information

MDPI

Article

Distributional and Knowledge-Based Approaches for Computing Portuguese Word Similarity †

Hugo Gonçalo Oliveira

Centre for Informatics and Systems of the University of Coimbra (CISUC), Department of Informatics Engineering, University of Coimbra, 3030-290 Coimbra, Portugal; hroliv@dei.uc.pt

† This paper is an extended version of our paper published in Progress in Artificial Intelligence, Proceedings of 18th EPIA Conference on Artificial Intelligence, Porto, Portugal, 5–8 September 2017; Volume 10423 of LNCS, Springer, pp. 828–840, entitled Unsupervised Approaches for Computing Word Similarity in Portuguese.

Received: 18 December 2017; Accepted: 4 February 2018; Published: 8 February 2018

Abstract: Identifying similar and related words is not only key in natural language understanding but also a suitable task for assessing the quality of computational resources that organise words and meanings of a language, compiled by different means. This paper, which aims to be a reference for those interested in computing word similarity in Portuguese, presents several approaches for this task and is motivated by the recent availability of state-of-the-art distributional models of Portuguese words, which add to several lexical knowledge bases (LKBs) for this language, available for a longer time. The previous resources were exploited to answer word similarity tests, which also became recently available for Portuguese. We conclude that there are several valid approaches for this task, but not one that outperforms all the others in every single test. Distributional models seem to capture relatedness better, while LKBs are better suited for computing genuine similarity, but, in general, better results are obtained when knowledge from different sources is combined.

Keywords: semantic similarity; word similarity; lexical knowledge bases; lexical semantics; word embeddings; distributional semantics

1. Introduction

Semantic similarity is a key problem in natural language processing (NLP) and understanding (NLU). Specifically, word similarity aims at determining the likeness of meaning transmitted by two words, and is generally a necessary step towards computing the semantic similarity of larger units, such as phrases or sentences. This is why there are many automatic approaches for computing word similarity, as well as several benchmarks that enable the assessment and comparison of distinct approaches for this purpose.

As it happens for other NLP tasks, most related work targets English, because it is widely spoken, which also results in more benchmarks in this language. Yet, especially since word embeddings–vector representations of words learned with a neural network [1]–became a trend in NLP, researchers have developed both benchmarks and approaches for computing semantic similarity in other languages, including Portuguese.

This work uses recently released word similarity tests in Portuguese and answers them with unsupervised approaches that either exploit the structure of existing lexical knowledge bases (LKBs) or distributional models of words, including word embeddings. The goal of the automatic procedures is to score the similarity of two words, which may then be assessed by comparison with the scores in the benchmark test, in this case, based on human judgements.

The main contribution of this work is the comparison of several unsupervised procedures employed for computing word similarity in Portuguese, which might support the choice of

approaches to adopt in more complex tasks, such as semantic textual similarity (for Portuguese, see ASSIN [2]) or other tasks involved in a NLU pipeline (useful for e.g., conversational agents), or in a semantically-enriched search engine. Indirectly, the resources underlying each approach end up also being compared. For instance, results provide cues on the most suitable LKBs for computing word similarity, also a strong hint on the quality and coverage of these resources. Overall, this work involved different procedures for computing word similarity and several resources, namely: two procedures applied to open Portuguese wordnets, including one fuzzy wordnet; two different procedures applied to available semantic networks for Portuguese, or to networks that result from their combination; one procedure based on the co-occurrence of words in articles of the Portuguese Wikipedia; and one final procedure that computes similarity from several different models of word embeddings currently available for Portuguese. The work is especially directed to those users that are interested in computing the similarity of Portuguese words but do not have the conditions for creating new broad-coverage semantic models from scratch. In fact, it can also be seen as a survey of resources–semantic models and benchmarks–currently available for this purpose.

The remainder of this paper starts with a brief overview on semantic similarity, variants, common approaches, and a focus on this topic for Portuguese. After that, the benchmarks used here are presented, followed by a description of the resources and approaches applied. Before concluding, the results obtained for each approach are reported and discussed, which includes a look at the state-of-the-art and the combination of different approaches towards better results. In general, the best results obtained are highly correlated with human judgements. Yet, depending on the nature of the dataset, both the best approach and underlying resource is different. An important conclusion is that the best results are obtained, first, with approaches that combine different resources of the same kind, then, by combining the former with models of a different kind.

2. Related Work

Semantic similarity measures the likeness of the meaning transmitted by two units, which can either be instances in an ontology or linguistic units, such as words or sentences. This involves comparing the features shared by each meaning, which sets their position in a taxonomy, and considers semantic relations such as synonymy, for identical meanings, or hypernymy, hyponymy and co-hyponymy, for meanings that share several features. Semantic relatedness goes beyond similarity and considers any other semantic relation that may connect meanings. For instance, the concepts of *dog* and *cat* are semantically similar, but they are not so similar to *bone*. On the other hand, *dog* is more related to *bone* than *cat* is, because *dogs* like and are often seen with *bones*.

Word similarity tests are collections of word pairs with a similarity score based on human judgements. To answer such tests, humans would either look for the words in a dictionary or search for their occurrence in large corpora, possibly with the help of a search engine. This has a parallelism with the common approaches for determining word similarity automatically and unsupervisedly: (i) corpus-based approaches, also known as distributional, resort to a large corpus and analyse the distribution of words; (ii) knowledge-based approaches exploit the contents of a dictionary or lexical knowledge base (i.e., a machine-friendly representation of dictionary knowledge). It should be noted that the distinction of similarity and relatedness is not very clear for everyone. Therefore, whether the test scores reflect similarity or relatedness is also not always completely clear. Nevertheless, approaches for one are often applied to the other.

Corpus-based approaches rely on the distributional hypothesis–words that occur in similar contexts tend to have similar meanings [3]–and often represent words in a vector space model [4]. Recent work uses neural networks to learn vectors from very large corpora, which are more accurate and computationally efficient at the same time. Sucessfull models of this kind include word2vec [1], GloVE [5], or fastText [6].

The similarity of two words may also be computed from their probability of occurrence in a corpus. Pointwise Mutual Information (PMI) [7] quantifies the discrepancy between the probability of two

words, a and b, co-occurring ($P(a,b)$), given their joint distribution and their individual distributions ($P(a)$ and $P(b)$), and assuming their independence. PMI can be computed according to Equation (1).

$$\text{PMI}(a,b) = log\frac{P(a,b)}{P(a) * P(b)} \quad (1)$$

Knowledge-based approaches for computing word similarity exploit the contents of a dictionary [8] or lexical knowledge-base (LKB), often WordNet [9], a resource where synonyms are grouped together in synsets and semantic relations (e.g., hypernymy, part-of) are held between synsets. Typical measures consider words used in the definition or example sentences, as well as the semantic connections between words.

Distributional word representations consider how language is used, while LKBs are more theoretical and often based on the work of lexicographers. In the former, several types of relation are present, but not explicit, while in LKBs semantic relations are explicit, but limited to a small set of types. For instance, despite the presence of a few other relations, WordNet is mainly focused on synonymy and hypernymy. This is why it is better-suited to measure similarity, but is outperformed by corpus-based approaches when it comes to measuring relatedness. Some authors have thus adopted hybrid approaches, where distributional and knowledge-based approaches are combined [10,11].

State of the art results for well-known English similarity tests can be found in the ACL Wiki (https://www.aclweb.org/aclwiki/index.php?title=Similarity_(State_of_the_art)). For Portuguese, however, research in the area is quite recent, because resources (e.g., LKBs and word vectors) and benchmarks, are only becoming available in recent years.

Related work for Portuguese has tackled mostly paraphrasing [12] or semantic textual similarity [13], which consists of computing the similarity of larger units of text. A shared task was recently organised on the latter [2]. In this case, supervised approaches typically perform better, by exploiting several features, including semantic features that might involve word similarity measures. In addition to semantic textual similarity, Portuguese word embeddings have also been recently used for part-of-speech tagging and to solve analogies [14].

When it comes to word similarity, Granada et al. [15] created a Portuguese test for this purpose, compared the judge agreement with other languages, and applied a distributional approach, based on Wikipedia, to answer it. Wilkens et al. [16] compiled the B^{2}SG test and used it to assess distributional similarity measures. However, B^{2}SG is slightly different from the tests used in this work because, instead of a similarity score, a related word is to be selected from a group where it is shuffled with distractors.

3. Benchmarks for Portuguese Word Similarity

Four Portuguese word similarity tests were used as benchmarks in the experiments reported in this paper. One is based on the RG-65 similarity test [17], which contains 65 pairs of nouns and their similarity of meaning, between 0 and 4, computed from 51 human judgements. RG-65 was translated to Portuguese by Granada et al. [15] and similarity was re-scored from the opinions of 50 native speakers of Portuguese. It was renamed to PT-65 and is freely available online, from http://www.inf.pucrs.br/linatural/wikimodels/similarity.html.

The other three tests–SimLex-999, WordSim-353 and RareWords–were also created originally for English, but recently adapted to Portuguese, having in mind their utilisation for assessing models of distributional similarity in this language [18]. They are also available online, from http://metashare.metanet4u.eu/.

