Assessing Lexical-Semantic Regularities in Portuguese Word Embeddings

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

[1] CISUC, Department of Informatics Engineering, Faculty of Sciences and Technology, University of Coimbra (Portugal)
[2] ISEC, Polytechnic Institute of Coimbra (Portugal)
[3] CISUC & ISEC, Polytechnic Institute of Coimbra (Portugal)

Received 23 November 2020 | Accepted 28 December 2020 | Published 19 February 2021

unir
LA UNIVERSIDAD
EN INTERNET

Abstract

Models of word embeddings 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 by some analogy test sets for English. Briefly, this paper aims to study how well pretrained Portuguese word embeddings capture such relations. For this purpose, we created a new test, dubbed TALES, with an exclusive focus on Portuguese lexical-semantic relations, acquired from lexical resources. With TALES, we analyse the performance of methods previously used for solving analogies, on different models of Portuguese word embeddings. Accuracies were clearly below the state of the art in analogies of other kinds, which shows that TALES is a challenging test, mainly due to the nature of lexical-semantic relations, i.e., there are many instances sharing the same argument, thus allowing for several correct answers, sometimes too many to be all included in the dataset. We further inspect the results of the best performing combination of method and model to find that some acceptable answers had been considered incorrect. This was mainly due to the lack of coverage by the source lexical resources and suggests that word embeddings may be a useful source of information for enriching those resources, something we also discuss.

Keywords

Natural Language Processing, Computational Semantics, Word Embeddings, Lexical Semantics, Analogy.

DOI: 10.9781/ijimai.2021.02.006

I. Introduction

Two main approaches have been followed for representing the words of a language according to their semantics: lexical-semantic knowledge bases (LKBs), such as wordnets [1]; and distributional models, like word embeddings. The former organise words and their meanings, often connected by explicit relations, such as Hypernymy or Part-of, and may include additional lexicographic information (part-of-speech, gloss). On the other hand, the latter follow the distributional hypothesis [2], which says that words that occur in the same contexts tend to convey similar meanings, 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 [3] or GloVe [4]. Besides their utility for computing word similarity, e.g., with the cosine similarity of the vector representations, such models are known for preserving several linguistic regularities, and have shown very interesting results when 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, from word inflections and derivations, to word knowledge relations like capital-country. Yet, we are interested in studying relations between word meanings that one would find, implicitly, in a language dictionary or, explicitly, in a LKB. Given that they connect general-language words according to their meanings, we refer to them as lexical-semantic relations. More precisely, our goal with this work is twofold. We aim to:

- Assess how lexical-semantic relations are preserved by Portuguese word embeddings;
- Analyse to what extent analogy solving methods could be useful for enriching LKBs.

Towards our goal, we needed an analogy test targeting lexical-semantic relations in Portuguese, which we created as described in this paper. It was baptised as *Teste para Analogias Léxico-Semânticas* (TALES, in English, Test for Lexical-Semantic Analogies) and is exclusively focused on these relations. Although some analogy tests already cover lexical-semantic relations [5], [6], they are for English and, while they could have been translated to Portuguese, as the Google Analogy Test was [7], we decided to create a new 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 [8]. This is important because, besides assessing word embeddings, TALES can provide training data for relation discovery in word embeddings, potentially useful for augmenting Portuguese LKBs, such as Portuguese wordnets [9].

* Corresponding author.

E-mail address: hroliv@dei.uc.pt

TALES follows a similar format to the English BATS test [5] and covers different types of lexical-semantic relation, with the same number of entries, 50, for each, which makes it a balanced test. The entries of TALES were selected based on their presence in several Portuguese lexical resources and their frequency in a corpus. We attempted at solving the lexical-semantic analogies of TALES by applying classic and more recent analogy solving methods [10] to pretrained word embeddings available for Portuguese. This included static word embeddings (word2vec and GloVe) but also static representations obtained from recent BERT [11] neural language models. As it happens for the lexical-semantic relations in BATS, accuracies are low, even if some relations are more challenging than others. However, in opposition to some relation types, namely syntactic and world knowledge relations, several entries in TALES have many acceptable answers (e.g., a hypernym generally has several hyponyms). And even though the adopted BATS format enables the inclusion of several answers, in many cases they are too many and it is just not possible to get them all from the lexical resources. Therefore, incorrect answers may include relations that are just missing, which makes word embeddings, potentially, a useful source of information for enriching those resources. Having this in mind, we analyse some of the results obtained and discuss this possibility. Indeed, missing examples were found for every relation type covered by TALES. Even if, for some types, these were a minority of cases, for others they represented almost 20%of the answers considered incorrect.

The remainder of the paper is structured as follows: next Section (II) overviews related work on available test sets for assessing word embeddings, in English and Portuguese, as well as some work on the automatic creation and enrichment of LKBs; Section III describes the creation of TALES, including all the decisions taken in the process, and shows examples of its contents; Section IV describes the models of pretrained word embeddings used and the analogy solving methods applied in our experimentation; Section V reports on the performance of solving the lexical-semantic analogies of TALES with word embeddings, using the different methods, also looking at the performance per-relation; before concluding, Section VI starts by analysing incorrect answers that would be acceptable and discusses the utility of word embeddings when it comes to enriching structured lexical resources.

II. Related Work

The quality of word embeddings is typically assessed with two kinds of test: word similarity and analogy. 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, i.e., the higher the correlation, the better the performance.

Popular tests of this kind, for English, include WordSim-353 [12] 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 [14] 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 (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. Its authors thus claim that it targets genuine similarity.

Anyway, despite the post-annotations in WordSim-353, relatedness scores are only a number that represents a strength, but tells nothing about the actual relation between the words or concepts they denote. To go further in distributional models, one may resort to analogies, i.e., look for pairs of words that are related similarly to a known pair of words or a set of pairs. When presenting word2vec, evaluation used what became known as the Google Analogy Test (GAT) [3]. 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 [5] 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). Moreover, BATS enables more than one possible answer for each question, which makes sense for some relation types (e.g., a hypernym will have more than one hyponym). The lexical-semantic relations in BATS were acquired from Princeton WordNet [1], a LKB where word senses are grouped in synonym sets, and semantic relations are established between the latter.

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 that consider more than one example for solving analogies, namely 3CosAvg and LRCos (see sub-section B of Section V).

DiffVec [6] 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 [16]. 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 [17], i.e., quantifying how many correct answers could be retrieved. For this purpose, measures like precision, recall, or Mean Average Precision (MAP) were used.

For assessing Portuguese word embeddings, some of the previous tests were translated to Portuguese [7], namely WordSim-353, SimLex-999 and GAT. Several approaches were tested for answering WordSim-353 and SimLex-999 [18], including knowledge and distributional approaches. GAT has been used for assessing Portuguese Word Embeddings [19] and, more recently, was translated to the BATS format [20], which enabled the application of alternative methods for analogy solving.

Another related dataset for Portuguese is B^{2}SG [21], which targets semantic relations, but has a different structure. It is similar to the Test Of English as a Foreign Language (TOEFL), but based on the Portuguese part of BabelNet [22], and was partially evaluated by humans. B^{2}SG 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. An important difference to the analogy tests is that B^{2}SG narrows the possible answers to the four candidates.

