BATS-PT: Assessing Portuguese Masked Language Models in Lexico-Semantic Analogy Solving and Relation Completion

Hugo Gonçalo Oliveira[a,b] (hroliv@dei.uc.pt), **Ricardo Rodrigues**[a,c], **Bruno Ferreira**[a,b], **Purificação Silvano**[d], and **Sara Carvalho**[e,f]
[a]CISUC, LASI, Portugal
[b]DEI, University of Coimbra, Portugal
[c]ESEC, Polytechnic Institute of Coimbra, Portugal
[d]CLUP / FLUP, University of Porto, Portugal
[e]CLLC / DLC, University of Aveiro, Portugal
[f]NOVA CLUNL, Portugal

Abstract

This paper presents BATS-PT, the manual translation of the lexicographic portion of the Bigger Analogy Test Set (BATS) to European Portuguese. BATS-PT covers ten types of lexico-semantic analogies and can be used for assessing word embeddings and language models. Following this, the dataset is showcased while assessing two pretrained language models for Portuguese, BERTimbau and Albertina, in two tasks: analogy solving and relation completion, both in zero- and few-shot mask-prediction approaches. Experiments reveal different performance across relations and, in both tasks, the best overall performance was achieved with BERTimbau, in a five-shot scenario. We further discuss the limitations of the reported experiments and directions towards future improvements in these tasks.

1 Introduction

A word analogy is a statement of the kind <a> *is to* <b> *as* <c> *is to* <d>, i.e., where the relation between a and b also holds between c and d. A classic example would be *man is to king as woman is to queen*. The goal of analogy solving is to predict d, given a, b and c. In the last ten years, this task has been widely adopted as a benchmark for models of distributional similarity (Mikolov et al., 2013). Following the evolution of technological trends in Natural Language Processing (NLP), it has also been used for assessing language models (Ushio et al., 2021).

The Bigger Analogy Test Set (BATS) (Gladkova et al., 2016) is a dataset that differs from previous datasets of analogies by being larger and balanced across relations of different categories and types. Another difference is that it addresses the possibility of several correct values of d, which is very common in some relations. However, as with other datasets, BATS was initially created only for English.

In this paper, we present BATS-PT, which results from translating a part of BATS, namely the lexico-semantic relations, to Portuguese. Traditionally found in *wordnets* (Fellbaum, 1998), these relations are important for representing the meaning of language. In fact, if language models do represent them well, they can be seen as an alternative to knowledge bases (Petroni et al., 2019), in this case, to existing Portuguese lexical knowledge bases (Gonçalo Oliveira, 2018). Lexico-semantic relations are one category of relations where it is crucial to accept more than a possible answer d, as enabled by BATS. For instance, in *apple is to fruit as dog is to* d, suitable values for d would include *animal*, *mammal*, or *vertebrate*.

For Portuguese, another analogy dataset has been translated (Querido et al., 2017), but it is neither focused on lexico-semantic relations nor on the aforementioned features of BATS. Moreover, TALES (Gonçalo Oliveira et al., 2020) is a dataset inspired by BATS, but created automatically, whereas BATS-PT was translated manually by native speakers of European Portuguese. The creation of BATS-PT was done in the scope of a larger effort that includes the translation of BATS to at least 15 languages (Gromann et al., 2024). It may thus be seen as a standard benchmark for assessing language models in different languages and, because alignments were kept in the process, it can also be used for cross-lingual tasks.

After describing the creation of BATS-PT, we report on its usage in two tasks: analogy solving and relation completion. The latter is a variation of analogy, for which the target relation is given. It is especially useful for knowledge base completion (Petroni et al., 2019). Both tasks are performed in zero- and few-shot scenarios, in two available masked language models (MLMs) pretrained for Portuguese: BERTimbau (Souza et al., 2020) and Albertina (Rodrigues et al., 2023). So, besides showcasing the dataset, we draw some conclusions

on both tasks, such as the impact of zero- and few-shot approaches on the performance of each model.

The main conclusion is that MLMs perform poorly in the tackled tasks, but interesting points remain for discussion. For instance, performance varies significantly across different relations, but generally improves in the few-shot scenario. BERTimbau performed more consistently and was, overall, the best model.

In the remainder of the paper, we review similar datasets and translations of BATS to other languages (Section 2), describe the creation of BATS-PT in more detail (Section 3), report on the performed experiments and discuss their results (Section 4), and present final conclusions, also pointing out future directions (Section 5).

2 Related Work

What is probably the most popular dataset for analogy solving, later known as the Google Analogy Test Set (GATS), was originally used for assessing regularities in *word2vec* (Mikolov et al., 2013). In such models, analogies are traditionally computed with the vector offset method, also known as 3CosAdd ($\vec{d} = \vec{b} + \vec{c} - \vec{a}$).