SimLex-999 [19] contains 999 word pairs (666 noun-noun, 222 verb-verb, 111 adjective-adjective) and their similarity score, based on the opinion of approximately 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. Although RG-65 also targets similarity, as far as we know, the process of differentiating similarity and relatedness was much more thorough in SimLex-999.

WordSim-353 [20] contains 353 word pairs and their relatedness score from 0 to 10, based on the judgement of 13 to 16 human judges. The original WordSim-353 was later manually split by Aguirre et al. [21] into similar and related pairs. For this purpose, they identified the semantic relation between the words of the pair, and split them into three sets: similar (synonyms, antonyms, identical, or hyponym-hyperonym); related (meronym-holonym, and none of the previous relations but average similarity higher than 5), and unrelated (the remaining pairs). Although the relation names are available in the Portuguese version, they were not used in this work.

RareWords [22] contains 2034 word pairs. The first words of each pair are rare, in the sense that they occur only 5000 to 10,000 times in the English Wikipedia, while the second words are related to the first in the widely-used LKB Princeton WordNet [23].

Tables 1 and 2 illustrate the contents of the Portuguese SimLex-999 and WordSim-353, respectively. The other two tests have a similar format, but each line contains no additional columns besides the words and the score.

Table 1. First two adjectives, nouns and verbs of the Portuguese SimLex-999, with original English translation.

Word#1	Word#2	POS	Similarity
velho (old)	*novo* (new)	A	0.00
esperto (smart)	*inteligente* (intelligent)	A	8.33
esposa (wife)	*marido* (husband)	N	5.00
livro (book)	*texto* (text)	N	5.00
ir (go)	*vir* (come)	V	3.33
levar (take)	*roubar* (steal)	V	6.67

Table 2. First five lines of the Portuguese WordSim-353 test with original English translation. Identified relations are: synonyms (S), identical tokens (i), second is part of the first (M).

Relation	Word#1	Word#2	Relatedness
S	*amor* (love)	*sexo* (sex)	6.77
S	*tigre* (tiger)	*gato* (cat)	7.35
i	*tigre* (tiger)	*tigre* (tiger)	10.00
M	*livro* (book)	*papel* (paper)	7.46
M	*computador* (computer)	*teclado* (keyboard)	7.62

4. LKBs and Knowledge-Based Approaches for Portuguese Word Similarity

The problem of computing word similarity was first tackled with the help of available lexical knowledge bases (LKBs) for Portuguese. LKBs are models of the mental lexicon that organise words of a language according to their meaning. The automatic procedures used here either exploit the structure of a wordnet (words, synsets, semantic relations) or other lexical networks that connect words by means of semantic relations, though without sense information. This section presents the resources of both kinds used and describes the procedures applied in their exploitation.

4.1. Wordnets

Princeton WordNet [23] is the paradigmatic LKB, for English, with a model also adapted to many other languages, including Portuguese, for which there are currently six wordnets [24]. Wordnets are structured in groups of synonymous words (synsets), which can be viewed as the possible lexicalisations of concepts, and semantic relations between synsets, including hypernymy, part-of, and possibly others. In this work, four open Portuguese wordnets were used, namely:

- OpenWordNet-PT (OWN.PT) [25], a Brazilian Wordnet aligned with Princeton WordNet, in the scope of the Open Multilingual WordNet project [26].

- PULO [27], a Portuguese wordnet, part of the Multilingual Central Repository [28], where wordnets of the Iberian languages are aligned.
- TeP [29], a wordnet for Brazilian with only synsets and antonymy relations.
- CONTO.PT [30], a fuzzy wordnet, created automatically from ten lexical resources for Portuguese, including the previous, and slightly different from the other wordnets, because words have variable (fuzzy) memberships to synsets, and synset connections have also got a variable number assigned, both based on their computed confidence.

Table 3 characterises the previous resources in terms of the number of words covered, number of synsets, and number of relation instances. These numbers were computed after conversion to a common format where minor incompatibilities were handled.

Table 3. Data of the wordnets used.

Wordnet	#Words	#Synsets	#Instances
OWN-PT	47,866	38,646	36,837
PULO	14,122	12,036	15,349
TeP	44,226	19,824	0
CONTO.PT	101,446	34,908	110,696

Experiments were also made with wordnet-oriented procedures on WordNet.Br [31] and OpenThesaurus.PT (available from http://paginas.fe.up.pt/~arocha/AED1/0607/trabalhos/thesaurus.txt) (OT.PT), but their results are not presented because the latter is very small and the former only covers verbs, while the majority of word pairs in the target tests are nouns.

In order to compute the similarity of two words, a and b, according to a fuzzy wordnet, one of the following procedures is applied:

- Memberships (μ): if there is a synset S with both a and b, $sim(a,b) = \mu_S(a) + \mu_S(b)$; if there are two synsets S_a and S_b, such that $a \in S_a \wedge b \in S_b$, and a relation instance R, such that $R = (S_a \text{ related-to } S_b)$, $sim(a,b) = \mu_{S_a}(a) * \mu_{S_b}(b) * \nu(R)$; otherwise, $sim(a,b) = 0$;
- Weighted adjacencies intersection: each word is represented by a vector, $\vec{a}$ and $\vec{b}$, where its adjacencies are weighted based on the memberships to the same synsets and confidence of direct connections. After this, $sim(a,b) = cos(\vec{a},\vec{b})$.
- Adjacencies intersection (Adj-Int): similar to the previous, but $\vec{a}$ and $\vec{b}$ are binary vectors, with 1, if the words are adjacent (no matter their weight) or 0, if they are not.

The previous procedures are applicable to the simple wordnets if all words w are considered to have a membership of 1 to each synsets they are included in: $\forall w \in S \rightarrow \mu_S(w) = 1$. The same is done for the semantic relation instances. For this reason, the second and the third procedures will be equivalent when OWN-PT, PULO or TeP are used.

4.2. Word-Based Semantic Networks

A simplified model of a LKB does not handle sense division and simply contains instances of the kind "*x related-to y*", where x and y are lexical items (nodes) and *related-to* is the name of a semantic relation (link). In this work, three kinds of such resources were used: (i) some already available in this format; (ii) some acquired from the open wordnets; (iii) others obtained by combining all or some resources of this kind.

For Portuguese, there are several word-based semantic networks, some extracted automatically from dictionaries, namely:

- PAPEL [32], extracted by exploiting the regularities in the definitions of a comercial Portuguese dictionary;

- Relations extracted from Dicionário Aberto (DA) [33] through a similar procedure as PAPEL;
- Relations extracted from Wiktionary.PT (http://pt.wiktionary.org), in this case, from the 2015 dump, through a similar procedure as PAPEL;
- Semantic relations available in the scope of Port4Nooj [34], a set of linguistic resources;
- Semantic relations held between two Portuguese terms in ConceptNet [11], a common-sense knowledge base.

The previous resources cover similar relation types, including synonymy, hypernymy, part-of, causation, purpose-of, property-of and location-of.

Resources with a similar format can be directly acquired from the available wordnets. For this purpose, synsets have to be deconstructed. For example, the instance {*porta, portão*} partOf {*automóvel, carro, viatura*} resulted in: (*porta* synonymOf *portão*), (*automóvel* synonymOf *carro*), (*automóvel* synonymOf *viatura*), (*carro* synonymOf *viatura*), (*porta* partOf *automóvel*), (*porta* partOf *carro*), (*porta* partOf *viatura*), (*portão* partOf *automóvel*), (*portão* partOf *carro*), (*portão* partOf *viatura*). Semantic networks were acquired this way from: OWN-PT, PULO, TeP and also OpenThesaurus.PT (OT.PT).

Finally, in order to analyse the benefits of combining the previous LKBs, they were combined in three additional semantic networks: (i) one with the relation instances in all LKBs used (All-LKB); (ii) one with only the relation instances in at least two of them (Redun2); (iii) and a network with all the relation instances extracted from dictionaries (CARTÃO = PAPEL+DA+Wikt.PT), for historical purposes [35].

Table 4 presents the LKBs used and their sizes, in terms of the number of distinct words and relation instances. It is clear that there are LKBs with significantly different sizes. A deeper analysis into the content and redundancy of these LKBs, and on the creation of redundancy-based LKBs, is found elsewhere [36].

Table 4. Size of the LKBs used in terms of words and relation instances.

LKB	#Words	#Instances
PAPEL	94,165	191,497
DA	95,188	139,404
Wikt.PT	45,345	80,071
Port4Nooj	12,641	20,340
ConceptNet	42,323	80,917
OWN-PT	40,940	151,731
PULO	12,135	154,906
TeP	40,499	480,932
OT	12,782	51,410
All-LKB	201,829	850,594
Redun2	58,192	150,952
CARTÃO	149,818	327,405

In order to compute word similarity from these semantic networks, two different algorithms were applied:

- Similarity of the adjacencies of each word in the target network, i.e., directly-connected words, computed with the Jaccard coefficient (Adj-Jac) and the cosine similarity (Adj-Cos);
- PageRank vectors, inspired by Pilehvar et al. [37]. For each word of a pair, Personalized PageRank is first run in the target LKB, for 30 iterations, using the word as context, and a vector is created with the resulting rank of each other word in the LKB. The similarity between the vectors of the two words is computed with the Jaccard coefficient between the sets of words in these vectors (PR-Jac) and with the cosine of the vectors (PR-Cos). Due to their large sizes, vectors were trimmed to the top $-N$ ranked words. Different sizes N were tested, from 50, to 3200.