Back to the analogy tests, we believe that, besides assessing word embeddings, they can be useful for developing models of relation discovery in the embedding space, especially considering lexical-semantic analogies, which often have more than one acceptable answer. More precisely, models trained in analogy tests could be useful for creating or enriching knowledge bases. The goal would be similar to earlier attempts for extracting relations from dictionaries [23], or from raw corpora, having in mind the enrichment of LKBs like WordNet [1], and tackled with handcrafted patterns [24], or patterns learned with weakly-supervised approaches, for extracting hypernymy [25] and other relations . The latter approaches would start with known seeds, which could be acquired from WordNet itself. An alternative way of enriching LKBs, which are focused on lexical knowledge, is to extend them with world knowledge, e.g., by linking them with Wikipedia, as in the BabelNet project [22]. For Portuguese, on this scope, Onto.PT is a wordnet [27] that combines information in existing thesauri with relations extracted from several Portuguese dictionaries [28].

III. Creating the TALES Test Set

In order to assess to what extent lexical-semantic relations are preserved in Portuguese word embeddings, we first needed a benchmark. For this purpose, we created a test set, dubbed TALES, that could be used in a similar way to other popular analogy test sets. This section describes the most important decisions taken in the creation of this test, starting with the adopted data format, target relations, and ending with decisions specifically concerning some relation types.

A. Data Format

We opted to represent TALES in a format similar to BATS, where included files have entries like those in Fig. 1. Specifically, for each covered 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 Fig. 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). We also note that, besides direct relations, BATS includes inherited relations in the possible answers, such as the inherited hypernyms in the first entries in Fig. 1.

As it happens in BATS, but not in GAT, when there is more than one possible answer, they are all included in the second column, split by '/'. 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 covered. Each file has the same number of entries, 50, which means that TALES is balanced between all of the relations covered.

B. Target Relations

For selecting the relation types to include in TALES, we initially targeted the more common types in wordnets, also included in BATS [5], namely Hypernymy, Meronymy, Synonymy and Antonymy. We then looked at relations of those and other types in a large set of relations extracted from ten lexical resources for Portuguese [29], covering both the European and the Brazilian variant[1], 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 decided to split synonymy in three types – Synonymy_n, between nouns, Synonymy_v, between verbs, and Synonymy_adj, between adjectives – and Hypernymy in 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 sub-section C). Finally, we also found enough Purpose-of relation instances and decided to included this type as well.

C. Instance Selection

Once we had decided on target relations, we wanted to select the most consensual 50 instances of each selected 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 [38], 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[2]. 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 [1]: Part, Member and Substance.

Despite being strict with the first relation argument, we dropped the frequency constraints for the second argument (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

[1] These resources were PAPEL [28], Dicionário Aberto [30], Wiktionary.PT [31], TeP [32], OpenThesaurus.PT, OpenWordNet-PT [33], PULO [34], WordNet.Br [35], Port4Nooj [36] and ConceptNet [37].

[2] CETEMPúblico was used only for ranking and filtering, based on the first argument of each relation instance, while all words still came from the lexical resources. In fact, we did not use CETEMPúblico directly, only the frequency lists available from AC/DC [39].

```
cat           feline/beast/animal/organism/fauna/placental/ carnivore/chordate/felid/eutherian/mammal/mammalian/...

hawk          raptor/bird/vertebrate/creature/beast/being/animal/organism/fauna/chordate/animate_being/craniate/...

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
```

Fig. 1. Example entries in a BATS files for 4_Lexicographic_semantics, namely *L01 [hypernyms - animals]* (first three lines) and *L03 [hyponyms - misc]* (last three lines).

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. At the same time, this may contribute to lower Mean Average Precision with some models (see examples in Section VI).

D. Non-symmetrical Relations

With initial experiments, we noticed that, in non-symmetrical relations, 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 many possible answers. As mentioned earlier, a hypernym will have several hyponyms, 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 non-symmetrical 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 then became 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.

E. 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 [40]. So, we 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 [41], 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 abstract 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 concepts.

F. Test Set Characterisation

Table I characterises TALES, the resulting test. It lists the relation types covered and their direction (D for direct, I, for inverse), the minimum *r* (higher for relations for which there were more instances) 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 '/'.

TABLE I. Characterisation of the Generated Lexical-semantic Relations Test

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)

As nothing was done to avoid semantic ambiguity, it is common to mix different senses of the same word, some of them figurative. 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.

IV. Experimentation Setup

TALES can be used for assessing Portuguese word embeddings, specifically, their ability to capture lexical-semantic relations. For this purpose, we first used three pretrained models of static word embeddings for Portuguese, where four methods were applied to solve TALES. Moreover, to embrace recent trends, we decided to test as well embeddings produced by pretrained neural language models, namely BERT [11]. For loading the embeddings and performing the tests, we used the Vecto[3] package, which supports analogy tests in the previously described BATS format, adopted by TALES, and includes the implementation of different analogy solving methods. This section describes the models and methods used in our experimentation in more detail.

A. Models of Word Embeddings

The analogy solving methods were first applied to three pretrained models of static word embeddings, all with 300 dimensions, but covering different learning algorithms, namely GloVe, word2vec CBOW and word2vec SKIP-GRAM. These models are part of NILC embeddings [19], a set of pretrained word embeddings for Portuguese, freely available for download[4].

However, the current trend in language representation are neural language models, like BERT [11], which rely on Transformer neural networks for encoding words and longer sequences in meaningful embedding vectors. An important difference towards static word embeddings is that BERT provides contextual word representations, meaning that, depending on its surrounding context, the same word might be represented by different vectors. Since, like any analogy test, the entries of TALES lack context and do not handle different senses of the same word, we were unsure whether we could take advantage of the contextual features of the previous models. Yet, recent work has showed that contextualised representations in a given layer (apparently, the lower, the better) can outperform static word embeddings in analogy solving [42]. Therefore, we decided to apply the analogy solving methods also to word representations by two BERT models covering Portuguese, namely: the Multilingual Cased BERT-Base model by Google[5], which includes Portuguese among 104 languages; and BERTimbau-Large, pretrained exclusively for Portuguese [43]. The former has 12 layers and encodes word sequences in 768-size vectors, the size of its hidden layers, while the version of BERTimbau used has 24 layers and encodes word sequences in 1,024-sized vectors.

Since we were dependent on Vecto for running the tests, and Vecto is only prepared to deal with static word embeddings, we had to convert BERT representations to a static format. With the help of the bert-as-a-service[6] tool, this conversion was made in three steps: (i) running through all the entries in the vocabulary of each BERT model, which includes words and subwords (word pieces); (ii) retrieve their representation in a selected layer of the model; (iii) use the resulting vector as the static representation of the vocabulary entry. BERT provides contextualised representations, but there is no context in the questions of analogy tests, therefore, we simply tested representations obtained from different layers and present results for: the first and the second, which should be less context-specific [42], [44]; and also the second to last, because the last layer is too close to the target functions during pretraining and may be biased to those targets. We also note that, with this adaptation, we might not be taking the most out of BERT. The main issue is related to the vocabulary coverage, which we are limiting to the entries in BERT's vocabulary file – 119,547 for BERT-ML and 29,794 for BERTimbau –, when we know that BERT relies on the WordPiece tokeniser and represents several words with the combination of two or more entries (subwords). At the same time, many subwords are used for obtaining inflections, which are scarce in the target lexical resources and thus in TALES. We leave alternative approaches on handling BERT for consideration in future work.

