Proceedings of the Workshop on Hybrid Intelligence for Natural Language Processing Tasks HI4NLP (co-located at ECAI-2020)
Santiago de Compostela, August 29, 2020, published at http://ceur-ws.org

TALES: Test Set of Portuguese Lexical-Semantic Relations for Assessing Word Embeddings

Hugo Gonçalo Oliveira [1] and **Tiago Sousa** [2] and **Ana Alves** [3]

Abstract. Static word embeddings, like word2vec or GloVe, are often assessed when solving syntactic and semantic analogies. Among the latter, we are interested in relations that one would find in lexical-semantic knowledge bases like WordNet, also covered in analogy test sets for English. This paper describes the creation of a new test for assessing Portuguese word embeddings, dubbed TALES, with an exclusive focus on lexical-semantic relations, acquired from lexical resources in Portuguese. It further reports on the performance of methods previously used for solving analogies, with pre-trained Portuguese word embeddings, when applied to the created dataset, an experiment that revealed that TALES is challenging to solve. Results achieved are briefly discussed, with conclusions that may be useful for developing new approaches for this problem, possibly new embeddings, as well as future versions of TALES.

1 Introduction

When it comes to computational representations of the semantics of a language, two main approaches have been followed: lexical-semantic knowledge bases (LKBs), such as wordnets [7]; and distributional models, like word embeddings. The former organise words and their meanings, often connected by relations, such as Hypernymy or Part-of, and may include additional lexicographic information (part-of-speech, gloss), while the latter follow the distributional hypothesis [11] and represent words as vectors of numeric features, according to the contexts they are found in large corpora. On distributional models, since 2013 the trend was to use efficient methods that learn dense-vector representations of words, like word2vec [18] or GloVe [21]. Besides their utility for computing word similarity, such models have shown very interesting results for solving analogies of the kind "*what is to b as a^* is to a*"? (e.g., what is to Portugal as Paris is to France?). So much that both previous tasks are extensively used for assessing word embeddings in different languages.

Popular analogy test sets cover syntactic and semantic relations of different types, with some [9, 27] covering lexical-semantic relations. Given our interest on this kind of relations, we created a similar test for them, but in Portuguese, which we baptised as *Teste para Analogias Léxico-Semanticas* (TALES, in English, Test for Lexical-Semantic Analogies). While English tests could be translated to Portuguese, as the Google Analogy Test was [23], we tackled the creation of such a test from scratch, because different languages represent different socio-cultural realities, they do not cover exactly the same part of the lexicon and, even where they seem to be common, several concepts are lexicalised differently [14]. This is important because the created test aims to be used in the computational processing of Portuguese, for assessing Portuguese word embeddings, even if with a focus on lexical-semantic relations, and it may provide training data for relation discovery in word embeddings, potentially useful for augmenting Portuguese lexical-knowledge bases, such as Portuguese wordnets [4].

Similarly to the English BATS test [9], TALES covers different types of lexical-semantic relation, with the same number of entries, 50, for each, and is thus a balanced test. In this case, entries were selected, first, according to their presence in several lexical resources for Portuguese, and, second, to their frequency in a corpus. Since relations are not explicit in word embeddings, these models have to be further explored for solving analogy tests, and several methods have been proposed for this task [6].

In the remainder of this paper, we review some related work on available test sets for word embeddings, in English and Portuguese. We then describe the creation of TALES, including all the decisions taken in the process, and show examples of its contents. Before concluding, we report on the results of solving the lexical-semantic analogies of TALES with available word embeddings pre-trained for Portuguese [12], using various methods for this purpose. As it happens for the lexical-semantic relations in BATS, accuracies are low, though more challenging for some relations than for others, which opens the door to developing better relation discovery methods or learning better word embeddings. We also provide a brief error analysis which might be useful towards both of the latter, as well as for better tests.

2 Related Work

The quality of word embeddings is typically assessed with word similarity and analogy tests. The former contain pairs of words and a score proportional to their semantic similarity. Given two words, scoring their semantic similarity becomes a matter of computing the cosine of their vectors. The correlation between the computed scores for all the pairs in the test and the ground-truth scores, may then be measured for evaluation. The higher the correlation, the better the performance.

Popular tests of this kind, for English, include WordSim-353 [8] and SimLex-999 [13]. WordSim-353 contains 353 word pairs and their relatedness score (0-10), based on the judgement of 13 to 16 human judges. Due to the known differences between similarity and relatedness, WordSim-353 was later [1] manually split into similar and related pairs. For this purpose, semantic relations between the words of the pair were identified, and pairs were split into: similar (synonyms, antonyms, identical, or hyponym-hyperonym); related

[1] CISUC, DEI, University of Coimbra, Portugal hroliv@dei.uc.pt
[2] ISEC, Polytechnic Institute of Coimbra, Portugal a21220135@alunos.isec.pt
[3] CISUC & ISEC, Portugal ana@dei.uc.pt

(meronym-holonym); none of the previous relations but average similarity higher than 5; unrelated (remaining pairs). SimLex-999 contains 999 word pairs (666 noun-noun, 222 verb-verb, 111 adjective-adjective) and their similarity score, based on the opinion of ≈50 judges. This is the only test where judges were specifically instructed to differentiate between similarity and relatedness and rate regarding the former only (genuine similarity).