The previous methods were tested with two different configurations: using all the relations of each LKB (All), or only synonymy and hypernymy relations (Syn+Hyp).

5. Distributional Models for Portuguese Word Similarity

Following the trend of using distributional models of words for computing similarity, models of this kind were used in this work. Besides testing some of the word embeddings that became available for Portuguese, PMI, a long-established measure, was also used.

5.1. Pointwise Mutual Information

Among other tasks, PMI has been used for mining synonyms from the Web [7] or for measuring the coherence of topic models [38]. Although word co-occurrence may consider words in the same window or sentence, in the latter case, PMI was computed from the presence of the words in Wikipedia articles, according to Equation (2).

$$\text{PMI}(a,b) = log\frac{Hits(a \cap b)}{Hits(a) * Hits(b)} \tag{2}$$

Following the previous idea, in this work, PMI was computed on the Portuguese Wikipedia, using its REST API. The result of the PMI measure was then normalised according to Bouma [39] (see Equation (3)).

$$\text{NPMI}(a,b) = \frac{PMI(a,b)}{-logHits(a \cap b)} \tag{3}$$

PMI was computed on the Portuguese Wikipedia, on March 2017. As Wikipedia was accessed remotely, through a REST API, we are not aware of the vocabulary size for this model.

5.2. Word Embeddings

Following the recent trend in using them for NLP, word embeddings have become available for Portuguese. Even though this kind of vector-based model has the disadvantage of not containing labelled semantic relations, it enables to compute the similarity of two words, a and b, as straightforward as computing the cosine of their vectors: $sim(a,b) = cos(\vec{a},\vec{b})$. This adds to the fact that there are several models available of this kind, learned from very large corpora, including some for Portuguese.

In this work, two collections of Portuguese embeddings were identified, namely: LX-DSemVectors [40], a collection of word2vec skip-gram models, trained with different parameters, and available through GitHub (https://github.com/nlx-group/lx-dsemvectors/); and NILC word embeddings [14], a collection of embeddings trained with different methods, and available on their own website (http://www.nilc.icmc.usp.br/nilc/index.php/repositorio-de-word-embeddings-do-nilc). From LX-DSemVectors, two models were selected: the vanilla LX model, trained with the default parameters, and LX model 17 (LX-p17), which has shown to be the most accurate [40]. While all the LX models were trained with word2vec, though with different parameters, there are NILC embeddings trained with different kinds of model. Besides the widely used word2vec, the models include: GloVE [5]; fastText [6], based on character n-grams, which can still be used to obtain word representations, especially relevant for morphologically rich languages; or Wang2vec [41], a modification of word2vec that considers word order. From those, a varied selection was made, according to best models in reported experiments [14] and in order to include models of different kinds: GloVE with size 300 (NILC-GloVE-300) and 600 (NILC-GloVE-600), fastText skip-gram with size 300 (NILC-fastText-sg300), word2vec CBOW with size 600 (NILC-word2vec-cb300) and Wang2vec skip-gram with size 600 (NILC-Wang2vec-sg600).

In addition to those previously mentioned, two other promising word embeddings for Portuguese were used, namely: the fastText Portuguese model by Facebook Research [6] (Facebook-fastText), trained in the Portuguese Wikipedia and available from GitHub (https://github.com/facebookresearch/

fastText/blob/master/pretrained-vectors.md); and the vectors of the Portuguese words picked from the ConceptNet Numberbatch model [11], which combines data from ConceptNet, word2vec, GloVE, and OpenSubtitles 2016. Also available from GitHub (https://github.com/commonsense/conceptnet-numberbatch), the latter model was recently used to achieve the best results in semantic similarity tasks for different languages, though not including Portuguese. While Numberbatch should be seen as a hybrid between a knowledge-based and a distributional model, we used it together with the other distributional models because it is available as a single resource and shares the word embedding vector representation of all the other distributional models but PMI.

Table 5 lists the embeddings used, together with their vocabulary size, vector size and kind of model. In terms of vocabulary, the NILC embeddings are the largest. On the other end, Numberbatch vocabulary is substantially smaller than all the others. This is however misleading, not only because just Portuguese words were selected but, mainly, because these words are all in the lemma form, which does not happen for the other embeddings, and compresses the vocabulary significantly. The main drawback is that words have to be lemmatised before the retrieval of their vector, but this is not necessary in the similarity tests, where words are already in their lemma form.

Table 5. Word embeddings used and their properties.

Resource	Vocabulary Size	Vector Size	Model
LX-vanilla	498,339	300	word2vec Skip-Gram
LX-p17	873,909	300	word2vec Skip-Gram
NILC-GloVE-300	934,963	300	GloVE
NILC-GloVE-600	934,963	600	GloVE
NILC-fastText-sg300	934,963	300	fastText Skip-Gram
NILC-word2vec-cb300	934,966	300	word2vec CBOW
NILC-Wang2vec-sg600	934,966	600	Wang2vec Skip-Gram
Facebook-fastText	592,108	300	fastText Skip-Gram
ConceptNet-Numberbatch	47,592	300	word2vec + GloVE

6. Results

This section presents and discusses the results obtained for the four Portuguese word similarity tests. As usual in this kind of tests, performance is assessed with the Spearman correlation (ρ) between the automatically-computed similarities and the gold ones in the test.

The section starts with the results using LKBs and then using the distributional models. After that, the global results are discussed and compared with the state-of-start results for English, for which the original versions of the target tests have been extensively used as benchmarks. A brief error analysis is then presented and discussed. Finally, following the trend of the best performing approaches for English, our results are further improved for three of the four tests, when the scores of the best knowledge-based approaches are combined with the best distributional models.

6.1. Results for LKBs

The selected knowledge-based approaches were adopted to answer the four similarity tests automatically. Yet, before computing similarities, the coverage of each similarity test by each LKB was analysed. Table 6 shows those numbers considering that, in order to cover a pair of words in a test, a LKB must include both words of this pair. The main conclusion is that different LKBs have a significantly different coverage of different tests. TeP is the only original LKB to cover all the pairs of a test (PT-65, the smallest one). Three of the combined LKBs (All-LKB, CARTÃO and CONTO.PT) also cover all the pairs of the latter test. Although no other test has all the pairs covered by a LKB, as expected, the highest coverages are those by the combined LKBs, always above 90% for PT-65, SimLex-999 and WordSim-353, with a single exception (Redun2 in WordSim-353). Again, as expected, coverage of the RareWords test is considerably lower, though always higher than 50% for the combined

LKBs, while only three of the original LKBs have more than 50% coverage for this test (TeP, PAPEL and DA). It is also clear that OT is not only the smallest resource, but has also the lowest coverages for every test.

In all the following experiments, similarity was set to 0 for every pair not covered by a LKB. Of course this will have a relevant impact on the results, but we believe this is the fairest way of comparing the performance of the LKB on these tests. In addition, we recall that coverage is one of the most important features to consider when selecting a LKB to use.

Table 6. Coverage of the pairs of each similarity test by each LKB used.

Resource	Covered Pairs			
	PT-65	**SimLex-999**	**WordSim-353**	**RareWords**
PAPEL	58 (89%)	920 (92%)	301 (86%)	1205 (59%)
DA	58 (89%)	878 (88%)	267 (76%)	1010 (50%)
Wikt.PT	59 (91%)	932 (93%)	298 (85%)	869 (43%)
Port4Nooj	37 (57%)	529 (53%)	193 (55%)	382 (19%)
ConceptNet	55 (85%)	903 (90%)	278 (79%)	728 (36%)
OWN-PT	55 (85%)	947 (95%)	292 (83%)	829 (41%)
PULO	48 (74%)	965 (97%)	297 (84%)	888 (44%)
TeP	65 (100%)	976 (98%)	320 (91%)	1300 (64%)
OT	18 (28%)	451 (45%)	86 (24%)	461 (23%)
All-LKB	65 (100%)	981 (98%)	335 (95%)	1453 (71%)
Redun2	62 (95%)	945 (95%)	309 (88%)	1125 (55%)
CARTÃO	65 (100%)	959 (96%)	321 (91%)	1323 (65%)
CONTO.PT	65 (100%)	964 (96%)	320 (91%)	1282 (63%)

After analysing coverage, the results obtained when computing word similarity with the LKBs are presented. Tables 7–10 show a selection of those results, respectively for PT-65, Simlex-999, WordSim-353 and RareWords. For each test, we present the top 5 or 6 results, depending on their differences, and the best results for each other resource not in the top 5. Besides the identification of each LKB, the relation set and the algorithm used are revealed for each result, plus the mean Spearman correlation ($\bar{\rho}$) and the standard deviation (σ). The latter were computed as suggested by Batchkarov et al. [42], who criticise how word similarity tests are used for assessing similarity models. What happens is that even the largest test, RareWords, is too small for taking conclusions about the performance of a broad-coverage resource, that aims at covering the whole language. Therefore, considering that human annotation of more word pairs is not possible, in order to better understand the variance of the applied methods, 500 random samples of the same test were created, which enabled the computation of the mean and standard deviation.