B. Analogy Solving Methods

In order to solve TALES, four different methods were applied to the selected models of word embeddings, all with implementation available in Vecto. For each method, Vecto outputs a JSON 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 [45] 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 [3], 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*.

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

Instead of considering only the word *b* (Similar-to-B) or this word plus a single pair of analogously-related words (*a*, a^*), the remaining two methods, both proposed by Drozd et al. [10], try to make the most out of the full test set. 3CosAvg computes the average offset between words in position *a* and respective 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 tested is LRCos (eq. 4), which considers the probability that a word *w* belongs to the same class as other words in position a^*, as well as the similarity between *w* and *b*, measured with the cosine. Although any classification algorithm could be used for this, the default implementation of LRCos, used by us, relies on logistic regression for computing the likelihood of a word belonging to the class of words a^*.

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

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

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

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

[5] https://github.com/google-research/bert

[6] https://github.com/hanxiao/bert-as-service

V. Experimentation Results

We tackled the challenge of solving the questions in TALES by applying the four methods described in sub-section B of Section IV — Similar-to-B (SIM), 3CosAdd (3CAD), 3CosAvg (3CAV), LRCos (LRC) — to the five models of word embeddings introduced earlier — GloVe, word2vec CBOW, word2vec SKIP-GRAM, BERT-ML, BERTimbau.

This section reports on the results of this experimentation. Besides revealing the accuracy of different methods in different models of embeddings, for different relations, performed experiments provide useful insights on the potential of this approach for discovering new relations, which is further discussed in section VI. To help us reach some conclusions, we first look at the overall performance of different configurations, measured with the accuracy and MAP@10, and then at the performance per relation.

A. Overall Accuracy

Table II has the overall performance of each method in the static word embeddings, considering all the 14 relations, only the symmetrical (synonymy and antonymy), and only the non-symmetrical, in terms of accuracy and MAP@10. Tables III and IV have similar information, respectively for the representations obtained from three different layers of the two BERT models used.

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 which the first answer given (b^*) was correct (i.e., it was one of the words in the second column of the entry for b). The MAP@10 considered not only the first answer, but the top-10 answers given by Vecto for each question.

TABLE II. Performance of Static Word Embedding Models Through Different Methods in TALES

	GloVe				**word2vec-CBOW**				**word2vec-SKIP**			
	SIM	**3CAD**	**3CAV**	**LRC**	**SIM**	**3CAD**	**3CAV**	**LRC**	**SIM**	**3CAD**	**3CAV**	**LRC**
					Accuracy							
Symmetrical	0.19	0.08	0.19	0.14	**0.22**	0.10	**0.22**	0.15	**0.22**	0.10	**0.22**	0.14
Non-Symmetrical	0.07	0.04	0.08	**0.12**	0.08	0.04	0.09	0.07	0.07	0.04	0.09	0.09
All	0.10	0.05	0.11	**0.13**	0.12	0.05	0.12	0.10	0.11	0.06	0.12	0.10
					MAP@10							
Symmetrical	0.28	0.14	**0.29**	0.21	**0.29**	0.15	0.28	0.20	0.28	0.15	0.26	0.20
Non-Symmetrical	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 III. Performance of Word Embeddings From Different Layers of BERT-ML Through Different Methods in TALES

	Layer 1				**Layer 2**				**Layer 11**			
	SIM	**3CAD**	**3CAV**	**LRC**	**SIM**	**3CAD**	**3CAV**	**LRC**	**SIM**	**3CAD**	**3CAV**	**LRC**
					Accuracy							
Symmetrical	0.02	0.00	0.02	0.01	0.02	0.00	0.02	0.01	0.02	0.00	0.02	0.01
Non-Symmetrical	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.02
All	0.01	0.00	0.01	0.02	0.01	0.00	0.01	0.02	0.01	0.00	0.01	0.02
					MAP@10							
Symmetrical	0.03	0.00	0.03	0.02	0.03	0.00	0.02	0.01	0.03	0.00	0.03	0.01
Non-Symmetrical	0.01	0.00	0.01	0.04	0.01	0.00	0.01	0.04	0.01	0.00	0.01	0.03
All	0.02	0.00	0.02	0.03	0.02	0.00	0.02	0.03	0.01	0.00	0.02	0.02

TABLE IV. Performance of Word Embeddings From Different Layers of BERTimbau Through Different Methods in TALE.

	Layer 1				**Layer 2**				**Layer 23**			
	SIM	**3CAD**	**3CAV**	**LRC**	**SIM**	**3CAD**	**3CAV**	**LRC**	**SIM**	**3CAD**	**3CAV**	**LRC**
					Accuracy							
Symmetrical	0.03	0.00	0.03	0.03	0.03	0.01	0.03	0.02	0.05	0.01	0.05	0.03
Non-Symmetrical	0.00	0.00	0.01	0.05	0.00	0.00	0.01	0.03	0.05	0.01	0.05	0.03
All	0.01	0.00	0.01	0.04	0.01	0.00	0.01	0.03	0.05	0.01	0.05	0.03
					MAP@10							
Symmetrical	0.07	0.02	0.07	0.03	0.07	0.02	0.07	0.03	0.08	0.02	0.08	0.03
Non-Symmetrical	0.02	0.01	0.02	0.07	0.02	0.01	0.02	0.06	0.08	0.02	0.08	0.04
All	0.03	0.01	0.04	0.06	0.03	0.01	0.04	0.05	0.08	0.02	0.08	0.04

When the word *b* is not in the embeddings vocabulary, the question is not answered. We consider these cases the same as giving an incorrect answer. While this would not have much impact on the comparison of GloVe and both word2vec models, which were learned from the same corpus and are expected to cover the same vocabulary, it is not the case of the BERT embeddings. So, this is required for making comparison fairer. We also note that, following its definition, 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. This means that 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 as the positive examples).

The main conclusion is that TALES is a very challenging test. Accuracies are way under the best figures for syntactic and semantic analogies using the same embeddings (i.e., between 40 and 60% [19], [20]). Yet, a similar situation happens for English, on the BATS dataset [10], where best accuracies for lexical-semantic relations are always below 30%, with the single exception for the opposites with GloVe.

Another important conclusion is that we could not improve the performance with the embeddings obtained from BERT. Accuracy is so low that the differences between different layers are minimal if any. As mentioned in sub-section A of Section IV, context is not used for this task, and we thus could not take advantage of this feature of BERT. Yet, the main negative impact should result from limiting the word coverage to the entries in BERT's vocabulary. Still, even when questions with out-of-vocabulary (OOV) words are ignored, these accuracies are still significantly below the best with the static word embeddings (e.g., highest accuracy would be 0.12, achieved with Similar-to-B in any layer of BERT-ML for the symmetrical relations, followed by 0.09 with LRCos in the first layer of BERT-PT for non-symmetrical relations).

Considering all relations, the method+model configuration with the best accuracy was LRCos+GloVe (13%), but by the minimal margin of a single percentage point. On the other hand, for the symmetrical relations, the highest accuracies are achieved with 3CosAvg and with the Similar-to-B baseline in both word2vec models. This happens because both synonymy and antonymy occur between similar concepts, for which this baseline is already a good estimation. For synonymy, we can say that there are no benefits of using more sophisticated methods. This result is an important contribution to the overall accuracy of Similar-to-B. On the other hand, for non-symmetrical relations, LRCos+GloVe is not only the most accurate configuration but also the only with an average accuracy higher than 10% in this scenario, suggesting that, despite its limitations, LRCos suits this kind of relation better. At least when applied to GloVe, because in word2vec LRCos performs better for the symmetrical relations.