Regarding analogy solving, when presenting word2vec, evaluation used what became known as the Google Analogy Test (GAT) [18]. It has analogies of the kind a *is to* a^* *as* b *is to* b^*, split between nine syntactic (e.g., adjective to adverb, opposite, comparative, verb tenses) and five semantic categories (e.g., capital-country, currency, male-female), with 20-70 unique example pairs per category, which may be combined in 8,869 semantic and 10,675 syntactic questions.

BATS [9] is a broader alternative to GAT, balanced between four types of relation – grammatical inflections, word-formation, lexical-semantic and world-knowledge relations – , with 10 categories of each type and 50 word pairs per category (overall 2,000 unique word pairs). Experiments using BATS have shown that some categories are more challenging than others, and lexical-semantic relations are among those with lower accuracy. This also motivated the experimentation with alternative methods. DiffVec [27] is another dataset for evaluating word embeddings. It covers 15 relation categories, including both grammatical (8) and lexical-semantic relations (7), obtained from several sources. Specifically, lexical-semantic relations were obtained from SemEval-2012 task 2 [15] and from the BLESS dataset [2]. With 12,458 questions in total, it is larger than GAT and, although covering less categories, also larger than BATS, but imbalanced.

Performance on analogy tests is typically measured with accuracy, i.e., the proportion of answers that match the expected word. Though, some researchers also assessed this task in a retrieval or classification scenario, using measures like precision, recall, or Mean Average Precision (MAP) [3].

For assessing Portuguese word embeddings, some of the previous tests were translated to Portuguese [23], namely WordSim-353, SimLex-999 and GAT. Another related dataset is B^{2}SG [28], which targets semantic relations, but has a different structure, similar to the Test Of English as a Foreign Language (TOEFL), but based on the Portuguese part of BabelNet [19], and partially evaluated by humans. It contains frequent Portuguese nouns and verbs (target), each followed by four candidates, from which only one is related, and is organised in six files: two for synonymy, two for hypernymy, and two for antonymy, between nouns and between verbs, respectively.

3 The Creation of TALES

Our aim was to create a test set that could be used as other popular analogy test sets, but in Portuguese and focused on lexical-semantic relations. This section describes the most important decisions taken in the creation of this test, dubbed TALES, starting with the data format and target relations, and ending with decisions specifically concerning some of those relation types.

3.1 Data Format

We opted to represent TALES in a format similar to BATS, where included files have entries like those in figure 1. Specifically, for each relation, there would be a file where each row corresponds to an entry and has two-columns: one with a word, to be used in the formulation of a question (b), and another with one or more words, to be used as the target answers (b^*). We recall that an analogy can be formulated as 'what is to b as a^* is to a', for which the answer is b^*. Considering the BATS entries in figure 1, possible questions would be: *what is to cat as reptile is to rattlesnake?* (i.e., hypernym-of cat), or *what is to citrus as turtleneck is to sweater?* (i.e., hyponym-of citrus).

As it happens in BATS, but not in GAT, when there is more than one acceptable answer, they are all included. This is relevant, especially in the context of lexical-semantic relations. For instance, a hypernym should have several hyponymys, or an object might have several parts. Also, as in BATS, we split the test into different files, one for each relation. Each file has the same number of entries, 50, which means that it is balanced between all of the relations covered.

3.2 Target Relations

For selecting the types of relation to include, we initially targeted the more common types in wordnets, also in BATS [9], namely Hypernymy, Meronymy, Synonymy and Antonymy. We then looked at relations of those and other types in a large set of the relations extracted from ten lexical resources for Portuguese [10], and at the number of instances of each kind in more than one resource. The number of resources that a relation instance is found in, hereafter r, can be seen as an indicator of its consensus, utility and, indirectly, of its quality, i.e., given that most of the exploited resources had some automatic step in their creation, r can also be used for avoiding incorrect relations.

When looking at available relations and how they were organised, we first split synonymy in three types – Synonymy_n, between nouns, Synonymy_v, between, verbs, and Synonymy_adj, between adjectives – and Hypernymy between two – Hypernymy_n, between nouns, and Hypernymy_v, between verbs. We further decided to use Antonymy and Meronymy, though only one type of each: Antonymy between adjectives, for being the most representative, and Part-of for Meronymy, because it was the only type for which there were enough instances (see section 3.3). We finally found enough Purpose-of relation instances and included this type as well.

3.3 Instance Selection

Once target relations were set, we wanted to select the most consensual 50 instances of each target type. These would be the 50 instances of each type with highest r. Yet, in most cases there would be ties, i.e., more than 50 instances had the same r. So, we also ranked instances by the frequency of their first argument (first column, to be used as b) in CETEMPúblico [24], a Portuguese corpus of news. As corpus frequency is an indicator of the commonality / usage frequency of words, it is also relevant for selecting words to include. Therefore, we only considered instances where the first argument occurred at least 100 times in CETEMPúblico. After this, not enough Member-of and Material-of relations were left, which is the main reason for our test covering only Part-of, whereas BATS covers three types of Meronymy, the same as in WordNet [7]): Part, Member and Substance.