GATS has about 19,000 analogy tuples (a, b, c, d) organised according to nine syntactic (e.g., adjective to adverb, opposite, comparative, verb tenses) and five semantic (e.g., capital-country, currency, male-female) categories, with between 20 and 70 examples per category. BATS (Gladkova et al., 2016) was created as a balanced alternative to GATS, while covering additional relations (e.g., lexico-semantic). It is organised into four categories of relations — inflexion morphology, derivational morphology, lexicographic semantics, and encyclopedic semantics —, and ten relations for each category. For each relation, there are exactly 50 entries of the type $source \rightarrow \{targets\}$, such that the relation holds between the source and each of its targets. Therefore, BATS supports analogies for which there is more than a single correct d, as it happens for many lexico-semantic relations. The data in BATS can be combined in a total of 99,200 analogy tuples.

BATS, originally developed for English, was translated to other languages, namely Japanese (Karpinska et al., 2018), Icelandic (Friðriksdóttir et al., 2022) and, more recently, six other languages, in a dataset christened as MATS (Multilingual Analogy Test Set) (Mickus et al., 2023). None of them was Portuguese.

GATS, on the other hand, was translated to Portuguese (Rodrigues et al., 2016). Also, TALES (Gonçalo Oliveira et al., 2020), with similar features to BATS, was created automatically, based on the contents of ten lexical resources for Portuguese. TALES adopts the format of BATS but targets lexico-semantic relations only, in a total of 14 files, also with 50 $source \rightarrow \{targets\}$ entries each, covering hypernymy, hyponymy, synonymy, antonymy, part-of, and purpose-of relations.

A related dataset for Portuguese is B2SG (Wilkens et al., 2016) where, given a lexico-semantic relation (hypernymy, synonymy, antonymy) and a source word, a target word has to be identified among four options. Another related dataset was created for studying how language models deal with homonymy and synonymy (Garcia, 2021), including sentences and target words in context. Part of the previous dataset can be used similarly to the Word-In-Context (WIC) (Pilehvar and Camacho-Collados, 2019) dataset.

To the best of our knowledge, work on analogy solving in Portuguese is limited to using word embeddings and the translation of GATS (Rodrigues et al., 2016; Hartmann et al., 2017; Sousa et al., 2020). Notwithstanding, relation completion has been tackled in TALES with BERTimbau (Gonçalo Oliveira, 2023). This takes advantage of the text completion capabilities of current language models, which have been tested in the acquisition of different kinds of knowledge, towards their utilisation as knowledge bases (Petroni et al., 2019; AlKhamissi et al., 2022). A set of patterns that indicate the relations in text (e.g., Hearst (1992) patterns) is first necessary. When instantiated with the source word and a mask instead of the target (e.g., *a dog is a type of [MASK].*), the goal is to predict suitable words for the mask (i.e., valid targets). Patterns can be handcrafted or discovered automatically from corpora (Bouraoui et al., 2020).

BERTimbau and Hearst (1992) patterns have also been used for classifying pairs of Portuguese words holding a relation of hypernymy or not (Paes, 2021). Training and evaluation data was extracted specifically for this work, automatically from two Portuguese knowledge bases.

For other languages, many approaches for analogy solving and related tasks are based on prompting pretrained models, in zero- or few-shot scenarios. This is mostly due to the size of the available datasets, but also because knowledge tends to be

forgotten during the fine-tuning process (Wallat et al., 2020).

Multilingual BERT (mBERT) was used for solving analogies in the seven languages of MATS (Mickus et al., 2023). In order to discriminate correct analogy pairs, another prompt-based approach for analogy solving computes the perplexity of analogy templates instantiated by analogy tuples (Ushio et al., 2021). The authors experimented with both MLMs and GPT-2 (Radford et al., 2019), with the latter performing better than mBERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020). Another example of using GPT-like models for analogy solving is GPT-3 (Brown et al., 2020), which was originally tested on a dataset of 374 analogies in English, in zero- and few-shot scenarios.

3 Dataset Creation

BATS-PT was created in the scope of a larger effort, which aimed at the translation of the lexicographic relations portion of BATS to several languages of different families (Gromann et al., 2024). The translation is currently concluded for a total of 13 languages, in alphabetical order: Albanian, Croatian, French, German, Greek, Hebrew, Italian, Lithuanian, Portuguese, Romanian, Slovak, Slovenian, and Spanish; and almost for two other languages: Bambara and Macedonian.

Lexicographic relations make up one-quarter of BATS and include the following ten relations:

- L01 [hypernyms – animals];
- L02 [hypernyms – misc];
- L03 [hyponyms – misc];
- L04 [meronyms – substance];
- L05 [meronyms – member];
- L06 [meronyms – part];
- L07 [synonyms – intensity];
- L08 [synonyms – exact];
- L09 [antonyms – gradable];
- L10 [antonyms – binary].

For each relation, there are exactly 50 source words, each with a variable number of targets.

