracy of analogy-solving methods in TALES. $^{\times}$GloVe; *word2vec-skip; $^{+}$word2vec-cbow.

Bouraoui et al., 2020). In any case, having in mind reproducibility and future improvements, the list of patterns was made available for anyone willing to use it.

Acknowledgements

This work was funded by national funds through FCT, within the scope of the project CISUC (UID/CEC/00326/2020) and by European Social Fund, through the Regional Operational Program Centro 2020. It is also based upon work from COST Action CA18209 Nexus Linguarum, supported by COST (European Cooperation in Science and Technology). http://www.cost.eu/.

References

Badr AlKhamissi, Millicent Li, Asli Celikyilmaz, Mona Diab, and Marjan Ghazvininejad. 2022. A review on language models as knowledge bases. https://arxiv.org/abs/2204.06031.

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.

Haw-Shiuan Chang, Ziyun Wang, Luke Vilnis, and Andrew McCallum. 2018. Distributional inclusion vector embedding for unsupervised hypernymy detection. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Com-*

putational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 485–495.

Anastasia Chizhikova, Sanzhar Murzakhmetov, Oleg Serikov, Tatiana Shavrina, and Mikhail Burtsev. 2022. Attention understands semantic relations. In *Proceedings of the Language Resources and Evaluation Conference*, pages 4040–4050, Marseille, France. European Language Resources Association.

Sandra Collovini de Abreu, Tiago Luis Bonamigo, and Renata Vieira. 2013. A review on relation extraction with an eye on Portuguese. *Journal of the Brazilian Computer Society*, 19(4):553–571.

Maria Cláudia de Freitas and Violeta Quental. 2007. Subsídios para a elaboração automática de taxonomias. In *Anais do XXVII Congresso da SBC*, pages 1585–1594.

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

Aleksandr Drozd, Anna Gladkova, and Satoshi Matsuoka. 2016. Word embeddings, analogies, and machine learning: Beyond king - man + woman = queen. In *Proc 26th International Conference on Computational Linguistics: Technical papers COLING 2016*, COLING 2016, pages 3519–3530.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the ACL*, 8:34–48.

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

Dayne Freitag, Matthias Blume, John Byrnes, Edmond Chow, Sadik Kapadia, Richard Rohwer, and Zhiqiang Wang. 2005. New experiments in distributional representations of synonymy. In *Procs 9th Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 25–32, Ann Arbor, Michigan. ACL.

Cláudia Freitas, Diana Santos, Hugo Gonçalo Oliveira, and Violeta Quental. 2015. VARRA: Validação, Avaliação e Revisão de Relações semânticas no AC/DC. In *Pesquisas e perspetivas em linguistica de corpus (Livro do IX Encontro de Linguística de Corpus, 2010)*, ELC 2010, pages 199–232. Mercado de Letras, Rio Grande do Sul, Brasil.

Yoav Goldberg. 2019. Assessing BERT's syntactic abilities. `https://arxiv.org/abs/1901.05287`.

Hugo Gonçalo Oliveira. 2022. Drilling lexico-semantic knowledge in Portuguese from BERT. In *Computational Processing of the Portuguese Language – 14th International Conf, PROPOR 2022*, volume 12037 of *LNCS*, pages 387—397. Springer.

Hugo Gonçalo Oliveira, Diana Santos, Paulo Gomes, and Nuno Seco. 2008. PAPEL: A dictionary-based lexical ontology for Portuguese. In *Proceedings of Computational Processing of the Portuguese Language - 8th International Conference (PROPOR 2008)*, volume 5190 of *LNCS/LNAI*, pages 31–40, Aveiro, Portugal. Springer.

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 *Proceedings 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 *Proceedings 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.

Nathan S. Hartmann, Erick R. Fonseca, Christopher D. Shulby, Marcos V. Treviso, Jéssica S. Rodrigues, and Sandra M. Aluísio. 2017. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In *Proc of 11th Brazilian Symposium in Information and Human Language Technology (STIL 2017)*.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proc 14th Conference on Computational Linguistics*, COLING 92, pages 539–545. ACL.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Ruixi Lin and Hwee Tou Ng. 2022. Does bert know that the is-a relation is transitive? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 94–99.

Ilia Markov, Nuno Mamede, and Jorge Baptista. 2014. Automatic Identification of Whole-Part Relations in Portuguese. In *Proceedings of 3rd Symposium on Languages, Applications and Technologies*, volume 38 of *OASICS*, pages 225–232. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

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

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

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

Patrick Pantel and Marco Pennacchiotti. 2006. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Procs of 21st International Conference on Computational Linguistics and 44th annual meeting of the Association for Computational Linguistics*, pages 113–120, Sydney, Australia. ACL Press.

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.

Diana Santos and Eckhard Bick. 2000. Providing Internet access to Portuguese corpora: the AC/DC project. In *Proceedings of 2nd International Conference on Language Resources and Evaluation*, LREC 2000, pages 205–210.

Rion Snow, Daniel Jurafsky, and Andrew Ng. 2005. Learning syntactic patterns for automatic hypernym discovery. *Advances in neural information processing systems*, 17:1297–1304.

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

Rodrigo Wilkens, Leonardo Zilio, Eduardo Ferreira, and Aline Villavicencio. 2016. B2SG: a TOEFL-like task for Portuguese. In *Procs 10th International Conference on Language Resources and Evaluation (LREC 2016)*. ELRA.