Despite being strict with the first relation argument, we dropped the frequency constraints for the second arguments (second column), which we recall could be more than one, and relaxed the r constraint for all but the first word. For the remaining words, the only constraint was that they occur in a relation of the target type with the first argument, in at least two resources ($r = 2$). Since some of the lexical resources considered included relations extracted from dictionaries, possibly not so common, and others were created automatically, setting $r = 2$ minimises the number of incorrect or unuseful relations.

```
cat          feline/beast/animal/organism/fauna/placental/ carnivore/chordate/felid/eutherian/mammal/...
rattlesnake  snake/reptile/pit_viper/serpent/ophidian
church       chapel/abbey/basilica/cathedral/duomo/kirk
citrus       lemon/orange/lime/mandarin/tangerine/yuzu
sweater      turtleneck/cardigan/pullover/slipover/turtle/polo-neck
```

Figure 1. Example entries in a BATS files for 4_Lexicographic_semantics (*L01 [hypernyms - animals]* and *L03 [hyponyms - misc]*).

3.4 Non-symmetrical relations

With initial experiments, we noticed that, in non-symmetrical relations (semantic), the challenge was different, depending on whether we were using direct (e.g., vehicle Hypernymy-of car) or inverse relations (car Hyponymy-of vehicle). This is mainly due to the fact that, in some directions, it is more common to have more than a single answer. As mentioned earlier, a hypernym will have several hyponymys, but a hyponym will often have a single (direct) hypernym. Or, something can be part of different things (e.g., blade part-of knife, axe, sower) or have different parts (e.g., parts of the body). Therefore, for each semantic relation, we created two different files, one with direct and another with inverse relations. In the latter, the order of the arguments was switched in the original relation set, which then went through the automatic creation process, including the application of the aforementioned constraints to the argument that was now the first. Since the switch was made in the original relation set, the instances in the file of direct relations are not necessarily the inverse of those in the direct.

3.5 Hypernymy and Concreteness

After Synonymy, Hypernymy_n is the second relation for which we had more instances, so we decided to further split them into more coherent sets. In BATS, there is a file for Hypernymy, another for its inverse, Hyponymy, and a third file for Hypernymy between animals only. For TALES, we did not create a file for a single class, but looked at another property of words: concreteness, i.e., the degree to which words refer to objects, persons, places, or things that can be experienced by the senses [20]. We further split the Hypernymy relations, direct and inverse, roughly into concrete (+concrete) and not concrete / abstract (-concrete). Concreteness values were obtained from the Minho Word Pool [25], where 3,800 Portuguese words have assigned values of concreteness and imageability, between 1 (minimum) and 7 (maximum). In this case, we empirically set that concrete words would have a minimum concreteness value of 6 (covering e.g., house, ball, money), whereas non-concrete would have 4.5 or less (covering e.g., age, space, energy). Again, to maximise the number of acceptable answers, this constraint was only applied to the first argument. Still, it is expectable that concrete concepts do relate with more concrete concepts and less concrete with less concrete.

3.6 Test Set Characterisation

Table 1 characterises TALES, the resulting test. It lists the relation types covered and their direction (D for direct, I, for inverse), the minimum r applied to the first-column argument, and examples of included relations, in Portuguese, with a rough English translation. As in BATS, for entries with more than one acceptable answer, the second argument has each possible answer split by '/'.

As nothing was done to avoid semantic ambiguity, it is common to mix different senses of the same word, some of them metaphorical. Yet, we do not see this as a problem. First, static word embeddings (e.g., word2vec, GloVe) also have a single vector per word, thus ignoring word senses. Second, in most cases, there are several acceptable answers, which might apply for different senses of the first argument. Such an example is the word *perna* (leg), for which four hypernyms are possible: *suporte/apoio*, related with the 'support' meaning, and *membro/segmento*, related to the 'limb' meaning.

4 Evaluation of Word Embeddings

TALES can be used for assessing Portuguese word embeddings, specifically, their ability to capture lexical-semantic relations. For demonstration purposes, we used three pre-trained models where four different methods were applied to solve TALES. Performed experiments, reported in this section, also provided useful insights on issues that might be fixed in the future. For loading the embeddings and performing the tests, we used the Vecto package[4], which supports analogy tests in the previously described BATS format, i.e., adopted by TALES.

4.1 Analogy Solving Methods

TALES was tackled with four analogy solving methods available in Vecto. For each method, Vecto outputs a report with information on each question, including a ranked list of candidate answers, a summary of the experimentation setup, and the accuracy of the test, computed from the first answer of each rank.