All translations were performed manually, by native speakers of the target languages. Since the context of the words in BATS is limited to the source and its targets, automatic translation would not be suitable.

During the translation process, correspondence between sources, targets and their English counterparts was kept. To some extent, this limits the initial range of source words. Nevertheless, it ensures that each language version of the dataset, including BATS-PT, is aligned with the English BATS, further enabling multilingual tasks. Despite the previous alignments, in this paper, we are focused on Portuguese, so we use BATS-PT in the original BATS format. Table 1 illustrates this format, which can be easily obtained from the aligned format.

The translation to European Portuguese was performed by four native speakers of this variety, all senior researchers: two linguists and two computer scientists with expertise in NLP. Each one was responsible for a part of the dataset, but a file comprising 20 entries (i.e., two randomly selected entries for each relation, specifically, those in Table 1) was translated by the four translators, independently, to measure inter-annotator agreement. Fleiss' kappa was 0.62, in the lower boundary of substantial agreement, which gives us confidence in the general consensus of the dataset.

General issues that arose during the translation process, and how to handle them, were discussed in meetings with the translators for other languages. To keep the dataset aligned, the following were marked: (*i*) translation to the target language is not possible or quite cumbersome (marked as *no translation* — e.g., *garden truck* or *hamdog* to Portuguese); (*ii*) translation of the target word was already used as the translation of another target of the same source (*duplicate translation* — e.g., *backpack*, *rucksack* and *knapsack*, all translated to *mochila* in Portuguese). We should note that both annotations were used exclusively for analysis purposes. For evaluation, untranslated words were not used as targets, whereas duplicates would result in a single target word. Moreover, translators were free to add additional target words, specific to their language, and not covered by the English targets. This was especially encouraged for sources with many duplicate targets or targets with no translation, as an attempt to keep a similar number of targets as in the original English dataset. Still, when all sources could be translated, the limitation of the original range could arise. For instance, in the first entry for L04 in Table 1, there would probably be more obvious sources (i.e., substances for box) than the original ones, but since translations were found for each original source (i.e., *cardboard*, *tin*, *boxwood*, *turkish_boxwood*), the translator felt no

File	Source	Targets
L01	coiote	canino/vertebrado/criatura/canídeo/../mamífero/../coisa_viva
	leão	felino/gato/animal/organismo/fauna/placentário/carnívoro/../grande_felino
L02	bolo	sobremesa/produtos_cozinhados/../alimento/../alimentação/mantimentos/..
	limão	citrino/fruto/fruto_comestível/../comida/matéria/objecto_natural/..
L03	igreja	capela/abadia/basílica/catedral
	joalharia	pulseira/conta/missangas/../bracelete/../botões_de_punho/brinco/gema/..
L04	caixa	cartão/estanho/madeira_de_buxo/madeira_de_buxo_turca
	nuvem	vapor/água/vapor_de_água
L05	elefante	manada
	agente	polícia
L06	dia	hora/manhã/entardecer/nanossegundo/meio-dia/fentossegundo/h/minutos/..
	rádio	receptor/sintonizador/../transmissor/../aparelho/amplificador/..
L07	lago	mar/oceano
	pónei	cavalo
L08	caminho_de_ferro	ferrovia
	margem	costa/praia/borda/orla
L09	consciente	desatento/inconsciente/insuspeito/a_dormir/../indiferente/desinformado
	barulhento	silencioso/não_comunicativo/mudo/desarticulado/calado/emudecido
L10	baixo	cima/acima/à_frente/../ressuscitado/brotado/ascendente/em_cima/subida
	subida	descida/declínio/queda/declive/inclinado_para_baixo

Table 1: Example entries in BATS-PT. Two entries are shown for each relation, corresponding to two entries in the respective file.

need of adding extra words (e.g., plastic or glass). Nonetheless, this option might be revisited in the future.

When necessary, meetings were also held between the four Portuguese translators to discuss specific issues of this language. In addition to the knowledge of the translators, available sources were consulted for the translations, including English–Portuguese dictionaries; Wikipedia and its cross-lingual links; automatic translation services like DeepL, which translate from English to European Portuguese; and even searching the Web for tentative translations to check if they do exist, mostly for multiword expressions.

In the end, all 500 source words were translated into Portuguese. Table 2 shows the main figures of the resulting dataset, including the number of translated sources, targets, Portuguese-specific targets added (Extra), untranslated targets (NT), and duplicate translations (Dup). We stress that, despite the balanced number of sources, the number of targets is variable across relations. We also note that, despite the inverse nature of some relations (e.g., hypernymy–hyponymy) and the symmetry of others (e.g., synonymy and antonymy), for purposes of uniformity, each entry of the dataset should be considered unidirectionally, i.e., *source* → *target*, thus reflecting the guidelines for the original BATS. After excluding duplicates and not translated targets, there are slightly more than 5,000 targets (4572 + 451) in total. Out of them, 1,123 are in the hyponymy relations (L03), whereas several relations have less than 200 targets (i.e., meronyms-substance, meronyms-member, synonyms-exact, antonyms-binary), a similar picture as in the original BATS.