We note that the method originally applied for solving analogies in word2vec [3], 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.

Together with other pretrained models, the static models used here have previously been used for solving analogies of different types, in Portuguese, with 3CosAdd [19] and, more recently, also 3CosAvg and LRCos [20]. For those attempts using 3CosAdd, it was always clear that GloVe was the most accurate model for semantic analogies. On syntactic analogies, it was generally outperformed by fastText-SKIP, which deals better with regular word terminations. Yet, for attempts using LRCos, the best method, GloVe was the best model for both semantic and syntactic analogies [20]. We note that, for all of those analogies, relations are not symmetrical. Therefore, even if, in our work, the selection of the best method and model could raise some doubts, based on previous work, we can say that the LRCos+GloVe combination is the best option for solving analogies.

This is also consistent with related research for English [6], [10], [17], where GloVe is often used for this purpose, and the methods that use more instances as reference (3CosAvg and LRCos) perform better than those that try to solve the analogy based on a single instance (3CosAdd). Yet, even if the previous works for English considered the synonymy and antonymy relations, they did not include the Similar-to-B baseline in their comparison. According to our experiments, that baseline could perform better than the other methods, thus constituting an exception in the preference for LRCos, especially if the embeddings are learned with word2vec.

B. Overall MAP@10

Although accuracy has been extensively used by others for assessing word embeddings in analogy solving [3], [10], this is a limited metric, because it does not discriminate between methods that still rank the correct answer high, and were thus closer to be correct, and methods that gave it a lower rank. This is especially important when tests include questions with more than one acceptable answer. For this task, ranking can be considered by adopting retrieval-based measures like precision and recall, with a threshold on the similarity score, or the Mean Average Precision (MAP) [17]. Therefore, towards a different perspective on evaluation, we also computed the MAP@10. This also had in mind the future exploitation of the methods used for improving TALES or, better, lexical resources in general, with new relations discovered (see Section VI).

As expected, MAP is higher than accuracy but, for most relations, it is not substantially higher. This means that, even if not that many, there are indeed correct answers ranked between second and tenth. Nevertheless, MAP scores support the idea that GloVe is a consistent model, not only for non-symmetrical analogies, using LRCos, but also for the symmetrical, using 3CosAvg or simply Similar-to-B. And it is for the non-symmetrical where differences towards other models are more clear.

C. Per-Relation Performance

Tables V, VI and VII present the MAP@10 for each relation with each method+model configuration, respectively for the static word embeddings, for different layers of the BERT-ML model, and for different layers of the BERT-PT model. Results make it clear that some relations pose different challenges than others. For instance, following the discussion in sub-section A, the Similar-to-B baseline outperformed all the other methods for Synonym-of. Though, when applied to different models, it also becomes clear that Similar-to-B is not so good for Antonymy, as this method is outperformed by 3CosAvg and LRCos in all the static word embeddings. This helps us narrow down the exceptions where Similar-to-B is enough for symmetrical relations to only synonymy. In fact, their high overall performance for symmetrical relations, in Table II, was influenced by the presence of three types of synonymy and only one of antonymy.

The best MAP for synonymy between nouns (0.27) and between verbs (0.37) was achieved in word2vec-CBOW, though the latter was tied with word2vec-SKIP, always with Similar-to-B. Between adjectives, the best MAP (0.25) was in GloVe, this time tied with Similar-to-B and 3CosAvg. For Antonym-of, the best MAPs resulted from applying the LRCos method to word2vec-SKIP (0.30) and to GloVe (0.29).

TABLE V. MAP@10 for Different Relations, with Static Word Embedding Models and Different Methods

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	0.25	0.13	0.27	0.15	0.26	0.08	0.25	0.14	0.25	0.13
Synonym-of_v		0.34	0.15	0.33	0.27	0.37	0.19	0.34	0.26	0.37	0.18	0.33	0.23
Synonym-of_adj		**0.25**	0.11	**0.25**	0.13	0.23	0.13	0.24	0.15	0.22	0.10	0.19	0.10
Antonym-of_adj		0.25	0.16	0.27	0.29	0.24	0.14	0.25	0.26	0.22	0.17	0.24	**0.30**
Hypernym-of_n	D	**0.20**	0.07	0.17	0.07	0.19	0.08	0.19	0.05	0.15	0.07	0.15	0.06
(+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	**0.19**	0.07	0.16	0.12	0.17	0.08	0.17	0.07	0.18	0.08	0.17	0.13
(-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
Hypernym-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	0.11	0.05	0.12	**0.13**	0.00	0.00	0.03	0.07	0.02	0.02	0.04	0.12
	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 VI. MAP@10 for Different Relations, with Word Embeddings from Different Layers of BERT ML and Different Methods

Relation		Layer 1				Layer 2				Layer 11			
		SIM	3CAD	3CAV	LRC	SIM	3CAD	3CAV	LRC	SIM	3CAD	3CAV	LRC
Synonym-of_n		0.01	0.00	0.01	0.01	0.01	0.00	0.01	0.00	0.01	0.00	0.02	0.00
Synonym-of_v		0.05	0.00	0.04	0.04	0.05	0.00	0.03	0.04	0.04	0.00	0.03	0.04
Synonym-of_adj		0.03	0.01	0.04	0.00	0.03	0.01	0.04	0.00	0.03	0.00	0.03	0.00
Antonym-of_adj		0.02	0.00	0.02	0.01	0.03	0.00	0.02	0.01	0.03	0.00	0.04	0.00
Hypernym-of_n	D	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.00
(+concrete)	I	0.01	0.00	0.01	0.09	0.01	0.00	0.02	0.09	0.00	0.00	0.02	0.09
Hypernym-of_n	D	0.01	0.00	0.01	0.02	0.01	0.00	0.01	0.02	0.01	0.00	0.01	0.00
(-concrete)	I	0.03	0.01	0.03	0.16	0.04	0.01	0.04	0.15	0.00	0.01	0.01	0.11
Hypernym-of_v	D	0.04	0.00	0.04	0.00	0.04	0.00	0.04	0.00	0.02	0.00	0.02	0.01
	I	0.04	0.00	0.04	0.08	0.03	0.00	0.04	0.06	0.03	0.00	0.03	0.02
Part-of	D	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00
	I	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Purpose-of	D	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	I	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.03

TABLE VII. MAP@10 for Different Relations, with Word Embeddings from Different Layers of BERTimbau and Different Methods