The first method, Similar-to-B (eq. 1), is often used for retrieving similar words, based on the cosine similarity of their vectors. Though not exactly an analogy-solving method, due to its simplicity, it has been used as a baseline [17] for this purpose. In fact, achieving the best accuracy with Similar-to-B means that more complex analogy solving methods are not doing any good.

$$b^* = \operatorname*{argmax}_{w \in V} \cos(b, w) \tag{1}$$

The second method, vector offset [18], was originally used for solving analogies with word2vec, and later became also known as 3CosAdd (eq. 2). It formulates the analogy as a *is to* a^* *as* b *is to* b^*, where b^* has to be inferred from a, a^* and b.

$$b^* = \operatorname*{argmax}_{w \in V} \cos(w, a^* - a + b) \tag{2}$$

The remaining two methods, both proposed by [6], try to make the most out of the full test set. 3CosAvg computes the average offset between words in position a and words in position a^*, in a set of relations of the target type (eq. 3). The answer, b^*, must maximise the cosine with the vector resulting from summing the average offset to b.

$$b^* = \operatorname*{argmax}_{w \in V} \cos(w, b + avg\_offset) \tag{3}$$

The final method for which we report results is LRCos (eq. 4). It considers the probability that a word w is of the same class as other words in position a^* as well as the similarity between w and b, measured with the cosine. A logistic regression is used for computing the likelihood of a word belonging to the class of words a^*.

$$b^* = \operatorname*{argmax}_{w \in V} P(w \in target\_class) * cos(w, b) \tag{4}$$

[4] https://github.com/vecto-ai

Relation		r	**Examples**
Synonym-of_n		7	(*local*, *sítio*) (*proposta*, *alvitre/sugestão/proposição*) (location, site), (proposal, suggestion/proposition)
Synonym-of_v		8	(*existir*, *viver/durar/...*) (*ouvir*, *perceber/entender/escutar/...*) (exist, live/last), (listen, feel/understand)
Synonym-of_adj		7	(*provisório*, *provisional/temporário*) (*rural*, *rústico/pastoril/...*) (provisional, temporary), (rural, rustic/pastoral)
Antonym-of_adj		5	(*estreito*, *largo*) (*velho*, *jovem/novo/moço*) (narrow, wide), (old, young/new/lad)
Hypernym-of_n (+concrete)	D	4	(*fruto*, *morango/ameixa/...*) (*veículo*, *jipe/monovolume/...*) (fruit, strawberry/plum), (vehicle, jeep/minivan)
	I	4	(*carro*, *veículo*) (*perna*, *suporte/segmento/membro/apoio*) (car, vehicle), (leg, support/segment/member)
Hypernym-of_n (-concrete)	D	4	(*regra*, *restrição/lei/etiqueta/...*) (*questão*, *pergunta/problema/...*) (rule, restriction/law/etiquette), (query, question/problem)
	I	4	(*futuro*, *tempo*) (*orgulho*, *satisfação/sentimento*) (future, time), (pride, satisfaction/feeling)
Hypernym-of_v	D	3	(*vir*, *chegar/desembarcar/cair*) (*contar*, *relatar/somar*) (come, arrive/land/fall), (count, report/sum)
	I	3	(*querer*, *ordenar/exigir*) (*pagar*, *subornar/dar/corromper*) (want, order/demand), (pay, bribe/give/pervert)
Part-of	D	2	(*mês*, *ano*) (*sala*, *casa/prédio/domicílio/edifício/habitação/...*) (month, year), (room, house/building/home)
	I	2	(*água*, *oxigénio/hidrogénio*) (*palavra*, *sílaba*) (water, oxygen/hydrogen), (word, syllable)
Purpose-of	D	3	(*levantar*, *guindaste*) (*desenhar*, *lapiseira/caneta/lápis/sombra/...*) (rise, crane), (draw, pencil/pen/shadow)
	I	3	(*lixa*, *polir*) (*fogão*, *aquecer/cozinhar*) (sandpaper, polish), (cooker, heat/cook)

Table 1. Characterisation of the generated lexical-semantic relations test.

We also experimented with other methods available for this purpose, namely 3CosMul and PairDirection [16], but concluded that they would not add much, and so left their results out of this paper. For instance, results of PairDirection were often 0 or very close.

4.2 Results

We tackled the challenge of solving the questions in TALES when the methods described in the previous section – Similar-to-B (SIM), 3CosAdd (3CAD), 3CosAvg (3CAV), LRCos (LRC) – were applied to three different models with 300 dimensions – GloVe, word2vec-CBOW and word2vec-SKIP-GRAM. All models are part of NILC embeddings [12], a set of pre-trained word embeddings for Portuguese, freely available for download[5]. We first look at the overall performance of different configurations, measured with the accuracy and MAP@10, and then at the performance per relation.

4.2.1 Overall Performance

Table 2 has the overall performance of each method+model configuration, considering all the 14 relations, only the lexical (synonymy and antonymy), and the semantic, in terms of accuracy and MAP@10. Given that TALES is balanced between the 14 relations, each in a different file with 50 entries, these are averages of the performance for each relation. Accuracy is given by the proportion of entries (b) for the given answer (b^*) was correct (i.e., it was one of the words in the second column of the entry for b). However, we recall that the figures for 3CosAdd imply not 50 but 2,450 questions (50×49), because they are based on averages of using each of the 50 entry pairs as $b : b^*$ when each of the remaining 49 entries is used as $a : a^*$. All methods were used with default parameters of the Vecto implementation. For instance, for LRCos, the logistic regression classifier was trained with 49 positive pairs (one from each entry, i.e., a and the first a^*, except the target one) and 49 negative pairs (each with two arguments from different entries, i.e., a is from an entry and a^* is from another, meaning that they are probably not related, at least not in as the positive examples).