Rel	Sources	Targets	Extra	NT	Dup
L01	50	726	+19	1	94
L02	50	687	0	2	105
L03	50	1123	+113	20	349
L04	50	192	0	0	5
L05	50	110	+5	0	3
L06	50	654	+7	5	177
L07	50	206	+62	1	46
L08	50	146	+52	3	39
L09	50	581	+178	14	280
L10	50	147	+15	4	38
All	500	4572	+451	50	1136

Table 2: BATS-PT in numbers.

4 Experiments

This section reports on two experiments using BATS-PT: analogy solving and relation completion. These were performed with two available language models pretrained for Portuguese, BERTimbau (Souza et al., 2020) and Albertina (Rodrigues et al., 2023), both described next.

4.1 Language Models

Methods for both tasks are based on prompting the selected language models, pretrained in the masked language modelling task. Models were accessed through the HuggingFace hub, using the transformers library.

We used the largest BERTimbau, BERTimbau-large[1], which is based on BERT (Devlin et al., 2019) and trained in Brazilian Portuguese (PTBR) texts. It has 24 layers and 335M parameters.

Albertina is a more recent model, also with 24 layers, but with 900M parameters. It is based on DeBERTA (He et al., 2020) and has two versions: one for European Portuguese (PTPT) and another for Brazilian Portuguese (PTBR). Since BATS-PT targets the European variety, we used Albertina PTPT[2].

4.2 Analogy Solving

BATS was originally created for assessing word embeddings in analogy solving tasks. Therefore, this was the first task we have addressed using BATS-PT.

Adopted approaches were based on prompting the models with a classic template for analogy. More precisely, in order to answer the question *What is to <c> as <a> is to <b>?*, the following prompt was used:

`<a> está para <b> assim como <c> está para [MASK].`

The goal of the model was to predict the most suitable token for the *[MASK]*.

This was performed for every combination of pairs $(a, b), (c, d)$ holding the same relation, i.e., since there were 50 sources for each relation, $50 \times 49 = 2,450$ analogies were computed for each relation, 24,450 in total[3].

[1] https://huggingface.co/neuralmind/bert-large-portuguese-cased

[2] https://huggingface.co/PORTULAN/Albertina-900m-portuguese-ptpt-encoder

[3] In fact, towards a balanced training data, we have used only the first target word for each source; otherwise, there would be many more combinations.

Shots	Prompt
0	`verdadeiro está para falso assim como saída está para [MASK].`
5	`dentro está para fora assim como sudeste está para sudoeste. sul está para norte assim como ocupado está para vago. cimo está para fundo assim como para a frente está para para trás. elevar está para afundar assim como para trás está para para a frente. seguir está para retirar assim como empregar está para demitir. verdadeiro está para falso assim como saída está para [MASK].`

Table 3: Prompts for the antonymy analogy: *verdadeiro está para falso assim como saída está para entrada*.

Moreover, tests were performed in a zero-shot, but also in a five-shot scenario, where the prompt was concatenated to five complete prompts, generated from ten other pairs in the dataset, holding the same relation. These pairs were selected automatically, but we made sure that, for every tested model, the shots for every (a, b, c, d) tuple were generated from exactly the same pairs. Table 3 has an example for a zero- and a five shot prompt for the analogy *verdadeiro está para falso assim como saída está para entrada* — in English, *true* is to *false* as *exit* is to *entry*.

Table 4 reports on the accuracy of each model, according to the scenario and relation. Since this was the first time BATS-PT was used, classic methods for analogy solving were also computed on 300-sized GloVe embeddings pretrained in Brazilian Portuguese text (Hartmann et al., 2017). These were the vector offset, also known as 3CosAdd ($d = argmax_{w \in vocab}(\vec{b} - \vec{a} + \vec{c})$); and 3CosAvg (Drozd et al., 2016), similar to 3CosAdd, but instead of a pair (a, b), it relies on the average vector in a set of given pairs. For each (c, d), 3CosAdd was computed for every (a, b) in the same file of the dataset. This was also true for 3CosAvg, however, $(\bar{a}, \bar{b})$ was the average of 11 vectors, i.e., (a, b) plus the same ten pairs used for the MLMs in the five-shot learning scenario.