Relation		Layer 1				Layer 2				Layer 23			
		SIM	3CAD	3CAV	LRC	SIM	3CAD	3CAV	LRC	SIM	3CAD	3CAV	LRC
Synonym-of_n		0.05	0.02	0.05	0.06	0.06	0.02	0.05	0.04	0.09	0.04	0.08	0.04
Synonym-of_v		0.02	0.00	0.02	0.06	0.03	0.00	0.02	0.07	0.08	0.01	0.10	0.06
Synonym-of_adj		0.08	0.02	0.08	0.00	0.07	0.02	0.07	0.00	0.05	0.01	0.04	0.00
Antonym-of_adj		0.11	0.03	0.11	0.00	0.11	0.03	0.12	0.00	0.12	0.04	0.12	0.01
Hypernym-of_n	D	0.03	0.00	0.03	0.00	0.03	0.00	0.04	0.00	0.04	0.00	0.03	0.00
(+concrete)	I	0.02	0.01	0.02	0.12	0.01	0.01	0.02	0.09	0.08	0.03	0.08	0.07
Hypernym-of_n	D	0.02	0.00	0.01	0.00	0.02	0.00	0.02	0.01	0.12	0.03	0.09	0.00
(-concrete)	I	0.03	0.02	0.04	0.19	0.03	0.03	0.04	0.17	0.15	0.08	0.17	0.16
Hypernym-of_v	D	0.01	0.00	0.01	0.05	0.01	0.00	0.01	0.03	0.13	0.02	0.09	0.01
	I	0.02	0.01	0.02	0.16	0.02	0.01	0.03	0.13	0.16	0.05	0.15	0.05
Part-of	D	0.00	0.00	0.00	0.06	0.01	0.00	0.00	0.04	0.01	0.01	0.01	0.02
	I	0.02	0.01	0.02	0.00	0.02	0.00	0.02	0.01	0.04	0.01	0.04	0.00
Purpose-of	D	0.02	0.00	0.02	0.03	0.02	0.00	0.02	0.02	0.03	0.00	0.05	0.03
	I	0.04	0.01	0.07	0.09	0.04	0.01	0.05	0.06	0.02	0.02	0.08	0.04

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

The best MAP for the direct Hypernym-of was again for the Similar-to-B, with GloVe for nouns, and word2vec-SKIP for verbs. To some extent, the performance of this baseline in these relations is explained by the high similarity in hypernym-hyponym pairs, as it happens for synonymy, i.e., hyponyms are very similar to their hypernyms, only more specific than synonyms. Yet, LRCos performs much better on the inverse direction than on the direct, suggesting that it is more difficult to find hyponyms (b^*) given their hypernym, when compared to the other way round. This reflects the higher number of hyponyms, especially when considering indirect hyponyms. Even though, in most cases, there was more than one acceptable answer, a list of hyponyms can be so extensive that some will frequently be missing (see Section VI). Besides the large number, the heterogeneity of the hyponyms also contributes to this low result, making it harder to learn a good representation of the hyponym class with the logistic regression classifier. This further explains, at least partially, why differences between the direct and inverse Hypernymy-of are smaller for 3CosAvg, which does not rely on a classifier, in some cases with better performance for the direct relation. This again suggests that different configurations are better suited for different goals.

Still on Hypernym-of, 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 general text and, for LRCos, to represent their class. Here, the main surprise was the performance of the BERTimbau embeddings, namely the first-layer encoding, where LRCos achieved the best MAP overall for the inverse Hypernym-of between abstract words (0.19), suggesting that the aforementioned contexts are better captured by BERT. Yet, what also contributed to this interesting performance is that this is the relation with the lower proportion of questions with OOV words, only two out of the 50. For the same relation in BERTimbau, we also highlight the MAP of the 3CosAvg in layer 23 (0.17). In fact, for all methods but LRCos the MAP of all relations improves for the upper layer. On the other hand, for LRCos, it decreases for most relations in that layer, which would be more in agreement with previous work [42], [44]. Still in layer 23 of BERTimbau, Similar-to-B achieved a MAP above 10% for five different relations: Antonym-of (0.12) and all direct and inverse hypernymy relations between verbs and abstract nouns. BERT-ML performed much worse, with a single MAP above 10%, namely the inverse Hypernym-of between abstract words, with LRCos, in any layer (0.16, 0.15, 0.11).

For Part-of, performance was the poorest of all relations. In GloVe, the MAP with LRCos for the direct relations (0.16) was twice the same measure for the inverse (0.08). With 3CosAvg, the latter was slightly higher (0.10). And even though it is the second relation for which there are less questions with OOV words in the BERT models, 17 and nine, respectively for BERT-ML and BERTimbau, it is far from achieving the second best MAP. 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 affected by other issues (see Section VI).

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 LRCos in GloVe, but it was also considerably higher than the baselines, and contrasting with the lower performance in word2vec-CBOW. This stresses that LRCos is well-suited for different kinds of semantic relations. Moreover, although not included in similar tests for English, this suggests that it would be interesting to include the Used-For relation (inverse of Purpose-of) in such benchmarks.

VI. On the Utility of Word Embeddings for Relation Discovery

As discussed in the previous section, when compared to syntactic or word knowledge analogies, solving lexical-semantic analogies from word embeddings is a challenging task. This happens for many reasons. For instance, lexical-semantic relations typically include very frequent words in language, which would result in better representations, if it were not for ambiguity also being higher, i.e., a significant number of these words has more senses than, for instance, names of cities and countries. Another reason is a great number of questions with many possible answers, in opposition, e.g., to syntactic or capital-country relations, for which there is a single answer. In fact, for some cases, there are so many possible answers that they simply cannot all be covered by the dataset, which, in our case, means that they are not in the source lexical resources as well (at least in more than one). Therefore, inspecting the answers by the different method+model configurations and identifying typical issues may, on the one hand, lead to a handful of fixes in future versions of the dataset (i.e., inclusion of more possible answers) and, potentially more interesting, result in important conclusions and suggestions regarding the utility of this kind of approach for enriching lexical resources.

In this section, we first look at the proportion of answers, in a sample, that were automatically considered incorrect, but were still acceptable. Then, we focus on some relation types to find typical issues and confirm that, despite missing from TALES, most acceptable answers constitute good candidates for enriching the lexical resources exploited in its creation. The main conclusion here is that word embeddings should be seen as useful sources of lexical-semantic information that, with analogy solving methods, might be ready for enriching structured lexical resources. Of course, given the still great amount of incorrect answers, the discovered relations should be seen as suggestions, to be considered, or not, for inclusion in the lexical resources, also depending on criteria set by the resource creators.

A. Acceptable Incorrect Answers

For better insights on the achieved performance and typical issues, we inspected the results of the configuration with the highest accuracy, LRCos+GloVe. Towards a more systematic analysis: (i) we focused on the top-10 answers, the same considered when computing MAP@10, for a sample of 15 randomly selected questions of each relation, totalling 210 questions and 2,100 answers; (ii) out of those considered to be incorrect, we manually identified those that were still acceptable, roughly meaning that the relation would make sense. For the target sample, Table VIII shows the proportion of correct answers side-by-side the proportion of answers that we considered acceptable, including one example for each relation type. In this sample, 259 answers (≈12.3%) automatically marked as incorrect were considered acceptable by us, which corresponds to more than twice the number of answers automatically marked as correct in the same sample (111, ≈5.3%).