The numbers show that using as much knowledge as possible leads to the best results. The All-LKB, which we recall, contains all the relation instances in all the other LKBs, clearly got the best performance in PT-65, SimLex-999 and WordSim-353, with second best in the RareWords tests, where CONTO.PT achieved the best $\bar{\rho}$. We recall that, though with a significantly different structure, CONTO.PT is also a combination of all the other LKBs. A possible explanation for the latter result is that the words in RareWords are less frequent, probably left out in smaller resources, but covered in larger resources and, in the case of CONTO.PT, also grouped with similar words.

Besides CONTO.PT and All-LKB, the other combined LKBs, with contents from more than a single original LKB, are CARTÃO and Redun2, both in the first half of the rank for the majority of the tests. The main exception is Redun2 which, in WordSim-353, is below this margin, possibly because, when considering only instances in two or more LKBs, important knowledge for computing word relatedness is discarded.

Table 7. Performance overview of LKBs in the PT-65 test.

Resource	Relations	Algorithm	$\bar{\rho}$	σ
All-LKB	All	PR-Cos$_{3200}$	**0.87**	0.03
All-LKB	All	PR-Cos$_{1600}$	**0.87**	0.03
All-LKB	All	PR-Cos$_{800}$	0.86	0.04
All-LKB	All	PR-Cos$_{400}$	0.86	0.04
All-LKB	All	Adj-Cos	0.86	0.04
CARTÃO	All	Adj+Cos	0.78	0.06
Redun2	Syn+Hyp	Adj-Cos	0.77	0.06
CONTO.PT	All	μ	0.74	0.05
ConceptNet	All	Adj-Cos	0.72	0.07
OWN-PT	Syn+Hyp	Adj-Cos	0.68	0.07
PAPEL	All	Adj-Cos	0.65	0.07
DA	All	Adj-Cos	0.63	0.07
TeP	All	Adj-Cos	0.62	0.07
Wikc.PT	Syn+Hyp	Adj-Jac	0.56	0.09
PULO	All	μ	0.51	0.08
OT	All	Adj-Int	0.42	0.08
Port4Nooj	Syn+Hyp	Adj-Cos	0.38	0.11

Best $\bar{\rho}$ is in bold for each test.

Table 8. Performance overview of LKBs in the SimLex-999 test.

Resource	Relations	Algorithm	$\bar{\rho}$	σ
All-LKB	Syn+Hyp	PR-Cos$_{400}$	**0.61**	0.02
All-LKB	Syn+Hyp	PR-Cos$_{800}$	**0.61**	0.02
All-LKB	Syn+Hyp	PR-Cos$_{200}$	**0.61**	0.02
All-LKB	Syn+Hyp	PR-Cos$_{1600}$	0.60	0.02
All-LKB	Syn+Hyp	PR-Cos$_{3200}$	0.60	0.02
All-LKB	Syn+Hyp	Adj-Cos	0.58	0.02
CARTÃO	Syn+Hyp	PR-Jac$_{1600}$	0.53	0.02
Redun2	Syn+Hyp	PR-Jac$_{50}$	0.49	0.03
PAPEL	All	PR-Jac$_{800}$	0.48	0.03
CONTO.PT	Syn+Hyp	μ	0.47	0.03
OWN-PT	Syn+Hyp	Adj-Int	0.44	0.03
ConceptNet	Syn+Hyp	Adj-Cos	0.43	0.03
Wikt.PT	All	PR-Jac$_{1600}$	0.42	0.03
PULO	Syn+Hyp	μ	0.41	0.03
DA	All	PR-Jac$_{400}$	0.38	0.03
TeP	Syn+Hyp	Adj-Jac	0.36	0.03
OT.PT	Syn+Hyp	Adj-Cos	0.34	0.03
Port4Nooj	All	Adj-Jac	0.19	0.03

Best $\bar{\rho}$ is in bold for each test.

On the best algorithm to use, the PageRank vectors got the highest performance in the first three tests. The size of the vectors did not seem to play a huge role, as the differences are not significant for different sizes. In order to further analyse this issue, Figures 1 and 2 show the evolution of the $\bar{\rho}$ for different LKBs, with different vector sizes and similarity computed with the cosine. At a first glance, the $\bar{\rho}$ of the word-based networks extracted from wordnets gets lower for larger vectors, while, for the other LKBs, performance is improved, at least until vectors of size 200 or 400, depending on the LKB. A possible explanation is that the synset deconstruction procedure used in the extraction of the former networks might result in undesired ambiguity, which makes those networks noisier. Since the PageRank vectors algorithm has a high time complexity, especially for large networks, we also show the best result with a simpler algorithm, Adj-Cos or Adj-Jac, which is within the range of one standard

deviation in PT-65 and WordSim-353, less than two in SimLex-999, and achieved the best performance of the All-LKB in the RareWords test.

Table 9. Performance overview of LKBs in the WordSim-353 test.

Resource	Relations	Algorithm	$\bar{\rho}$	σ
All-LKB	All	PR-Cos$_{200}$	**0.46**	0.05
All-LKB	All	PR-Cos$_{3200}$	**0.46**	0.05
All-LKB	All	PR-Cos$_{1600}$	0.45	0.05
All-LKB	All	PR-Cos$_{800}$	0.45	0.05
All-LKB	All	PR-Cos$_{400}$	0.45	0.05
All-LKB	All	Adj-Cos	0.44	0.05
OWN-PT	All	Adj-Cos	0.37	0.05
ConceptNet	All	Adj-Cos	0.34	0.05
CARTÃO	All	Adj-Cos	0.33	0.05
CONTO.PT	All	μ	0.30	0.05
PULO	Syn+Hyp	μ	0.29	0.04
PAPEL	Syn+Hyp	Adj-Cos	0.29	0.05
Redun2	All	Adj-Cos	0.28	0.05
Port4Nooj	Syn+Hyp	Adj-Cos	0.28	0.05
DA	All	Adj-Jac	0.27	0.05
Wikt.PT	All	Adj-Jac	0.26	0.05
TeP	All	μ	0.22	0.05
OT.PT	Syn+Hyp	Adj-Cos	0.15	0.06

Best $\bar{\rho}$ is in bold for each test.

Table 10. Performance overview of LKBs in the RareWords test.

Resource	Relations	Algorithm	$\bar{\rho}$	σ
CONTO.PT	All	μ	**0.41**	0.02
All-LKB	All	Adj-Jac	0.38	0.02
TeP	All	Adj-Cos	0.38	0.02
TeP	All	Adj-Jac	0.38	0.02
TeP	All	Adj-Int	0.38	0.02
CARTÃO	All	Adj-Jac	0.36	0.02
Redun2	All	Adj-Jac	0.34	0.02
PAPEL	All	Adj-Jac	0.33	0.02
OWN-PT	All	Adj-Int	0.30	0.02
Wikt.PT	All	Adj-Jac	0.29	0.02
DA	Syn-Hyp	Adj-Jac	0.28	0.02
ConceptNet	All	Adj-Cos	0.28	0.02
PULO	All	Adj-Int	0.28	0.02
OT.PT	Syn+Hyp	Adj-Cos	0.27	0.02
Port4Nooj	All	Adj-Cos	0.07	0.02

Best $\bar{\rho}$ is in bold for each test.

On the selected relation types, for most LKBs, using only synonymy and hypernymy relations resulted in best results in SimLex-999, which makes sense because these are also the most relevant relations for computing genuine similarity. For the other tests, the majority of the best results were achieved using all the relations, also relevant when computing relatedness. This trend is less clear in PT-65, where there might be some confusion between the concepts of similarity and relatedness.

Looking at individual LKBs, PAPEL, OWN-PT and ConceptNet performed generally well, when compared to the others. Although with modest performances in the first three tests, TeP was within two standard deviations to the best performance in the RareWords dataset. Again, although TeP only covers synonymy and antonymy, it covers a large number of words and, when the synsets are deconstructed, it becomes the second LKB with more relation instances, which might have played a positive role.