Performance varies across relations, but it is clear that solving lexico-semantic analogies automatically is still challenging with the used models. Even when not limited to a single answer (d), as in BATS, accuracy is always lower than 0.50. Nevertheless, using MLMs is a better option than traditional word embeddings. This is especially true for BERTimbau, which achieved the best performance in nine relations and overall. Seven of those were achieved in the five-shot scenario, which shows

Relation	GloVe		BERTimbau		Albertina	
	3CAdd	3CAvg(11)	0-shot	5-shot	0-shot	5-shot
L01	0.09	0.12	0.06	0.16	0.65	**0.73**
L02	0.05	0.08	0.12	**0.22**	0.02	0.04
L03	0.05	0.10	0.10	**0.19**	0.06	0.13
L04	0.05	0.06	0.32	**0.34**	0.12	0.10
L05	0.03	0.06	0.22	**0.30**	0.08	0.08
L06	0.02	0.00	0.08	**0.12**	0.08	0.06
L07	0.04	0.12	0.12	**0.16**	0.02	0.04
L08	0.03	0.07	**0.14**	0.10	0.00	0.00
L09	0.05	0.15	0.39	**0.47**	0.14	0.16
L10	0.16	0.27	**0.46**	0.41	0.22	0.26
Average	0.06	0.10	0.20	**0.25**	0.14	0.16

Table 4: Accuracy of Analogy Solving in BATS-PT, according to model, scenario and relation.

that the model can learn from a small number of examples. Exceptions are in L08 (exact synonyms) and L10 (binary antonyms), where BERTimbau performs better in zero-shot, and L01 (animals hypernyms), where Albertina achieved an impressive performance of 0.73 in the five-shot scenario.

A closer inspection of the previous results shows that, for many analogies, Albertina predicts the word *animal*, which is a valid d for most analogies of this relation. As for L08 and L10, after L05, they are the relations with the lower number of targets, which limits the number of correct answers. Specifically in L08, we also observe some confusion with co-hyponyms (e.g., *criança* for *bebé*; or *carro* and *moto* for *bicicleta*), which increases with five-shot learning. For L10, our explanation is that it contains many adverbs (e.g., *após* $\rightarrow$ *antes* or *dentro* $\rightarrow$ *fora*), which may occur in many different contexts, but less naturally in the analogy pattern (e.g., *dentro está para fora assim como após está para antes*), also resulting in additional confusion with five-shot learning, where this pattern is repeated six times. This could, perhaps, be minimised if the related words were quoted in the prompts, as tested by Mickus et al. (2023), but we leave this analysis for future work.

The best performance of BERTimbau was for L09 (gradable antonyms), while it performed worst in L08 (exact synonyms) and L06 (part meronyms). These are followed by L07 (intensity synonyms) and L01, where Albertina performed the best.

We note that the reported results are limited by using MLMs, which predict tokens for the mask. However, some targets in the dataset have more than one token, starting with multiword expressions. Still, we also note that every source word has at least one single-word target, so the impact of the previous should not be too high.

These results are in line with those in BATS and in its translation to other languages (Mickus et al., 2023), which vary between 0.05 (Chinese) and 0.22 (English). However, a deeper analysis of the previous work tells us that the approach is not directly comparable to ours. On the one hand, it uses a multilingual model instead of a monolingual one and does not test few-shot learning. On the other hand, in the previous translations, multiword expressions were excluded. Moreover, when looking at their code, we notice another important difference: instead of computing a single analogy for each tuple (a, b, c, d), they compute analogies with all the possible targets of a in the position of b, and with a variable number of masks, based on the tokenization of all correct answers d. If at least one of the previous predictions is correct, the analogy for the tuple is considered correct, which has a positive bias on accuracy.

4.3 Relation Completion

The second tackled task was relation completion, where BATS can also be used as a benchmark. The main difference to analogy solving is that instead of an analogous pair (a, b), a relation is provided — for instance, in the form of a pattern. Specifically, given a relation r and a word a, the goal becomes to predict b, such that r holds between a and b.

For Portuguese, relation completion has previously been assessed in TALES (Gonçalo Oliveira et al., 2020), a dataset with a similar structure as BATS-PT, though created automatically and not

covering exactly the same lexico-semantic relations. Different approaches for this task have been tested in TALES, including prompting BERTimbau in a zero-shot scenario (Gonçalo Oliveira, 2023).

Here, we adopt a similar approach, but include also the model Albertina and few-shot learning. For this of approach, the relation was expressed in text. Since there are many ways of doing it, we devised two groups of prompts, and, for each relation, tested one prompt from each group. In the first group, hereafter relation prompts, the relation is explicitly mentioned (see the templates in Table 5).

Relation	Prompt
L01 / L02	`[MASK] é hiperónimo de <a>.`
L03	`[MASK] é hipónimo de <a>.`
L04	`[MASK] é substância de <a>.`
L05	`<a> é membro de [MASK].`
L06	`[MASK] é parte de <a>.`
L07 / L08	`[MASK] é sinónimo de <a>.`
L09 / L10	`[MASK] é antónimo de <a>.`

Table 5: Relation prompts used for each relation in BATS-PT.