We recall that incorrect answers are those that did not meet our criteria for inclusion in TALES, i.e., they correspond to relations that are not in any of the exploited lexical resources or they are in a single one, so confidence on them is low. More than suggesting their addition to future versions of TALES, these answers highlight that the exploited lexical resources are limited in terms of coverage, and that this kind of approach can be useful for enriching such resources. An exception

TABLE VIII. Proportion of Acceptable Relation Instances Considered Incorrect in the Answers of LRCos+GloVe

Relation		Correct	Acceptable	Example (PT)	(EN)
Synonym-of_n		6.0%	6.7%	(*luta, combate*)	(fight, combat)
Synonym-of_v		6.7%	8.7%	(*voltar, retornar*)	(come back, return)
Synonym-of_adj		5.3%	6.7%	(*antigo, primitivo*)	(ancient, primitive)
Antonym-of_adj		4.7%	6.0%	(*legítimo, ilegal*)	(legitimate, illegal)
Hypernym-of_n	D	2.7%	10.7%	(*sala, auditório*)	(room, auditorium)
(+concrete)	I	8.7%	16.7%	(*edifício, prédio*)	(edifice, building)
Hypernym-of_n	D	4.0%	18.7%	(*regra, analogia*)	(rule, analogy)
(-concrete)	I	4.7%	7.3%	(*memória, lembrança*)	(memory, reminder)
Hypernymy-of_v	D	5.3%	13.3%	(*receber, acolher*)	(receive, welcome)
	I	10.7%	16.0%	(*mostrar, demonstrar*)	(show, demonstrate)
Part-of	D	3.3%	17.3%	(*porta, armário*)	(door, wardrobe)
	I	4.0%	13.3%	(*humano, cérebro*)	(human, brain)
Purpose-of	D	5.3%	17.3%	(*aquecer, forno*)	(heat, hoven)
	I	2.7%	14.7%	(*camisola, vestir*)	(sweater, wear)

regards a minority of acceptable answers that are the plural (2.7% of the acceptable) or feminine form (1.1% of the acceptable) and are thus not expected to be found in the lexical resources used, because their entries are typically lemmatised. This happens, for instance, in Part-of relations, with *segundos* (seconds) Part-of *minuto* (minute); *minutos* (minutes) Part-of *hora* (hour); or *alunos* (students) Part-of *escola* (school).

Another situation regards transitive relations (e.g., Hypernymy and Part-of), because some lexical resources only make direct connections explicit, not indirect (e.g., inherited hypernyms). This also depends on the taxonomy adopted by the lexical resource and is much noisier in resources extracted from dictionaries. Even though BATS includes relations inherited through transitivity (see e.g., Fig. 1), we did not consider them in the creation of TALES, both due to the aforementioned issue and to the lack of information on word senses, in some resources.

We note that, for each relation, the proportion of acceptable incorrect answers is not correlated with the proportion of correct answers (Pearson coefficient is 0.08). The former is higher for all relations, but this difference ranges from 0.7 points (Synonym-of_n) to 14.7 (direct Hypernym-of_n abstract). On the other hand, the proportion of acceptable incorrect answers is related to the lack of coverage of the instance by TALES, and thus, indirectly, by the lexical resources. By manual inspection, we confirmed that the average number of possible answers, i.e., words related in the target way, is an important contribution to the proportion of acceptable answers not in TALES. For Antonym-of, the relation for which this number is lower, as well as for the other symmetrical relations, the coverage of the lexical resources is not as low as for the other relations. Even in a broad interpretation of antonymy and synonymy, the set of antonyms and synonyms is not as large as for other relations. This is also the case of the inverse Hypernym-of_n for abstract words, but not for the remaining non-symmetrical relations. In fact, the universe of instances of the non-symmetrical relations is considerably larger. As mentioned earlier, a hypernym has several hyponyms, but an object might also have many parts or be used for different purposes. This number increases if inherited relations are considered, namely for hypernymy (e.g., animal Hypernym-of mammal Hypernym-of dog) and part-of (e.g., minute Part-of hour Part-of day Part-of month, ...).

B. Typical Issues

A deeper error analysis was made for the relation with a lower MAP in TALES (the inverse Part-of) and for those with a higher proportion of acceptable answers. Yet, recalling the recurrently given example of hypernymy – a concept might have a huge number of hyponyms – we first focus on the inverse Hypernym-of relation.

As expected, they can be so many that TALES does not cover all possible hyponyms of most Hypernym-of entries. For instance, it includes five types of *escola* (school) but not 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 are in any of the lexical resources used. Some, in fact, can be used just as modifiers of *escola*, often appearing together (e.g., *escola preparatória* or *escola secundária*), but they can also be used alone, with the same meaning. 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 LKBs. 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, again 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).

Considering the inverse Part-of (Has-Part) relation, for which MAP was very low, we came to the conclusion that the test for this relation includes several difficult entries. Some have multiple senses that can be significantly different, such as *ser* (to be / living being), *câmara* (camera, chamber), or *programa* (program, show). Others refer to abstract concepts, like *todo* (whole), *mundo* (world), *espaço* (space), *organização* (organization), *vida* (life) or *coisa* (thing). On the one hand, the issue of ambiguity is minimised by the presence of several acceptable answers. On the other hand, ambiguous words are used in different contexts, making the relations less obvious in the embedding 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 enough for this.

Two other issues were noted regarding the confusion of this relation with:

- Hyponymy, i.e., some answers were hyponyms of *b* and not part. For instance, 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).

- Its inverse, i.e., some answers were not the parts, but the whole of *b*. For instance, 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).

Looking at the relations for which more acceptable answers were found, they include again many cases for which there is a large set of acceptable answers, and not all are in TALES. Examples of such answers include: words for which *sentimento* (feeling) is a hypernym, namely *otimismo* (optimism) and *ansiedade* (anxiety); words for which *porta* (door) is a part, namely *prédio* (building), *casa* (house), *armário* (wardrobe) or *banheiro* (bathroom); or words for which *cozinhar* (to cook) is a purpose-of, namely *forno* (hoven), *molho* (sauce) or *caldo* (broth). These examples also show that, despite acceptable answers, not all are the most obvious and their inclusion would probably require better-defined criteria. We would say that it is virtually impossible to name all possible feelings, all things which have a door, or everything used for cooking, which shows why a lexical resource will never be fully complete regarding some relations. Nevertheless, we believe to have shown that this can be minimised by exploiting word embeddings learned from large corpora.

VII. Concluding Remarks

Towards better insights on how lexical-semantic relations are preserved in pretrained models of word embeddings for Portuguese, we have presented the following contributions:

- TALES, a new analogy-like test covering 14 types of lexical-semantic relations, created automatically with information in Portuguese lexical resources;
- An evaluation covering four different analogy solving methods in TALES, when applied to five pretrained models of Portuguese word embeddings, including static word embeddings as well as embeddings obtained from BERT models;
- An analysis of the obtained results, having in mind the application of the adopted methods for relation discovery in word embeddings and their utility for enriching lexical resources.

TALES is freely available from https://github.com/NLP-CISUC/PT-LexicalSemantics, for anyone willing to use it. As we have shown, it is a challenging test, for which high performances will require better solving methods or different models of word embeddings. Interested researchers may also want to assess other models for Portuguese or alternative ways of exploiting the models used here. According to our experiments, better results are achieved with static word embeddings than with BERT. However, the performance of the latter can most certainly be improved, if this model is used differently. To leverage on Vecto, the platform we used for loading the embeddings and running the tests, we had to get static word representations from BERT, based on its vocabulary file, which makes it impossible to get representations for OOV words and, more importantly, to words obtained from a combination of subwords. While context does not seem to be important in this kind of text, recent work for English has shown that BERT models can still outperform static word embeddings when solving analogies [42]. For better analysing if this is also the case for Portuguese, in the future, we will study alternative ways of handling BERT. In order to keep using Vecto, one possibility would be to include not only a representation for each entry in BERT's vocabulary, but also for all the words in TALES. However, if we just do this, results would probably be positively biased, due to less confusion. Another possibility is to include the encodings of words in a large representative list, starting, for instance, with the vocabulary of the static word embeddings. We should also look at previous work on using BERT for solving lexical tasks (e.g., [44]).