Results show that TALES is a challenging test. Accuracies are way under the best figures for syntactic and semantic analogies using the same embeddings (i.e., between 40 and 60% [12]). Yet, a similar situation happens for English, on the BATS dataset [6], where best accuracies for lexical-semantic relations are always below 30%, with the single exception for the opposites with GloVe.

Considering all relations, four configurations are tied with best accuracy (13%), using different methods and models. One of them is the Similar-to-B baseline in word2vec-CBOW. The overall high accuracy of this baseline is mainly influenced by its performance for the symmetrical relations (lexical), where it achieved the best accuracy in word2vec-SKIP, tied with 3CosAvg. This is achieved because both Synonymy and Antonymy occur between similar concepts, for which this baseline is already a good estimation. Thus, for these lexical relations, benefits of using more sophisticated methods are residual, if any. On the other hand, accuracy of Similar-to-B is lower for non-symmetrical relations (semantic), where LRCos in GloVe achieves the best accuracy (13%). Though not very high, this is the only configuration with an average accuracy higher than 10% in this scenario. On a final note, the method originally applied for solving analogies in word2vec [18], 3CosAdd, is generally the one with worst performance, worse than Similar-to-B. This is also a consequence of how accuracy is computed for this method, which predicts b^* from a single pair $a : a^*$. Although this might work well for some relations, for the target ones, results show that it normally does not.

Although accuracy has been extensively used by others [18, 6], when tests have more than one acceptable answer, it makes sense to adopt metrics that look further than just the first given answer. This includes retrieval-based measures like precision and recall, with a threshold on the similarity score, or the Mean Average Precision (MAP) [3]. Since this is the case of TALES, towards a different perspective, we also computed the MAP@10.

For most relations, MAP is not significantly higher, suggesting

[5] http://nilc.icmc.usp.br/embeddings

	GloVe				word2vec-CBOW				word2vec-SKIP			
	SIM	**3CAD**	**3CAV**	**LRC**	**SIM**	**3CAD**	**3CAV**	**LRC**	**SIM**	**3CAD**	**3CAV**	**LRC**
Accuracy												
Lexical	0.20	0.08	0.20	0.15	0.22	0.10	**0.23**	0.16	**0.23**	0.10	0.22	0.15
Semantic	0.08	0.04	0.09	**0.13**	0.09	0.04	0.09	0.08	0.07	0.04	0.09	0.09
All	0.11	0.05	0.12	**0.13**	**0.13**	0.05	**0.13**	0.10	0.12	0.06	**0.13**	0.11
MAP@10												
Lexical	0.28	0.14	**0.29**	0.21	**0.29**	0.15	0.28	0.20	0.28	0.15	0.26	0.20
Semantic	0.16	0.08	0.16	**0.18**	0.11	0.06	0.12	0.11	0.12	0.07	0.13	0.13
All	0.19	0.10	**0.20**	0.19	0.16	0.09	0.17	0.14	0.16	0.09	0.17	0.15

Table 2. Performance different models and methods in TALES.

that not many correct answers are ranked between second and tenth. Nevertheless, MAP scores show more clearly that GloVe is the most consistent model for this kind of analogy. It always leads to the best MAP, though with different methods: 3CosAvg for the lexical relations and overall, and LRCos for the semantic. This is also consistent with related research for English [6, 27, 3], where GloVe is often used for this purpose, and the methods that use more instances (3CosAvg and LRCos) perform better than those that try to solve the analogy based on a single instance (3CosAdd) [6].

4.2.2 Per-Relation Performance

Table 3 presents the MAP@10 for each relation with each method+model configuration. Results make it clear that some relations pose different challenges than others. For instance, as expected, the Similar-to-B baseline outperformed all the other methods for Synonym-of, though with different models. Between nouns, the best MAP (0.27) was achieved in word2vec-CBOW, also between verbs (0.38), but tied with word2vec-SKIP. Between adjectives, the best MAP (0.28) was in GloVe, where 3CosAvg achieved the same result as Similar-to-B. For Antonym-of, the best configurations used the LRCos method (0.30) in both GloVe and word2vec-SKIP. This shows that, although Antonymy is also symmetrical, it behaves differently than synonymy, and might benefit from considering a larger set of relations.

For six out of 10 semantic relations, the best MAP was achieved by LRCos in GloVe, confirming that this configuration is a good choice for such relation types. This happened to the inverse of Hypernym-of, between concrete nouns (0.29), abstract nouns (0.16) and verbs (0.25), to Part-of (0.16), and Purpose-of, in both direct (0.15) and inverse direction (0.35).

Results of the latter configuration for Hypernymy-of are probably due to the higher difficulty of finding the hyponym given its hypernym, when compared to the other way round, mainly because a hypernym can have multiple hyponyms. Even though, in most cases, there was more than one acceptable answer, a list of hyponyms can be so extensive that several are probably missing (see section 4.3). For the direct Hypernym-of relation between both concrete and abstract nouns, the performance of the Similar-to-B baseline achieved the best MAP, also with GloVe. This is not surprising, not only due to the previous reason, but also because hyponyms are very similar to their hypernyms, only more specific.