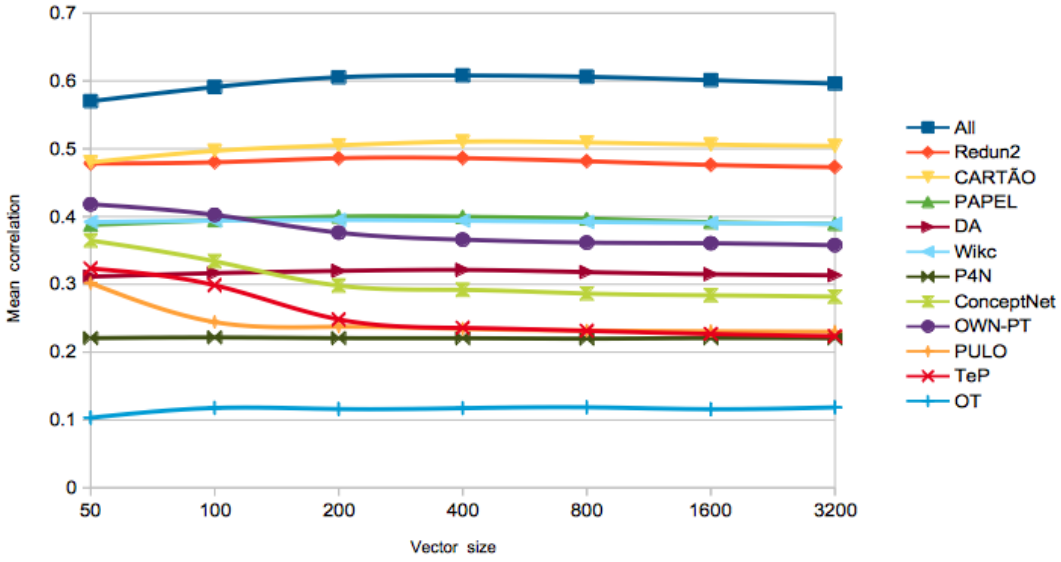

Figure 1. Evolution of $\bar{\rho}$ in SimLex-999, for different LKBs, using only synonymy and hypernymy relations, with growing vector sizes in PageRank.

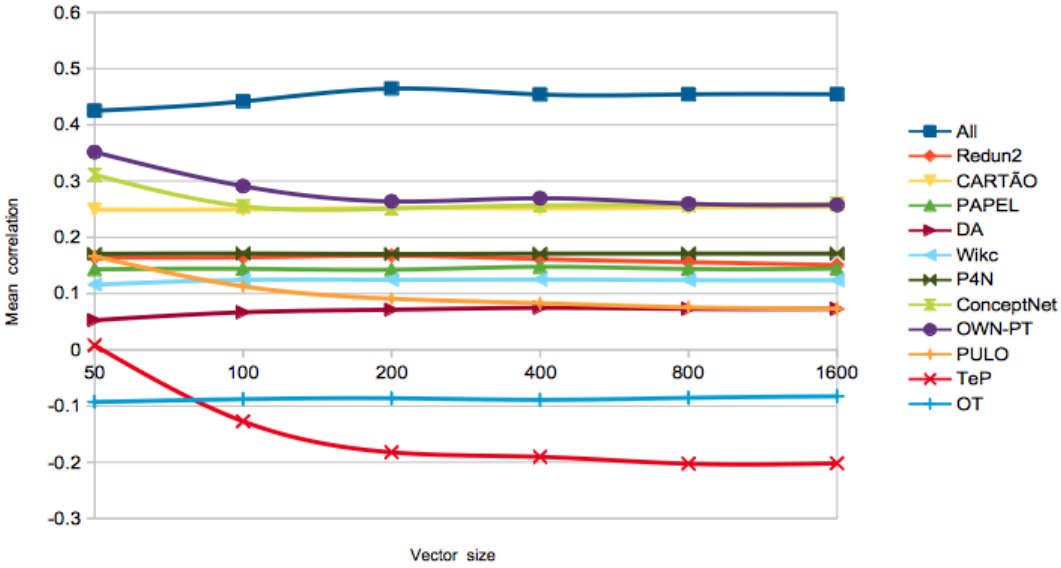

Figure 2. Evolution of $\bar{\rho}$ in WordSim-353, for different LKBs, using all relations, with growing vector sizes in PageRank.

A remark should be given on the presented results, which are not only influenced by the contents and structure of the LKBs, but also by the applied algorithms. Therefore, although they provide useful hints, our results are only valid for the tested algorithms, and better results could possibly be obtained by alternative ways of exploiting the LKBs.

6.2. Results for Distributional Models

Similarly to the experiments with the LKBs, the coverage of the tests by the distributional models was also analysed. Results, in Table 11, show that most coverages are very high, for all the tests, which makes sense because the embeddings are learned from large collections of text and PMI is computed in the full Portuguese Wikipedia. As expected, coverages are lower for the RareWords test, but still higher than 80%, except for Numberbatch. The latter is probably due to the lower coverage of Portuguese words by ConceptNet, with a higher impact in less frequent (rare) words.

Table 11. Coverage of each dataset by each distributional model.

Resource	PT-65	SimLex-999	WordSim-353	RareWords
LX-vanilla	65 (100%)	966 (97%)	345 (98%)	1624 (80%)
LX-p17	65 (100%)	974 (97%)	345 (98%)	1767 (87%)
NILC-GloVE-300	65 (100%)	975 (98%)	328 (93%)	1810 (89%)
NILC-GloVE-600	65 (100%)	975 (98%)	328 (93%)	1810 (89%)
NILC-fastText-sg300	65 (100%)	975 (98%)	328 (93%)	1810 (89%)
NILC-word2vec-cb600	65 (100%)	975 (98%)	328 (93%)	1810 (89%)
NILC-Wang2vec-sg600	65 (100%)	975 (98%)	328 (93%)	1810 (89%)
Facebook-fastText	64 (98%)	963 (96%)	325 (92%)	1663 (82%)
Numberbatch	64 (98%)	962 (96%)	318 (90%)	958 (47%)
PMI	65 (100%)	997 (≈100%)	353 (100%)	1962 (96%)

With the distributional models, computing similarity of two words is just a matter of computing the cosine of their vectors, for the embeddings, or their PMI, which means that there was no variation in the algorithm used. The obtained mean Spearman correlation ($\bar{\rho}$) and the standard deviation (σ), computed on the same 500 random samples as for the LKBs, are presented in Table 12 for each distributional model and similarity test. The models are ordered according to their performance in the PT-65 test, but the best results of each test are in bold.

Table 12. Spearman correlation (ρ) for the distributional approaches in the different tests.

Resource	PT-65		SimLex-999		WordSim-353		RareWords	
	$\bar{\rho}$	σ	$\bar{\rho}$	σ	$\bar{\rho}$	σ	$\bar{\rho}$	σ
Numberbatch	**0.80**	0.06	**0.63**	0.02	**0.50**	0.05	0.31	0.02
NILC-fastText-sg300	0.77	0.06	0.33	0.03	0.41	0.05	**0.42**	0.02
NILC-Wang2vec-sg600	0.75	0.06	0.39	0.03	0.41	0.05	**0.42**	0.02
NILC-GloVE-600	0.74	0.07	0.30	0.03	0.30	0.05	0.37	0.02
Facebook-fastText	0.74	0.07	0.34	0.03	0.42	0.05	0.34	0.02
NILC-GloVE-300	0.73	0.07	0.30	0.03	0.31	0.05	0.38	0.02
LX-p17	0.65	0.08	0.33	0.03	**0.48**	0.05	0.35	0.02
PMI	0.65	0.08	0.22	0.03	**0.49**	0.04	0.28	0.02
NILC-word2vec-cb600	0.60	0.09	0.25	0.03	0.33	0.05	0.36	0.02
LX-vanilla	0.56	0.10	0.23	0.03	0.36	0.05	0.27	0.02

Best $\bar{\rho}$ is in bold for each test.

Numberbatch achieved the best results in the three first tests, though it was in SimLex-999 where it clearly outperformed the other models. Even though distributional models are typically better suited for computing relatedness and not genuine similarity, the underlying semantic network of ConceptNet might have played an important role here. On the other hand, Numberbatch performed much worse in RareWords. This is primarily caused by the lower coverage of the words pairs of this test by this model, as seen in Table 11. On the WordSim-353 test, the $\bar{\rho}$ obtained by PMI and by the LX-17 model are not significantly different from those of Numberbatch, which shows that those models are particularly well-suited for computing word relatedness. The second and third best models for PT-65 were respectively NILC-fastText-sg300 and NILC-Wang2vec-sg600, though within a single standard deviation from Numberbatch. They were also the models that achieved the best performance in the RareWords test, not only for the distributional models, but overall. As it happened for the LKBs, those are not only among the embeddings with a larger vocabulary, but are also the resource with the highest coverage of the word pairs in RareWords (89%).

6.3. Global Performance: Knowledge-Based vs. Distributional Approaches

Looking at the global results, ConceptNet Numberbatch achieved the best performance in two tests, SimLex-999 and WordSim-353, even though the best result of the All-LKB is within the range of Numberbatch's standard deviation, and that the differences in WordSim-353 are not significant from PMI and LX-P17. This is still quite surprising, because the aforementioned test sets are significantly different: one targets genuine similarity and the other relatedness. In general, the best approaches for SimLex-999 perform poorly in WordSim-353 and vice-versa. Even when the position in both tests is comparable, the type of relations considered is different–synonymy and hypernymy for SimLex-999 and all for WordSim-353. However, Numberbatch is able to combine the strengths of a lexical knowledge base–theoretical model of the mental lexicon, captures similarity better–with those of a distributional model–practical model of language, captures relatedness better.

Apart from this exception, the best results on SimLex-999 are obtained with LKBs. In fact, the majority of the LKBs performed better than all the distributional models on this test. On WordSim-353, results are more mixed, but the top-3 approaches rely on distributional models. In PT-65, the best approach used the All-LKB, followed by Numberbatch. Finally, in RareWords, the best performances were those of the NILC-fastText-sg300 and NILC-Wang2vec-sg600 models, but their result is not significantly higher than the result of the fuzzy wordnet CONTO.PT. Numberbatch did not perform so well here, especially due to its limited coverage of unfrequent words.