Future experiments may also include alternative analogy solving methods. While we did not get improvements with 3CosMul and PairDistance [46], more recent methods, like the Translation and the Regression Model [17], are not included in Vecto, and were thus not tested.

Based on the analysis of incorrect answers, namely on the proportion of acceptable answers, we are looking forward to using this kind of approach for suggesting new relation instances to Portuguese lexical resources and thus contributing to their semi-automatic enrichment. If focused on a single lexical resource, it is perhaps advisable to use a new test obtained exclusively from its relations, to better capture the criteria followed in its creation. For some LKBs, we could possibly leverage on the word sense organisation and, if desired, include inherited relations in the test. After this, it should be a matter of going through all the incorrect answers and consider their addition to the LKB or not. As we have shown, even though many may be definitely incorrect, some might be acceptable instances that are simply missing from the resource. This will, or course, contribute to better structured lexical resources, with higher coverage.

Acknowledgment

This is an extended version of a paper describing the creation of the TALES dataset and its usage for assessing three models of static word embeddings [47], presented at the ECAI 2020 Workshop on Hybrid Intelligence for Natural Language Processing Tasks. Besides some more detailed discussions, the main additions are the inclusion of the BERT models in our experiment and a more thorough inspection of the results, in order to complement our discussion on the utility of this kind of approach for enriching computational lexical resources.

We would like to thank the reviewers for their useful comments that lead to a better version of this manuscript.

References

[1] C. Fellbaum Ed., *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, 1998.

[2] Z. Harris, "Distributional structure," *Word*, vol. 10, no. 2-3, pp. 1456–1162, 1954.

[3] T. Mikolov, K. Chen, G. Corrado, J. Dean, "Efficient estimation of word representations in vector space," in *Proceedings of the Workshop track of ICLR*, 2013.

[4] J. Pennington, R. Socher, C. D. Manning, "GloVe: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, EMNLP 2014, 2014, pp. 1532–1543, ACL.

[5] A. Gladkova, A. Drozd, S. 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*, 2016, pp. 8–15, ACL.

[6] E. Vylomova, L. Rimell, T. Cohn, T. 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, 2016, pp. 1671–1682, ACL.

[7] A. Querido, R. Carvalho, J. Rodrigues, M. Garcia, J. Silva, C. Correia, N. Rendeiro, R. Pereira, M. Campos, A. Branco, "LX-LR4DistSemEval: a collection of language resources for the evaluation of distributional semantic models of Portuguese," *Revista da Associação Portuguesa de Linguística*, vol. 3, no. 3, pp. 265–283, 2017.

[8] G. Hirst, "Ontology and the lexicon," in *Handbook on Ontologies*, S.

Staab, R. Studer Eds., International Handbooks on Information Systems, Springer, 2004, pp. 209–230.

[9] V. de Paiva, L. Real, H. Gonçalo Oliveira, A. Rademaker, C. Freitas, A. Simões, "An overview of Portuguese wordnets," in *Proceedings of 8th Global WordNet Conference*, GWC'16, Bucharest, Romania, 2016, pp. 74–81.

[10] A. Drozd, A. Gladkova, S. 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, 2016, pp. 3519–3530.

[11] J. Devlin, M.-W. Chang, K. Lee, K. 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, 2019, pp. 4171–4186, ACL.

[12] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, E. Ruppin, "Placing search in context: The concept revisited," *ACM Trans. Inf. Syst.*, vol. 20, pp. 116–131, Jan. 2002.

[13] F. Hill, R. Reichart, A. Korhonen, "Simlex-999: Evaluating semantic models with genuine similarity estimation," *Computational Linguistics*, vol. 41, no. 4, pp. 665–695, 2015.

[14] E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Paşca, A. 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, 2009, pp. 19–27, ACL.

[15] D. Jurgens, S. Mohammad, P. Turney, K. 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)*, 2012, pp. 356–364, ACL.

[16] M. Baroni, A. Lenci, "How we BLESSed distributional semantic evaluation," in *Proceedings of GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, Edinburgh, UK, 2011, pp. 1–10, ACL.

[17] Z. Bouraoui, S. Jameel, S. Schockaert, "Relation induction in word embeddings revisited," in *Proceedings of the 27th International Conference on Computational Linguistics*, COLING 2018, Santa Fe, New Mexico, USA, Aug. 2018, pp. 1627–1637, ACL.

[18] H. Gonçalo Oliveira, "Distributional and Knowledge-Based Approaches for Computing Portuguese Word Similarity," *Information*, vol. 9, no. 2, 2018.

[19] N. S. Hartmann, E. R. Fonseca, C. D. Shulby, M. V. Treviso, J. S. Rodrigues, S. 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.

[20] T. Sousa, H. G. Oliveira, A. Alves, "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*, vol. 83 of OASIcs, 2020, pp. 9:1–9:14, Schloss Dagstuhl - Leibniz-Zentrum für Informatik.

[21] R. Wilkens, L. Zilio, E. Ferreira, A. Villavicencio, "B2SG: a TOEFL-like task for Portuguese," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, LREC 2016, Paris, France, 2016, ELRA.

[22] R. Navigli, S. P. Ponzetto, "BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network," *Artificial Intelligence*, vol. 193, pp. 217–250, 2012.

[23] N. Calzolari, L. Pecchia, A. Zampolli, "Working on the italian machine dictionary: a semantic approach," in *COLING 1973 Volume 2: Computational And Mathematical Linguistics: Proceedings of the International Conference on Computational Linguistics*, 1973, Association for Computational Linguistics.

[24] M. A. Hearst, "Automatic acquisition of hyponyms from large text corpora," in *Proceedings of 14th Conference on Computational Linguistics*, COLING 92, Morristown, NJ, USA, 1992, pp. 539–545, Association for Computational Linguistics.

[25] R. Snow, D. Jurafsky, A. Ng, "Learning syntactic patterns for automatic hypernym discovery," *Advances in neural information processing systems*, vol. 17, pp. 1297–1304, 2005.

[26] P. Pantel, M. Pennacchiotti, "Espresso: Leveraging generic patterns for automatically harvesting semantic relations," in *Procs of 21st International Conference on Computational Linguistics and 44th annual meeting of the Association for Computational Linguistics*, Sydney, Australia, 2006, pp. 113–120, ACL Press.

[27] H. Gonçalo Oliveira, P. Gomes, "ECO and Onto.PT: A flexible approach for creating a Portuguese wordnet automatically," *Language Resources and Evaluation*, vol. 48, no. 2, pp. 373–393, 2014.

[28] H. Gonçalo Oliveira, D. Santos, P. Gomes, N. Seco, "PAPEL: A dictionary-based lexical ontology for Portuguese," in *Proceedings of Computational Processing of the Portuguese Language - 8th International Conference (PROPOR 2008)*, vol. 5190 of LNCS/LNAI, Aveiro, Portugal, September 2008, pp. 31–40, Springer.

[29] H. Gonçalo Oliveira, "A survey on Portuguese lexical knowledge bases: Contents, comparison and combination," *Information*, vol. 9, no. 2, 2018.

[30] A. Simões, Á. I. Sanromán, J. J. Almeida, "Dicionário-Aberto: A source of resources for the Portuguese language processing," in *Proceedings of Computational Processing of the Portuguese Language, 10th International Conference (PROPOR 2012), Coimbra Portugal*, vol. 7243 of LNCS, April 2012, pp. 121–127, Springer.