A curious situation was that, in opposition to LRCos in GloVe, for 3CosAvg in word2vec-CBOW the performance for direct Hypernym-of_n relations was higher than for the inverse. This again suggests that different configurations are better suited for different goals.

Still on Hypernym-o_n, performance is generally better when it is between concrete concepts than for those more abstract. This should be due to the nature of abstract nouns, with which one cannot interact directly, making it also difficult to generalise the contexts they occur in. A similar cause might help explaining the results for Hypernym-of_v, for which the highest MAP was achieved by Similar-to-B in word2vec-SKIP (0.22).

For Part-of, performance was the poorest. In LRCos+GloVe, the direct relations (0.16) got twice the MAP of the inverse (0.08), but 3CosAvg+GloVe had a slightly higher MAP (0.10). Similarly to Hypernym-of, this might be affected by the fact that an object might have several parts and it may be a part of different objects. Yet, in this case, the low MAP is also due to other issues (see section 4.3).

On the other hand, one of the highest MAPs in the test was achieved for Purpose-of in the inverse direction (0.35). Not only its accuracy was high with the LRCos+GloVe configuration, but it was also considerably higher than the baselines, and contrasting with the lower performance in word2vec-CBOW. This suggests that, although not included in similar tests for English, the Used-For relation (inverse of Purpose-of) suits this kind of test well.

4.3 Brief Error Analysis

For better insights on the achieved performance and typical issues, we inspected the results of two different methods for two different relations. Some of the identified issues might be addressed in future versions of TALES, while others will hopefully contribute to the development of better methods for relation discovery, or even models that capture these relations better.

First, issues concerning the inverse Hypernym-of relation, recalling that a concept might have a huge number of hyponyms. Although TALES includes five types of *escola* (school), it does not cover others given by LRCos+GloVe as an answer, namely *preparatória* (preparatory), *conservatório* (conservatory), *secundária* (secondary) or *liceu* (high school). This happens because none of the aforementioned connections is in any of the lexical resources used as the source of TALES. In fact, some of them are often used as modifiers of *escola*, often appearing together (e.g., *escola preparatória* or *escola secundária*), with the "simple" version not covered by the lexical resources. Another example is the word *jornal* (newspaper), for which the first answer was *semanário* (weekly newspaper), not accepted because, despite being correct, the instance *jornal* Hypernym-of *semanário* was found in a single lexical resource, and thus not included in TALES. Other issues are related to the presence of world-knowledge, much of which not included in dictionaries and lexical knowledge bases. This happens, for instance, for the word *moeda* (currency), with the first answer 'ecu', the former European currency, precursor of the euro, not in the source lexical resources. The word 'euro' came in second, but is also not in TALES, because it was in a single lexical resource. A second example of this kind occurred for *automóvel* (car), for which many answers were brands of cars, starting with *fiat*, followed by *volkswagen* (rank #4), *renault* (#5), *bmw* (#6) and *audi* (7).

We also inspected the results of GloVe+3CosAvg for Part-of I (Has-Part), for which MAP was very low. We came to the con-

Relation		GloVe				word2vec-CBOW				word2vec-SKIP			
		SIM	**3CAD**	**3CAV**	**LRC**	**SIM**	**3CAD**	**3CAV**	**LRC**	**SIM**	**3CAD**	**3CAV**	**LRC**
Synonym-of_n		0.23	0.12	0.26	0.13	**0.27**	0.15	0.26	0.09	0.26	0.14	0.26	0.13
Synonym-of_v		0.36	0.15	0.34	0.28	**0.38**	0.19	0.36	0.28	**0.38**	0.18	0.34	0.24
Synonym-of_adj		**0.28**	0.11	**0.28**	0.14	0.26	0.13	0.26	0.17	0.25	0.10	0.21	0.12
Antonym-of_adj		0.26	0.16	0.27	**0.30**	0.25	0.14	0.26	0.27	0.23	0.17	0.25	**0.30**
Hypernym-of_n	D	**0.24**	0.07	0.20	0.08	0.23	0.08	0.23	0.06	0.19	0.07	0.18	0.07
(+concrete)	I	0.18	0.15	0.25	**0.29**	0.15	0.09	0.20	0.19	0.14	0.09	0.16	0.22
Hypernym-of_n	D	**0.23**	0.07	0.19	0.14	0.20	0.08	0.20	0.08	0.21	0.08	0.19	0.15
(-concrete)	I	0.11	0.07	0.10	**0.16**	0.07	0.04	0.06	0.08	0.10	0.07	0.11	0.12
Hypernymy-of_v	D	0.17	0.09	0.14	0.16	0.20	0.12	0.17	0.21	**0.22**	0.11	0.19	0.11
	I	0.21	0.12	0.16	**0.25**	0.20	0.13	0.20	0.24	0.22	0.12	0.20	0.21
Part-of	D	0.10	0.05	0.09	**0.16**	0.02	0.02	0.02	0.04	0.02	0.02	0.02	0.03
	I	0.09	0.05	**0.10**	0.08	0.05	0.04	0.05	0.02	0.05	0.04	0.05	0.01
Purpose-of	D	0.13	0.05	**0.15**	**0.15**	0.00	0.00	0.03	0.08	0.02	0.02	0.05	**0.15**
	I	0.11	0.13	0.25	**0.35**	0.00	0.02	0.06	0.10	0.02	0.06	0.15	0.18