As a brief conclusion, either Numberbatch or the All-LKB are a good choice for one willing to compute genuine similarity in Portuguese. Regular distributional models are not the best choice for this task, but are better-suited for computing relatedness. Nevertheless, we can say that Numberbatch is probably the safest choice when it is not clear if the goal is to compute similarity or relatedness. This is only not valid when dealing with a large amount of unfrequent words. In this case, one should probably resort either to the largest distributional models or rely on the fuzzy synsets of CONTO.PT.

6.4. Comparison with State-of-the-Art

To the best of our knowledge, there are not published results for the Portuguese versions of SimLex-999, WordSim-353 nor RareWords, because they were only translated recently. On the other hand, our results for PT-65 clearly outperform experimental results by Granada et al. [15], who used a LSA distributional model based on the Portuguese Wikipedia and achieved $\rho = 0.53$, which is substantially lower than the 0.87 achieved with the All-LKB.

State-of-the-art results for the English versions of RG-65, SimLex-999 and WordSim-353 can be found in the ACL Wiki (https://aclweb.org/aclwiki/Similarity_(State_of_the_art), checked in November 2017). Although with different distances, all our results (so far) are below the best results for English, which are all quite recent. For RG-65, the best performance ($\rho = 0.920$) was achieved with a knowledge-based approach [37] that exploits Wiktionary [43], and is 4 points higher than our best. The best results for SimLex-999 ($\rho = 0.642$), obtained by combining distributional knowledge with WordNet [10], are not far from our results obtained with the All-LKB ($\rho = 0.61$) and are even closer to those obtained with the Portuguese Numberbatch ($\rho = 0.63$). The best results for WordSim-353 ($\rho = 0.828$) were obtained with hybrid approaches, including Numberbatch (word2vec, GloVe and ConceptNet) [11]. In this case, our best results were also obtained with Numberbatch, but with a significantly lower performance ($\rho = 0.5$). Although not in the ACL Wiki, a $\rho = 0.46$ was recently reported for the English RareWords test, using the fastText embeddings [6], which is again higher than the best results reported here ($\rho = 0.42$ for two of the NILC embeddings and $\rho = 0.41$ for CONTO.PT). As in our experiments, this is also the test with lower state-of-the-art results.

We believe that the main reason for the difference in the performances between Portuguese and English is that languages are different, which means that approaches in Portuguese had to resort to different resources than those of English. Moreover, this kind of research is very recent for Portuguese, while there has been much work on this topic for English. On top of this, there might be translation issues. While at the conceptual level, similarity scores should be the same, word meanings might "shift"

after translation. However, analysing the quality of the translations and the scores of the tests, which were not created by us, is out of the scope of this paper. Some issues are, nevertheless, presented in the following section. Yet, from our superficial observation, and given the size of tests, we believe that they should have no more than a residual impact on the final results.

6.5. Error Analysis

In order to analyse where the best approaches were still failing and could possibly be improved, an error analysis was conducted. For this purpose, the word pairs of each test were ranked, based on their gold similarity scores, and compared to the ranking of the same pairs after the automatic computation of the similarity scores. The pairs with a higher difference between these rankings were analysed for a selection of approaches in the four similarity tests. The difference of the similarity scores was not considered because the scales used by each test and by different approaches is also different.

Tables 13–16 show the previous analysis, respectively for each similarity test. Several pairs are presented, together with their gold rank, their rank in the best approaches of each kind (knowledge-based and distributional) and the difference between the latter and the gold rank (Δ). The presented pairs are the top-5 with higher Δ for each test and approach. Some tables have less than five pairs per approach because some pairs are the same for different approaches.

PT-65 is the smallest test, where all or the majority of the word pairs are covered by the best approaches, so there is no clear reason for the problem with the presented pairs. The exception is *meio-dia*, not covered by Numberbatch, and thus put in a lower rank by this resource. About the other pairs in a lower rank, we can just say that they are not as strongly related in the exploited resources as they probably should.

In SimLex-999, several pairs include one multiword expression, with less probability of being covered by the exploited resources. In the English version of SimLex-999, these were not multiwords, but they became so after translation, because there was no suitable Portuguese singleword with exactly the same meaning, at least for those presented. A related issue results from a highly debatable option in the creation of the All-LKB, where prepositions and pronouns in the end of the multiwords were removed, for uniformisation purposes. This means that the All-LKB contains the multiword *ter_direito* (roughly, "have the right to") and several others matching *ter_direito_a_X*, where X is a verb, but not *ter_direito_a*. The same LKB does contain *reunir*, but not *reunir-se*. For *cama_de_bebé* the problem is different. The All-LKB covers this term, but written in the Brazilian way (*cama_de_bebê*). Another writing issue occurs for the pair {*visão*, *perceção*}. Although the All-LKB covers both words, visão is not directly related to *perceção*, but to *percepção*, a synonym written according to the old Portuguese norm. This analysis also showed that word embeddings do not deal very well with antonyms (namely {*levar*,*trazer*}, {*lembrar*, *esquecer*}, {*perder*, *ganhar*}, {*aceitar*, *rejeitar*}).

In WordSim-353, in addition to the multiword expressions, the problematic pairs for the LKBs involve named entities (e.g., Maradona, OPEC, Freud), which are not expected to be covered by LKBs, as the latter typically cover language and not world knowledge. For the distributional models, the issues with the presented pairs is less clear. There are multiword expressions and others, such as: again *meio-dia*, not covered by Numberbatch; a word that, in English is often used in the plural (clothes $\rightarrow$ *roupas*), but has a singular in Portuguese (*roupa*), which is its only form in Numberbatch; or the name of two drinks that are translated to Portuguese (vodka $\rightarrow$ vodca, brandy $\rightarrow$ brande), even though the translation is rarely used.

Finally, in RareWords, two of the problematic pairs for CONTO.PT include words that are covered but not related (*vil* and *cruel*; *constringir* and *adstringir*), and the other three include an uncovered word. Two of the latter are words (*repórteres* and *consensos*) that, for some reason, appear in the plural, though, as well as the other LKBs, CONTO.PT stores words in their lemma form.

Table 13. Word pairs with higher rank difference (Δ) in procedures using three different databases and gold score in PT-65.

Word#1	Word#2	Gold	All-LKB		Numberbatch		NILC-fastText	
		#	#	(Δ)	#	(Δ)	#	(Δ)
cálice (glass)	*mágico* (magician)	36	57	(21)	41	(5)	31	(5)
cemitério (graveyard)	*hospício* (madhouse)	40	20	(20)	36	(4)	18	(22)
costa (shore)	*viagem* (voyage)	29	48	(19)	26	(3)	49	(20)
risada (grin)	*rapaz* (lad)	43	62	(19)	54	(11)	29	(14)
pássaro (bird)	*bosque* (woodland)	25	43	(18)	23	(2)	24	(1)
meio-dia (midday)	*almoço* (noon)	20	19	(1)	58	(38)	14	(6)
pássaro (bird)	*grua* (crane)	61	47	(14)	28	(33)	53	(8)
grua (crane)	*instrumento* (implement)	24	21	(3)	51	(27)	59	(35)
litoral (coast)	*floresta* (forest)	32	35	(3)	59	(27)	33	(1)
rapaz (lad)	*bruxo* (wizard)	30	25	(5)	53	(23)	19	(11)
cálice (glass)	*taça* (tumbler)	4	16	(12)	8	(4)	36	(32)
autógrafo (autograph)	*assinatura* (signature)	11	11	(0)	20	(9)	38	(27)
carro (car)	*jornada* (journey)	31	39	(8)	52	(21)	58	(27)
almofada (cushion)	*bijuteria* (jewel)	53	53	(0)	46	(7)	27	(26)

Table 14. Word pairs with higher rank difference (Δ) in procedures using three different databases and gold score in SimLex-999.

Word#1	Word#2	Gold	All-LKB		Numberbatch		NILC-Wang2vec	
		#	#	(Δ)	#	(Δ)	#	(Δ)
cama_de_bebé (crib)	*berço* (cradle)	52	981	(929)	937	(885)	981	(929)
ter_direito_a (deserve)	*merecer* (earn)	62	990	(928)	961	(899)	993	(931)
juntar (gather)	*reunir-se* (meet)	108	987	(879)	958	(850)	474	(366)
visão (vision)	*perceção* (perception)	150	985	(835)	951	(801)	284	(134)
pensar (think)	*racionalizar* (rationalize)	103	903	(800)	997	(894)	680	(577)
levar (carry)	*trazer* (bring)	942	372	(570)	118	(824)	69	(873)
lembrar (remind)	*esquecer* (forget)	968	912	(56)	230	(738)	19	(949)
perder (lose)	*ganhar* (get)	976	632	(344)	430	(546)	40	(936)
aceitar (accept)	*rejeitar* (deny)	941	879	(62)	259	(682)	18	(923)

Table 15. Word pairs with higher rank difference (Δ) in procedures using three different databases and gold score in WordSim-353.