In the second group, hereafter corpora prompts, the prompt is a pattern where one would commonly find the related words in raw corpora, for instance, like Hearst (1992) patterns. Since many different patterns could be used for the same relation, we selected the best of this kind in equivalent relations in TALES, with BERTimbau (Gonçalo Oliveira, 2023). As the previous did not consider meronymy relations, the corpora prompts for relations L04, L05 and L06 were selected empirically (see templates in Table 6). Some of the patterns used were obtained from VARRA (Freitas et al., 2015), a service for searching for and validating instances of lexico-semantic relations by resorting to Portuguese corpora.

Relation	Prompt
L01 / L02	`<a>, isto é, um tipo de [MASK].`
L03	`[MASK] é um tipo de <a>.`
L04	`<a> é constituído por [MASK].`
L05	`[MASK] tem <a>.`
L06	`<a> tem [MASK].`
L07 / L08	`<a> é o mesmo que [MASK].`
L09 / L10	`<a> é o contrário de [MASK].`

Table 6: Corpora prompts used for each relation in BATS-PT.

The performance of the models is summarised in Tables 7 and 8, respectively using the relation and the corpora prompts.

Relation	BERTimbau		Albertina	
	0-shot	5-shot	0-shot	5-shot
L01	0.00	**0.80**	0.00	**0.80**
L02	0.00	**0.26**	0.00	0.00
L03	0.00	0.09	0.09	**0.22**
L04	0.00	**0.21**	0.00	0.04
L05	0.04	**0.21**	0.00	0.04
L06	0.00	**0.16**	0.12	0.12
L07	0.00	**0.04**	**0.04**	0.00
L08	0.00	**0.13**	0.00	0.00
L09	0.04	**0.30**	0.00	0.04
L10	0.09	**0.48**	0.09	0.22
Average	0.02	**0.27**	0.03	0.15

Table 7: Accuracy of Relation Completion in BATS-PT, using relation prompts, according to model, scenario and relation.

Relation	BERTimbau		Albertina	
	0-shot	5-shot	0-shot	5-shot
L01	0.40	0.08	**0.48**	0.12
L02	**0.30**	0.22	0.00	0.00
L03	0.00	0.09	0.00	**0.26**
L04	**0.17**	**0.17**	0.00	0.04
L05	0.00	**0.13**	0.08	0.04
L06	0.00	**0.08**	0.04	**0.08**
L07	0.00	**0.04**	0.00	0.00
L08	**0.13**	**0.13**	0.00	0.00
L09	0.30	**0.43**	0.00	0.09
L10	0.22	**0.43**	0.09	0.22
Average	0.15	**0.18**	0.07	0.08

Table 8: Accuracy of Relation Completion in BATS-PT, using corpora prompts, according to model, scenario and relation.

Relation completion seems to be even more challenging than analogy solving for MLMs. Performance is also variable across relations, it also improves in the five-shot scenario, and BERTimbau is again the best overall model. Another conclusion is that corpora prompts are the best for zero-shot, but the improvements of few-shot learning are more reflected in the relation prompts. So much so that the best overall performance is achieved with these prompts in the five-shot scenario. One possible explanation is that corpora prompts are closer to what the models learned from, thus the best performance in zero-shot. At the same time, relation prompts

are shorter and more structured, thus helping the model to learn a pattern in the few-shot scenario.

Even in few-shot, the most challenging relation is L07 (intensity synonyms). A possible reason is that it opens the notion of synonym, while the dataset still has a limited number of correct targets. As it has happened for analogy, antonymy relations (L09, L10) are among the best performing. Nonetheless, we would highlight two relations that deviate from the average: L01 (animal hypernyms) and L03 (hyponyms). In both models, the fact that most entries in L01 have *animal* has a hypernym has a positive impact on the performance of few-shot with the relation prompt. However, when it comes to the corpora prompt, accuracy is substantially higher in the zero-shot scenario. This is mostly a consequence of the prompt used, which is long enough to capture the relation, but, when concatenated with more sequences alike, confuses the model. In fact, using the same prompt in L02 has a similar effect.

Relation L03 is the only one for which Albertina achieves top accuracy. After inspecting the results, we note that, with the used prompt, BERTimbau predicts many functional words like, for instance, *não*, *este*, *ele*, or *pois*, whereas Albertina does not suffer so much from this. This could be fixed by adding an article to the start of the prompt, but it would bias the predictions towards the gender of the article. This is why we have used only gender-neutral prompts, but they end up having their limitations. Another option would be to add quotes both around <*a*> and around the *[MASK]*, as Mickus et al. (2023) did for analogy.

So, the used prompts do have an impact on the results. We stress that the reported scores are based on a single prompt for each relation, and that some of those prompts were selected based on their performance in a different dataset, but with BERTimbau. Accuracy could possibly be improved with other prompts (for instance, selected specifically for Albertina), or by combining the predictions of different prompts. This adds to the aforementioned limitation of MLMs, which predict single tokens only. Since the main goal of this paper is to present and showcase the dataset, we leave prompt engineering and alternative approaches for future work.