[31] L. Anton Pérez, H. Gonçalo Oliveira, P. Gomes, "Extracting lexical-semantic knowledge from the Portuguese Wiktionary," in *Proceedings of the 15th Portuguese Conference on Artificial Intelligence*, EPIA 2011, Lisbon, Portugal, October 2011, pp. 703–717, APPIA.

[32] E. G. Maziero, T. A. S. Pardo, A. D. Felippo, B. C. Dias-da-Silva, "A Base de Dados Lexical e a Interface Web do TeP 2.0 - Thesaurus Eletrônico para o Português do Brasil," in *VI Workshop em Tecnologia da Informação e Linguagem Humana*, TIL, 2008, pp. 390–392.

[33] V. de Paiva, A. Rademaker, G. de Melo, "OpenWordNet-PT: An Open Brazilian WordNet for Reasoning," in *Proceedings of 24th International Conference on Computational Linguistics*, COLING (Demo Paper), 2012.

[34] A. Simões, X. G. Guinovart, "Bootstrapping a Portuguese wordnet from Galician, Spanish and English wordnets," in *Advances in Speech and Language Technologies for Iberian Languages, Proceedings of 2nd International Conference, IberSPEECH 2014, Las Palmas de Gran Canaria, Spain*, vol. 8854 of LNCS, 2014, pp. 239–248, Springer.

[35] B. C. Dias-da-Silva, "Wordnet.Br: An exercise of human language technology research," in *Proceedings of 3rd International WordNet Conference (GWC)*, GWC 2006, South Jeju Island, Korea, January 2006, pp. 301–303.

[36] A. Barreiro, "Port4NooJ: an open source, ontology-driven portuguese linguistic system with applications in machine translation," in *Proceedings of the 2008 International NooJ Conference (NooJ'08)*, Budapest, Hungary, 2010, Newcastle-upon-Tyne: Cambridge Scholars Publishing.

[37] R. Speer, J. Chin, C. Havasi, "Conceptnet 5.5: An open multilingual graph of general knowledge," in *Proceedings of Thirty-First Conference on Artificial Intelligence (AAAI)*, 2017, pp. 4444–4451.

[38] P. A. Rocha, D. 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),* São Paulo, 19-22 de Novembro 2000, pp. 131–140, ICMC/USP.

[39] D. Santos, E. Bick, "Providing Internet access to Portuguese corpora: the AC/DC project," in *Proceedings of 2nd International Conference on Language Resources and Evaluation*, LREC 2000, 2000, pp. 205–210.

[40] A. Paivio, J. C. Yuille, S. A. Madigan, "Concreteness, imagery, and meaningfulness values for 925 nouns," *Journal of Experimental Psychology monograph supplement*, vol. 76, no. 1, pp. 1–25, 1968.

[41] A. P. Soares, A. S. Costa, J. Machado, M. Comesaña, H. M. Oliveira, "The Minho Word Pool: Norms for imageability, concreteness, and subjective frequency for 3,800 Portuguese words," *Behavior Research Methods*, vol. 49, no. 3, pp. 1065–1081, 2017.

[42] K. Ethayarajh, "How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 55–65.

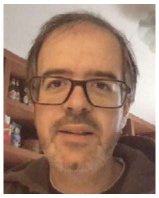
[43] F. Souza, R. Nogueira, R. Lotufo, "Bertimbau: Pretrained bert models for brazilian portuguese," in *Proceedings of the Brazilian Conference on Intelligent Systems (BRACIS 2020)*, vol. 12319 of LNCS, Cham, 2020, pp. 403–417, Springer.

[44] I. Vulić, E. M. Ponti, R. Litschko, G. Glavaš, A. Korhonen, "Probing pretrained language models for lexical semantics," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 7222–7240.

[45] T. Linzen, "Issues in evaluating semantic spaces using word analogies," in *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, Berlin, Germany, Aug. 2016, pp. 13–18, ACL.

[46] O. Levy, Y. Goldberg, "Linguistic regularities in sparse and explicit word representations," in *Proceedings of 18th Conference on Computational Natural Language Learning*, CoNLL 2014, 2014, pp. 171–180, ACL.

[47] H. Gonçalo Oliveira, T. Sousa, A. Alves, "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)*, vol. 2693 of CEUR Workshop Proceedings, 2020, pp. 41–47, CEUR-WS.org.



Hugo Gonçalo Oliveira

Hugo Gonçalo Oliveira is an Assistant Professor at the Department of Informatics Engineering of the University of Coimbra and a researcher of the Center of Informatics and Systems of the same university (CISUC), in the Cognitive and Media Systems lab. His main research interests lie in Artificial Intelligence and sub-domains of Computational Creativity (CC) and Natural Language Processing (NLP), with a focus on the Portuguese language. He has developed some prototypes for generating creative text in Portuguese, such as Tra-la-Lyrics and PoeTryMe, and participated in the FET European projects ConCreTe and PROSECCO. His research also led to the development of several lexical resources for Portuguese, such as PAPEL and Onto.PT, the main scope of his PhD thesis (2013), which was awarded as the best in the computational processing of Portuguese (2011-2014). His current research also covers topics like Semantic Textual Similarity and Conversational Agents, following his leading role in two national-funded projects on the latter (AIA, FLOWANCE). He is the author of more than 100 peer-reviewed scientific papers, has participated in national-funded projects (InfoCrowds, REMINDS, Socialite), in the organisation of scientific conferences and workshops (CC-NLG 2017-19, SLATE 2019, PROPOR 2018) and shared tasks (Second HAREM, TweetMT, ASSIN-2), and is a regular member of the program committee of some of the main scientific events on CC (ICCC) and NLP (ACL, EMNLP, COLING).

Tiago Sousa

Tiago Sousa has a degree in Informatics Engineering from Institute of Engineering of Coimbra and is a student of the Master in Informatics and Systems at the Institute of Engineering of Coimbra, where his main research focus is around Portuguese word embeddings and their applications to analogy solving and relation discovery. Since 2016, he also works at Present Technologies as a software developer.

Ana Alves

Ana Alves is, since 2007, a member of the Ambient Intelligence Laboratory (AmILab) of the Cognitive and Media Systems (CMS) group, integrated in the Center for Informatics and Systems of the University of Coimbra (CISUC). Her research is dedicated mainly to the way urban spaces are organized and how people use them by influencing patterns of mobility and land use analysis. The information to support this analysis is mined primarily from the Web and social networks and consists mostly of textual data. Learning to interconnect and represent this information is another research challenge she is devoted to. She is also an Assistant Professor at Polytechnic Institute of Coimbra (IPC), Coimbra Institute of Engineering (ISEC), since 2000, lecturing courses on programming, operating systems and ubiquitous computing. She holds a Ph.D. in Science and Information Technology since 2012, MSc. in Informatics and Systems in 2004 (pre-bolonha), and a 5-year Bachelor degree in Informatics Engineering in 2000 from the University of Coimbra. She is a member of scientific associations such as: Portuguese Artificial Intelligence Association (APPIA - Associação Portuguesa Para a Inteligência Artificial); Association for Computational Linguistics (ACL); and Registered researcher at the Foundation for Science and Technology (Fundação para a Ciência e Tecnologia).