Word#1	Word#2	Gold	All-LKB		Numberbatch		PMI		LX-p17	
		#	#	(Δ)	#	(Δ)	#	(Δ)	#	(Δ)
homicídio (murder)	*homicídio_involuntário* (manslaughter)	23	350	(327)	329	(306)	8	(15)	351	(328)
computador (computer)	*programa_informático* (software)	25	348	(323)	324	(299)	113	(88)	348	(323)
Maradona	*futebol* (football)	20	337	(317)	301	(281)	96	(76)	67	(47)
OPEC	*petróleo* (oil)	21	329	(308)	323	(302)	44	(23)	64	(43)
psicologia (psychology)	*Freud*	40	343	(303)	311	(271)	14	(26)	71	(31)
meio_ambiente (environment)	*ecologia* (ecology)	16	233	(217)	328	(312)	20	(4)	350	(334)
meio-dia (midday)	*meio-dia* (noon)	3	3	(0)	308	(305)	3	(0)	3	(0)
tipo (type)	*género* (kind)	10	50	(40)	20	(10)	351	(341)	34	(24)
roupeiro (closet)	*roupas* (clothes)	57	113	(56)	103	(46)	315	(258)	96	(39)
vodca (vodka)	*brande* (brandy)	42	342	(300)	28	(14)	289	(247)	38	(4)
declaração (announcement)	*esforço* (effort)	311	279	(32)	261	(50)	68	(243)	261	(50)
meio-dia (noon)	*linha* (string)	349	280	(69)	309	(40)	111	(238)	340	(9)
pedra_preciosa (gem)	*joia* (jewel)	11	6	(5)	307	(296)	87	(76)	345	(334)
natureza (nature)	*meio_ambiente* (environment)	37	235	(198)	327	(290)	90	(53)	349	(312)

Table 16. Word pairs with higher rank difference (Δ) in procedures using three different databases and gold score in RareWords.

Word#1	Word#2	Gold	CONTO.PT		NILC-fastText		NILC-Wang2vec	
		#	#	(Δ)	#	(Δ)	#	(Δ)
vil (villainous)	*cruel* (wicked)	58	2015	(1957)	216	(158)	186	(128)
repórteres (reporters)	*repórter* (reporter)	7	1940	(1933)	120	(113)	86	(79)
desfavorecido (disadvantaged)	*desprivilegiado* (underprivileged)	57	1943	(1886)	335	(278)	323	(266)
constringir (constrict)	*adstringir* (astringe)	130	1985	(1855)	2027	(1897)	2027	(1897)
consensos (concurrencies)	*acordo* (agreement)	92	1918	(1826)	1240	(1148)	1141	(1049)
pouco_convincente (unconvincing)	*pouco_persuasivo* (unpersuasive)	89	1883	(1794)	2006	(1917)	2,006	(1917)
em_combustão (combusting)	*em_chamas* (ablaze)	53	1631	(1578)	1963	(1910)	1963	(1910)
incombustível (incombustible)	*à_prova_de_fogo* (fireproof)	88	1849	(1761)	1997	(1909)	1997	(1909)
inequívoco (unequivocal)	*não_ambíguo* (unambiguous)	52	1594	(1542)	1956	(1904)	1956	(1904)

6.6. Combining LKBs and Distributional Models

The state-of-the-art strongly suggests that benefits may arise from the combination of similarity models of different kinds. Therefore, although the results so far were very positive, we decided to investigate if they could be further improved. For this purpose, a simple experiment was performed, where the similarity score was computed from the average of two distinct approaches. More precisely, for each test, the scores of the best performing LKBs were combined with those of the best distributional models.

After this experimentation, no combination was found to outperform the All-LKB in PT-65. Combining the latter with NILC-Wang2vec-sg300 or Numberbatch achieves, respectively, $\bar{\rho} = 0.82 \pm 0.05$ or $\bar{\rho} = 0.81 \pm 0.05$, both lower than $\bar{\rho} = 0.87 \pm 0.03$.

On the other hand, interesting improvements were achieved for all the other tests. For SimLex-999, there was an improvement of up to 3 points on Numberbatch alone, when it is combined with the All-LKB, and minor improvements when combined with PAPEL and Redun2 (see Table 17). For WordSim-353, there was a substantial improvement of 10 points when the All-LKB is combined with PMI, but improvements are also obtained with other combinations (see Table 18). For RareWords, improvements of up to 5 points are obtained when combining the All-LKB, CONTO.PT or TeP with NILC-fastText-sg300, and slightly lower when the same LKBs are combined with NILC-Wang2vec-sg600 (see Table 19).

Table 17. Top Spearman correlations (ρ) for SimLex-999, based on the average of two different models.

Resources	$\bar{\rho}$	σ
All-LKB + Numberbatch	**0.66**	0.02
PAPEL + Numberbatch	0.64	0.02
Redun2 + Numberbatch	0.64	0.02
CARTÃO + Numberbatch	0.60	0.02

Best $\bar{\rho}$ is in bold for each test.

Table 18. Top Spearman correlations (ρ) for WordSim-353, based on the average of two different models.

Resources	$\bar{\rho}$	σ
All-LKB + PMI	**0.60**	0.04
OWN-PT + PMI	0.57	0.04
All-LKB + Numberbatch	0.57	0.04
ConceptNet + PMI	0.56	0.04
All-LKB + LX-p17	0.54	0.04

Best $\bar{\rho}$ is in bold for each test.

Table 19. Top Spearman correlation (ρ) for RareWords, based on the average of two different models.

Resources	$\bar{\rho}$	σ
All-LKB + NILC-fastText-sg300	**0.47**	0.02
CONTO.PT + NILC-fastText-sg300	0.46	0.02
TeP + NILC-fastText-sg300	0.46	0.02
All-LKB + NILC-Wang2vec-sg600	0.46	0.02
CONTO.PT + NILC-Wang2vec-sg600	0.45	0.02

Best $\bar{\rho}$ is in bold for each test.

7. Concluding Remarks

Several approaches for computing word similarity in Portuguese, using different available resources, were presented and compared. We believe that the presented results can be seen as a reference for future work on the development of new word models, LKBs, or algorithms for computing semantic similarity in Portuguese.

Results, assessed through word similarity tests that recently became available for Portuguese, confirm that there are several valid approaches for this purpose, but the best depends on the nature of the test. Nevertheless, some relevant conclusions were taken or confirmed. For instance, LKBs are better suited for computing genuine similarity, while distributional models suit better the computation of semantic relatedness [21]. Although the best results for genuine similarity were achieved by a distributional model, it is not a typical one (Numberbatch), as it also exploited a semantic network (ConceptNet) in the learning process. The latter model is probably the safest choice, unless there are many unfrequent words, because it has a limited coverage of those. In this case, either larger word embeddings or a fuzzy wordnet revealed to be a good option. Excluding the singularity of Numberbatch, better results are typically obtained with larger resources, due to their broader coverage.

Yet, the best results are obtained not only by combining knowledge from different sources, but also knowledge organised differently. In fact, as it happens for English, the combination of LKBs with distributional approaches was revealed to the best solution for computing word similarity in Portuguese. This was confirmed with a simple experiment, where similarity scores computed with different models were averaged and contributed equally to a combined score. Further work might focus on the issues identified with the presented error analysis. For instance, if an unlemmatised word is not covered by the semantic model, its lemma could be used instead. Or, even if a word is covered, semantic information on its inflections, typically available in the distributional models, could be combined with information on its lemma. On the other hand, some of the identified words that are not covered might suggest future additions or potential fixes on the creation of the LKBs used. It would as well be interesting to explore how the scores of two or more approaches could be weighted in order to achieve higher performances.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. In Proceedings of the Workshop track of the International Conference on Learning Representations (ICLR), Scottsdale, AZ, USA, 2–4 May 2013.
2. Fonseca, E.R.; dos Santos, L.B.; Criscuolo, M.; Aluísio, S.M. Visão Geral da Avaliação de Similaridade Semântica e Inferência Textual. *Linguamática* **2016**, *8*, 3–13.
3. Harris, Z. Distributional structure. *Word* **1954**, *10*, 146–162.
4. Turney, P.D.; Pantel, P. From Frequency to Meaning: Vector Space Models of Semantics. *J. Artif. Int. Res.* **2010**, *37*, 141–188.