We can still say that the accuracy of BERTimbau in zero-shot hypernymy and antonymy completion is similar to that of the same model and same relations in TALES (Gonçalo Oliveira, 2023). On the contrary, it is much lower for hyponymy (0.28–0.40 in TALES) and synonymy (0.20–0.34 in TALES). This suggests that, due to its automatic creation, TALES has a higher coverage of hyponyms and a broader sense of synonyms, which positively impacts accuracy. In fact, this is supported by the total number of targets in the synonymy relation files, much greater in TALES (533, 1,240, 615) than in BATS-PT (269, 196).

5 Conclusion

We presented a new test set of lexico-semantic analogies in Portuguese, BATS-PT, resulting from the manual translation of the same analogies in BATS. We described the translation process, part of a multilingual effort, discussed the options taken and provided some figures on the dataset.

BATS-PT was then used for benchmarking two MLMs pretrained for Portuguese, BERTimbau and Albertina, in two language comprehension tasks: analogy solving and relation completion. We saw that performance varies across relations, and the highest is achieved in a five-shot scenario, where BERTimbau performed the best overall. This is somewhat surprising, given that BERTimbau has only one-third of the parameters of Albertina. Nevertheless, the best average accuracy was only 0.27, for analogy solving, and 0.18, for relation completion, showing that there is still much room for improvement in both tackled tasks.

Future approaches with MLMs should invest more in prompt engineering, consider multiple masks, as well as the combination of prompts. Generative language models should also be explored for both tasks, analogy solving and relation completion. In this case, the prompts must be adapted for text completion instead of mask prediction. Preliminary results of relation completion with GPT-3, in TALES and in an earlier version of BATS-PT, suggest that the performance of large generative models is far superior to that of MLMs, even when the latter consider a combination of prompts (Gonçalo Oliveira and Rodrigues, 2023). Specifically, with direct prompts like `lista os 10 hiperónimos, em português, da palavra <a>`, GPT-3 achieved an overall accuracy of 0.42 and 0.52, respectively in the zero- and five-shot scenarios. Stronger conclusions should follow experimentation with other models, ideally open source (e.g., BLOOM (Scao et al., 2022) or Llama2 (Touvron et al., 2023)), also in analogy solving.

In addition to BATS-PT, the adopted approaches

could be applied to other Portuguese datasets, such as TALES. So far, this dataset has only been used to assess zero-shot relation completion with BERTimbau and older models. In the future, Albertina may also be used for relation completion, while approaches for analogy solving may be tested with both models. This may help make stronger conclusions on the quality of TALES, which was created automatically.

The current version of BATS-PT is publicly available[4] for anyone willing to test other models, approaches or perform other experiments. For instance, in addition to analogy solving, a dataset like BATS enables further studies to understand better language models, such as analysing their ability to understand relations, their types and directionality (Rezaee and Camacho-Collados, 2022).

We should add that we are still discussing how to handle some of the issues in the original dataset. Once fixed, these might be reflected in a minority of differences in BATS-PT. We may also consider the translation of the files for the remaining relations in BATS to Portuguese: inflexion, derivational and encyclopedic relations. In fact, these have fewer targets and should be even more consensual, thus taking less time to translate.

Acknowledgements: This work was based upon activities carried out in the COST Action CA18209 Nexus Linguarum, supported by COST (European Cooperation in Science and Technology): http://www.cost.eu/; and financially supported by the Portuguese Recovery and Resilience Plan (PRR) through project C645008882-00000055, Center for Responsible AI.

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 *Proceedings of AAAI Conference on Artificial Intelligence*, pages 7456–7463. AAAI Press.

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

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of 2019 Conf of North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186. Association for Computational Linguistics.

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

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

Cláudia Freitas, Diana Santos, Hugo Gonçalo Oliveira, and Violeta Quental. 2015. VARRA: Validação, Avaliação e Revisão de Relações semânticas no AC/DC. In Ana Maria T. Ibaños, Lívia Pretto Mottin, Simone Sarmento, and Tony Berber Sardinha, editors, *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.

Steinunn Rut Friðriksdóttir, Hjalti Daníelsson, and Steinþór Steingrímsson. 2022. IceBATS: An Icelandic adaptation of the Bigger Analogy Test Set. In *Proceedings of the 13th Language Resources and Evaluation Conference*, LREC 2022, pages 4227–4234, Marseille, France. ELRA.

Marcos Garcia. 2021. Exploring the representation of word meanings in context: A case study on homonymy and synonymy. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint*

[4] https://github.com/NLP-CISUC/PT-LexicalSemantics/tree/master/BATS-PT

Conference on Natural Language Processing (Volume 1: Long Papers), pages 3625–3640, Online. Association for Computational Linguistics.

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 *Proceedings of NAACL 2016 Student Research Workshop*, pages 8–15. ACL.