Table 3. MAP@10 for different relations, with different models and methods.

clusion that the test for this relation includes several difficult entries, some of which with multiple senses, some of which significantly different, like *ser* (to be / living being), *câmara* (camera, chamber), or *programa* (program, show); and others that refer to vague concepts, like *todo* (whole), *mundo* (world), *espaço* (space), *organização* (organization), *vida* (life) or *coisa* (thing). While the issue of ambiguity is minimised by the presence of several acceptable answers, what might increase the difficulty is that ambiguous words are used in different contexts, making the relations less obvious in the geometric space. Vagueness could possibly be minimised if, as we did for Hypernym-of, we split concrete and abstract nouns, but available Part-of instances are not as many.

Though not necessarily more difficult, the Part-of inverse test also covers several time-related words, like *minuto* (minute), *hora* (hour), *dia* (day), *semana* (week), *mês* (month), or *ano* (year). Moreover, on the many incorrect answers, the main issues noted were:

- Confusion between the direct and inverse relation, i.e., some answers were not the parts, but the whole of *b*, e.g., for *dia* (day), answers included *semana* (week) and *mês* (month); for *palavra* (word), *expressão* (expression) and *frase* (sentence); or, for *texto* (text), *documento* (document) and *comentário* (comment).
- Confusion with hyponymy, i.e., some answers were hyponyms of *b*, e.g., for *homem* (man), answers included *rapaz* (boy), *jovem* (young) and *garoto* (kid); for *casa* (house), *apartamento* (apartment) and *mansão* (mansion); or, for *mês* (month), names of months, like *abril* (April), *maio* (May), *março* (March) and *fevereiro* (February).
- Most plural forms are not covered by the lexical resources and are thus not in TALES. Yet, some answers were in the plural form, often making sense, e.g., *segundos* (seconds) for *minuto* (minute); *minutos* (minutes) for *hora* (hour); or *alunos* (students) for *escola* (school).
- Correct answers that are not in TALES, e.g., *madrugada* (dawn) for *noite* (night); *texto* (text) for *documento* (document); *cérebro* (brain) and *genoma* (genome) for *humano* (human); or *porta* (door) and *mesa* (table) for *sala* (room). Not all are the most obvious relations and none was in any of the exploited lexical resources, but they could probably be in TALES. In parallel, they could be seen as suggestions for augmenting the aforementioned lexical resources.

5 Concluding Remarks

We have presented TALES, a new test for assessing Portuguese word embeddings in the domain of lexical-semantic relations, i.e., how well are such relations captured by the embeddings. Decisions taken in the creation of this test were first explained. Then, methods commonly used for solving analogy tests were used with pre-trained word embeddings for Portuguese to answer the questions in TALES, which lead to some conclusions, here discussed.

TALES is available from `https://github.com/hgoliv/PT-LexicalSemantics`. As we have shown, it is a challenging test, for which high performances will require better methods or models of word embeddings. Interested researchers may want to assess other models for Portuguese, such as Numberbatch [26], for which we have recently performed initial experiments with the highest performances achieved. However, it might be unfair to compare Numberbatch with embeddings learned exclusively from text, because the creation of the former also considered the structure of ConceptNet, which includes some well-structured lexical-semantic knowledge. It would also be interesting to test more recent pre-trained language models, also known as contextual embeddings, like ELMo [22] and BERT [5]. Yet, we are unsure whether we can take advantage of the contextual features of the previous because the entries of TALES lack context and do not handle different senses of the same word.

Other analogy solving methods may as well be tested. As mentioned earlier, 3CosMul and PairDistance [16] were far from outperforming the reported results. But there are other promising methods proposed recently, such as the Translation and the Regression Model [3]. Finally, TALES may be used for training models of relation discovery in word embeddings, which may then be used for augmenting existing Portuguese lexical knowledge-bases [4].

In the future, we might as well look into some of the issues noted in our error analysis and create an improved version of TALES. For instance, we may consider including also instances that occur in a single lexical resource, especially if at least one relation per line is in more. Besides Hypernym-Of, we may split other relations according to the concreteness of their arguments. Yet, for most relation types, this might result in less than 50 instances. Other possible gains will probably require to use a lexical knowledge base with sense information, such as a wordnet. In fact, relying on a single knowledge base, ideally a manually-curated one, may avoid possible inconsistencies regarding organisation and other criteria (e.g., definitions adopted for each semantic relation). We may also consider removing *b* words with many senses, or increase the number of hypernyms and hyponyms by expanding the hypernym taxonomy. In fact, something similar was done for BATS [9]. We will also look at other performance metrics, possibly looking at the average ranking of the correct answer(s) or the distance or similarity of the given answer(s) to the correct.

REFERENCES

[1] Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa, 'A study on similarity and relatedness using distributional and WordNet-based approaches', in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pp. 19–27. ACL, (2009).

[2] Marco Baroni and Alessandro Lenci, 'How we BLESSed distributional semantic evaluation', in *Proceedings of GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pp. 1–10, Edinburgh, UK, (2011). ACL.