5. Pennington, J.; Socher, R.; Manning, C.D. GloVe: Global Vectors for Word Representation. In Proceedings of the Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
6. Bojanowski, P.; Grave, E.; Joulin, A.; Mikolov, T. Enriching Word Vectors with Subword Information. *arXiv* **2016**, arXiv:1607.04606.
7. Turney, P.D. Mining the Web for Synonyms: PMI–IR versus LSA on TOEFL. In Proceedings of the 12th European Conference on Machine Learning, ECML 2001, Freiburg, Germany, 5–7 September 2001; Volume 2167, pp. 491–502.
8. Lesk, M. Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. In Proceedings of the 5th Annual International Conference on Systems Documentation (SIGDOC '86) , Toronto, ON, Canada, 8–11 June 1986; pp. 24–26.
9. Budanitsky, A.; Hirst, G. Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Comput. Linguist.* **2006**, *32*, 13–47.
10. Banjade, R.; Maharjan, N.; Niraula, N.B.; Rus, V.; Gautam, D. Lemon and Tea Are Not Similar: Measuring Word-to-Word Similarity by Combining Different Methods. In Proceedings of the 16th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2015) Part I, Cairo, Egypt, 14–20 April 2015; Volume 9041, pp. 335–346.
11. Speer, R.; Chin, J.; Havasi, C. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In Proceedings of the 31st AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; pp. 4444–4451.
12. Barreiro, A. ParaMT: A Paraphraser for Machine Translation. In Proceedings of the 8th International Conference on Computational Processing of the Portuguese Language (PROPOR 2008), Aveiro, Portugal, 8–10 September 2008; Volume 5190, pp. 202–211.
13. Pinheiro, V.; Furtado, V.; Albuquerque, A. Semantic Textual Similarity of Portuguese-Language Texts: An Approach Based on the Semantic Inferentialism Model. In Proceedings of the 11th International Conference on Computational Processing of the Portuguese Language (PROPOR 2014), São Carlos, Brazil, 6–8 October 2014; Volume 8775, pp. 183–188.
14. Hartmann, N.S.; Fonseca, E.R.; Shulby, C.D.; Treviso, M.V.; Rodrigues, J.S.; Aluísio, S.M. Portuguese Word Embeddings: Evaluating on Word Analogies and Natural Language Tasks. In Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology (STIL 2017), Uberlandia, Brazil, 2–4 October 2017.
15. Granada, R.; Trojahn, C.; Vieira, R. Comparing semantic relatedness between word pairs in Portuguese using Wikipedia. In Proceedings of the 11th International Conference on Computational Processing of the Portuguese Language (PROPOR), São Carlos, Brazil, 6–8 October 2014; Volume 8775, pp. 170–175.
16. Wilkens, R.; Zilio, L.; Ferreira, E.; Villavicencio, A. B^2SG: A TOEFL-like Task for Portuguese. In Proceedings of the 10th International Conference on Language Resources and Evaluation (ELRA), Portoroz, Slovenia, 23–28 May 2016.
17. Rubenstein, H.; Goodenough, J.B. Contextual Correlates of Synonymy. *Commun. ACM* **1965**, *8*, 627–633.
18. Querido, A.; Carvalho, R.; Rodrigues, J.; Garcia, M.; Silva, J.; Correia, C.; Rendeiro, N.; Pereira, R.; Campos, M.; Branco, A. LX-LR4DistSemEval: A collection of language resources for the evaluation of distributional semantic models of Portuguese. *Rev. Assoc. Port. Linguíst.* **2017**, 265–283, doi:10.26334/2183.
19. Hill, F.; Reichart, R.; Korhonen, A. Simlex-999: Evaluating Semantic Models with Genuine Similarity Estimation. *Comput. Linguist.* **2015**, *41*, 665–695.
20. Finkelstein, L.; Gabrilovich, E.; Matias, Y.; Rivlin, E.; Solan, Z.; Wolfman, G.; Ruppin, E. Placing Search in Context: The Concept Revisited. *ACM Trans. Inf. Syst.* **2002**, *20*, 116–131.
21. Agirre, E.; Alfonseca, E.; Hall, K.; Kravalova, J.; Paşca, M.; Soroa, A. A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches. In Proceedings of the Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Boulder, CO, USA, 1–3 June 2009; ACL Press: Stroudsburg, PA, USA, 2009; pp. 19–27.
22. Luong, T.; Socher, R.; Manning, C. Better Word Representations with Recursive Neural Networks for Morphology. In Proceedings of the Seventeenth Conference on Computational Natural Language Learning (CoNLL), Sofia, Bulgaria, 8–9 August 2013; pp. 104–113.

23. Fellbaum, C. (Ed.) *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*; The MIT Press: Cambridge, MA, USA, 1998.
24. De Paiva, V.; Real, L.; Gonçalo Oliveira, H.; Rademaker, A.; Freitas, C.; Simões, A. An overview of Portuguese Wordnets. In Proceedings of the 8th Global WordNet Conference (GWC'16), Bucharest, Romania, 27–30 January 2016; pp. 74–81.
25. De Paiva, V.; Rademaker, A.; de Melo, G. OpenWordNet-PT: An Open Brazilian WordNet for Reasoning. In Proceedings of the 24th International Conference on Computational Linguistics (COLING), Bombay, India, 8–15 December 2012.
26. Bond, F.; Foster, R. Linking and Extending an Open Multilingual Wordnet. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria, 4–9 August 2013; ACL Press: Sofia, Bulgaria, 2013; Volume 1, pp. 1352–1362.
27. Simões, A.; Guinovart, X.G. Bootstrapping a Portuguese WordNet from Galician, Spanish and English Wordnets. In *Advances in Speech and Language Technologies for Iberian Languages, Proceedings of the 2nd International Conference, IberSPEECH 2014, Las Palmas de Gran Canaria, Spain, 19–21 November 2014*; Springer: Berlin, Germany, 2014; Volume 8854, pp. 239–248.
28. Gonzalez-Agirre, A.; Laparra, E.; Rigau, G. Multilingual Central Repository version 3.0. In Proceedings of the 8th International Conference on Language Resources and Evaluation (ELRA), Istanbul, Turkey, 21–27 May 2012; pp. 2525–2529.
29. Maziero, E.G.; Pardo, T.A.S.; Felippo, A.D.; Dias-da-Silva, B.C. 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*; STIL: Vila Velha, Brazil, 2008; pp. 390–392.
30. Gonçalo Oliveira, H. CONTO.PT: Groundwork for the Automatic Creation of a Fuzzy Portuguese Wordnet. In Proceedings of the 12th International Conference on Computational Processing of the Portuguese Language (PROPOR), Tomar, Portugal, 13–15 July 2016; Volume 9727, pp. 283–295.
31. Dias-da-Silva, B.C. Wordnet.Br: An exercise of human language technology research. In Proceedings of the 3rd International WordNet Conference (GWC), Seogwipo, Korea, 22–26 January 2006; pp. 301–303.
32. Gonçalo Oliveira, H.; Santos, D.; Gomes, P.; Seco, N. PAPEL: A Dictionary-Based Lexical Ontology for Portuguese. In Proceedings of the 8th International Conference on Computational Processing of the Portuguese Language (PROPOR), Aveiro, Portugal, 8–10 September 2008; Volume 5190, pp. 31–40.
33. Simões, A.; Sanromán, Á.I.; Almeida, J.J. Dicionário-Aberto: A Source of Resources for the Portuguese Language Processing. In Proceedings of the 10th International Conference on Computational Processing of the Portuguese Language (PROPOR 2012), Coimbra, Portugal, 17–20 April 2012; Volume 7243, pp. 121–127.
34. Barreiro, A. 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, 8–10 June 2008; Cambridge Scholars Publishing: Cambridge, UK, 2010.
35. Gonçalo Oliveira, H.; Pérez, L.A.; Costa, H.; Gomes, P. Uma rede léxico-semântica de grandes dimensões para o português, extraída a partir de dicionários electrónicos. *Linguamática* **2011**, *3*, 23–38.
36. Gonçalo Oliveira, H. Comparing and Combining Portuguese Lexical-Semantic Knowledge Bases. In Proceedings of the 6th Symposium on Languages, Applications and Technologies (SLATE 2017), Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, Vila do Conde, Portugal, 26–27 June 2017; Springer International Publishing: Cham, Switzerland; pp. 16:1–16:14.
37. Pilehvar, M.T.; Jurgens, D.; Navigli, R. Align, Disambiguate and Walk: A Unified Approach for Measuring Semantic Similarity. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria, 5 August 2013; ACL Press: Sofia, Bulgaria, 2013; Volume 1, pp. 1341–1351.
38. Newman, D.; Lau, J.H.; Grieser, K.; Baldwin, T. Automatic Evaluation of Topic Coherence. In Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Human Language Technologies, Los Angeles, CA, USA, 1–6 June 2010; pp. 100–108.
39. Bouma, G. Normalized (pointwise) mutual information in collocation extraction. *In From Form to Meaning: Processing Texts Automatically, Proceedings of the Biennial GSCL Conference Potsdam, Germany 2009*; Narr Francke Attempto Verlag GmbH: Tübingen, Germany, 2009.
40. Rodrigues, J.A.; Branco, A.; Neale, S.; Silva, J.R. LX-DSemVectors: Distributional Semantics Models for Portuguese. In Proceedings of the 12th International Conference on the Computational Processing of the Portuguese Language (PROPOR), Tomar, Portugal, 13–15 July 2016; Volume 9727, pp. 259–270.

41. Ling, W.; Dyer, C.; Black, A.W.; Trancoso, I. Two/Too Simple Adaptations of Word2Vec for Syntax Problems. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, CO, USA, 31 May–5 June 2015; pp. 1299–1304.
42. Batchkarov, M.; Kober, T.; Reffin, J.; Weeds, J.; Weir, D.J. A critique of word similarity as a method for evaluating distributional semantic models. In Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP, RepEval@ACL 2016, Berlin, Germany, 12 August 2016; pp. 7–12.
43. Pilehvar, M.T.; Navigli, R. From senses to texts: An all-in-one graph-based approach for measuring semantic similarity. *Artif. Intell.* **2015**, *228*, 95–128.