Hugo Gonçalo Oliveira. 2018. Distributional and Knowledge-Based Approaches for Computing Portuguese Word Similarity. *Information*, 9(2).

Hugo Gonçalo Oliveira. 2023. On the acquisition of WordNet relations in Portuguese from pretrained masked language models. In *Proceedings of 12th Global WordNet Conference*, GWC, San Sebastian, Spain. ACL.

Hugo Gonçalo Oliveira, Tiago Sousa, and Ana Alves. 2020. TALES: Test set of Portuguese lexical-semantic relations for assessing word embeddings. In *Proceedings of the 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 and Ricardo Rodrigues. 2023. GPT3 as a Portuguese lexical knowledge base? In *Proceedings of the 4th Conference on Language, Data and Knowledge*, pages 358–363, Vienna, Austria. NOVA CLUNL, Portugal.

Dagmar Gromann, Hugo Gonçalo Oliveira, Lucia Pitarch, Elena-Simona Apostol, Jordi Bernad, Eliot Bytyçi, Chiara Cantone, Sara Carvalho, Francesca Frontini, Radovan Garabik, Jorge Gracia, Letizia Granata, Fahad Khan, Timotej Knez, Penny Labropoulou, Chaya Liebeskind, Maria Pia di Buono, Ana Ostroški Anić, Sigita Rackevičienė, Ricardo Rodrigues, Gilles Sérasset, Linas Selmistraitis, Mahammadou Sidibé, Purificação Silvano, Blerina Spahiu, Enriketa Sogutlu, Ranka Stanković, Ciprian-Octavian Truică, Giedrė Valūnaitė Oleškevičienė, Slavko Zitnik, and Katerina Zdravkova. 2024. MultiLexBATS: Multilingual Dataset of Lexical Semantic Relations. Submitted to LREC-COLING 2024.

Nathan Hartmann, Erick Fonseca, Christopher Shulby, Marcos Treviso, Jéssica Silva, and Sandra Aluísio. 2017. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In *Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology*, pages 122–131.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. DeBERTa: Decoding-enhanced BERT with disentangled attention. *arXiv preprint arXiv:2006.03654*.

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

Marzena Karpinska, Bofang Li, Anna Rogers, and Aleksandr Drozd. 2018. Subcharacter information in Japanese embeddings: When is it worth it? In *Proceedings of the Workshop on the Relevance of Linguistic Structure in Neural Architectures for NLP*, pages 28–37.

Timothee Mickus, Eduardo Calò, Léo Jacqmin, Denis Paperno, and Mathieu Constant. 2023. „Mann" is to "Donna" as 「国王」 is to « Reine » Adapting the Analogy Task for Multilingual and Contextual Embeddings. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 270–283, Toronto, Canada. Association for Computational Linguistics.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

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.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. 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 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

Andreia Querido, Rita Carvalho, João Rodrigues, Marcos Garcia, João Silva, Catarina Correia, Nuno Rendeiro, Rita Valadas Pereira, Marisa Campos, and António Branco. 2017. LX-LR4DistSemEval: A collection of language resources for the evaluation of distributional semantic models of Portuguese. *Revista da Associação Portuguesa de Linguística*, (3):265–283.

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, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

João Rodrigues, António Branco, Steven Neale, and João Silva. 2016. Lx-DSemVectors: Distributional semantics models for Portuguese. In *Computational Processing of the Portuguese Language: 12th International Conference, PROPOR 2016, Tomar, Portugal, July 13-15, 2016, Proceedings 12*, pages 259–270. Springer.

João Rodrigues, Luís Gomes, João Silva, António Branco, Rodrigo Santos, Henrique Lopes Cardoso, and Tomás Freitas Osório. 2023. Advancing neural encoding of Portuguese with transformer AlbertinaPT-*. In *Progress in Artificial Intelligence – 22nd EPIA Conference on Artificial Intelligence, EPIA 2023, Faial Island, Azores, September 5-8, 2023, Proceedings, Part I*, volume 14115 of *LNCS*, pages 441–453. Springer.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Tiago Sousa, Hugo Gonçalo Oliveira, and Ana Alves. 2020. Exploring different methods for solving analogies with Portuguese word embeddings. In *Proceedings 9th Symposium on Languages, Applications and Technologies, SLATE 2020, July 13-14, 2020, School of Technology, Polytechnic Institute of Cávado and Ave, Portugal*, volume 83 of *OASIcs*, pages 9:1–9:14. Schloss Dagstuhl - Leibniz-Zentrum für Informatik.

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

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 *Proceedings 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.

Rodrigo Wilkens, Leonardo Zilio, Eduardo Ferreira, and Aline Villavicencio. 2016. The portuguese b2sg: A semantic test for distributional thesaurus. In *Proceedings of 12th International Conference on Computational Processing of the Portuguese Language (PROPOR 2016)*, volume 9727 of *LNAI*, pages 115–121, Tomar, Portugal. Springer.