[3] Zied Bouraoui, Shoaib Jameel, and Steven Schockaert, 'Relation induction in word embeddings revisited', in *Proceedings of the 27th International Conference on Computational Linguistics*, COLING 2018, pp. 1627–1637, Santa Fe, New Mexico, USA, (August 2018). ACL.

[4] Valeria de Paiva, Livy Real, Hugo Gonçalo Oliveira, Alexandre Rademaker, Cláudia Freitas, and Alberto Simões, 'An overview of Portuguese wordnets', in *Proceedings of 8th Global WordNet Conference*, GWC'16, pp. 74–81, Bucharest, Romania, (2016).

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, 'BERT: Pre-training of deep bidirectional transformers for language understanding', in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, NAACL-HLT 2019, pp. 4171–4186. ACL, (2019).

[6] Aleksandr Drozd, Anna Gladkova, and Satoshi Matsuoka, 'Word embeddings, analogies, and machine learning: Beyond king - man + woman = queen', in *Proceedings the 26th International Conference on Computational Linguistics: Technical papers (COLING 2016)*, COLING 2016, pp. 3519–3530, (2016).

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

[8] Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin, 'Placing search in context: The concept revisited', *ACM Trans. Inf. Syst.*, **20**(1), 116–131, (January 2002).

[9] Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka, 'Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't.', in *Proceedings of the NAACL 2016 Student Research Workshop*, pp. 8–15. ACL, (2016).

[10] Hugo Gonçalo Oliveira, 'A survey on Portuguese lexical knowledge bases: Contents, comparison and combination', *Information*, **9**(2), (2018).

[11] Zelig Harris, 'Distributional structure', *Word*, **10**(2-3), 1456–1162, (1954).

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

[13] Felix Hill, Roi Reichart, and Anna Korhonen, 'Simlex-999: Evaluating semantic models with genuine similarity estimation', *Computational Linguistics*, **41**(4), 665–695, (2015).

[14] Graeme Hirst, 'Ontology and the lexicon', in *Handbook on Ontologies*, eds., Steffen Staab and Rudi Studer, International Handbooks on Information Systems, 209–230, Springer, (2004).

[15] David Jurgens, Saif Mohammad, Peter Turney, and Keith Holyoak, 'SemEval-2012 task 2: Measuring degrees of relational similarity', in **SEM 2012: The 1st Joint Conference on Lexical and Computational Semantics – Vol 1: Proc. of main conference and shared task, Vol 2: Proc. of 6th (SemEval 2012)*, pp. 356–364. ACL, (2012).

[16] Omer Levy and Yoav Goldberg, 'Linguistic regularities in sparse and explicit word representations', in *Proceedings of 18th Conference on Computational Natural Language Learning*, CoNLL 2014, pp. 171–180. ACL, (2014).

[17] Tal Linzen, 'Issues in evaluating semantic spaces using word analogies', in *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pp. 13–18, Berlin, Germany, (August 2016). ACL.

[18] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, 'Efficient estimation of word representations in vector space', in *Proceedings of the Workshop track of ICLR*, (2013).

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

[20] A. Paivio, J. C. Yuille, and S. A. Madigan, 'Concreteness, imagery, and meaningfulness values for 925 nouns', *Journal of Experimental Psychology monograph supplement*, **76**(1), 1–25, (1968).

[21] Jeffrey Pennington, Richard Socher, and Christopher D. Manning, 'GloVe: Global vectors for word representation', in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, EMNLP 2014, pp. 1532–1543. ACL, (2014).

[22] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer, 'Deep contextualized word representations', in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 2227–2237. ACL, (2018).

[23] Andreia Querido, Rita Carvalho, João Rodrigues, Marcos Garcia, João Silva, Catarina Correia, Nuno Rendeiro, Rita Pereira, Marisa Campos, and António Branco, '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, (2017).

[24] Paulo Alexandre Rocha and Diana Santos, 'CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa', in *V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR 2000)*, ed., Maria das Graças Volpe Nunes, pp. 131–140, São Paulo, (19-22 de Novembro 2000). ICMC/USP.

[25] Ana Paula Soares, Ana Santos Costa, João Machado, Montserrat Comesaña, and Helena Mendes Oliveira, 'The Minho Word Pool: Norms for imageability, concreteness, and subjective frequency for 3,800 Portuguese words', *Behavior Research Methods*, **49**(3), 1065––1081, (2017).

[26] Robert Speer, Joshua Chin, and Catherine Havasi, 'Conceptnet 5.5: An open multilingual graph of general knowledge', in *Proceedings of Thirty-First Conference on Artificial Intelligence (AAAI)*, pp. 4444–4451, (2017).

[27] Ekaterina Vylomova, Laura Rimell, Trevor Cohn, and Timothy Baldwin, 'Take and took, gaggle and goose, book and read: Evaluating the utility of vector differences for lexical relation learning', in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2016, pp. 1671–1682. ACL, (2016).

[28] Rodrigo Wilkens, Leonardo Zilio, Eduardo Ferreira, and Aline Villavicencio, 'B^{2}SG: a TOEFL-like task for Portuguese', in *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, LREC 2016, Paris, France, (2016). ELRA.