LX-DSemVectors: Distributional Semantics Models for Portuguese

João Rodrigues(✉), António Branco, Steven Neale, and João Silva

Department of Informatics,
Faculty of Sciences, University of Lisbon, Lisbon, Portugal
{joao.rodrigues,antonio.branco,steven.neale,jsilva}@di.fc.ul.pt

Abstract. In this article we describe the creation and distribution of the first publicly available word embeddings for Portuguese. Our embeddings are evaluated on their own and also compared with the original English models on a well-known analogy task. We gathered a large Portuguese corpus of 1.7 billion tokens, developed the first distributional semantic analogies test set for Portuguese, and proceeded with the first parametrization and evaluation of Portuguese word embeddings models.

Keywords: Distributional semantics · Word embeddings · Portuguese

1 Introduction

Current research trends focusing on distributional semantics are sparking interest in possible ways to enrich the resources and tools used for natural language processing (NLP) tasks. Researchers and practitioners are exploring possible improvements that can be achieved from integrating distributional vectors semantics, also known as word embeddings, in a range of syntactic and semantic tasks, including speech recognition [16], semantic similarity of words [15] part-of-speech (POS) tagging, named entity recognition, sentiment analysis [13] and logical semantics [4].

Experimenting with word embeddings in such tasks requires large data sets to extract word embeddings in a specific language. At the time of writing we have found no such freely available or evaluated data set for Portuguese to exist. There is therefore a need to create word embeddings in the Portuguese language that can be explored in the types of tasks mentioned above for the English language.

In this paper we describe our results in training, parameterizing and evaluating word embeddings for Portuguese – a computationally intensive and time consuming undertaking – as well as comparing them with an implementation for the English language. Our contribution is in making available a set of trained word embeddings for the computational processing of Portuguese, as well as a set of instructions for getting them running quickly and easily.

In Sect. 2 we briefly describe word embeddings and current methods for obtaining them, followed by a description of our own implementation of the

J. Silva et al. (Eds.): PROPOR 2016, LNAI 9727, pp. 259–270, 2016.
DOI: 10.1007/978-3-319-41552-9_27

models in Sect. 3 and the set of experiments we run to improve their accuracy. The resulting models are evaluated and analyzed in Sect. 4 against the English models, before we draw our conclusions and outline our plans for future work in Sect. 5.

2 Related Work

As concisely stated in [8], "distributional semantics is predicated on the assumption that linguistic units with certain semantic similarities also share certain similarities in the relevant environments". Addressing this so-called 'relevant environment' using distributional semantics methods is based on two key paradigms – count-based and prediction-based methods.

Both count and prediction-based methods generate a set of distributional vectors (also known as word embeddings or distributed word representations) that are able to reflect the semantic similarity between words, with the meaning of each word possibly characterized by a vector of real values. For example, cosine similarity can be used to find the similarity between two word vectors, and hence the meaning of the words they represent. Interestingly, it is possible to perform algebraic operations using the vectors, as demonstrated by the typical example of $vector(king) - vector(man) + vector(woman)$ resulting in a similar vector to $vector(queen)$, the distributed word representation of the word queen [15].

In this article we focus on the prediction-based methods, which – inspired by neural network-based probabilistic language models [3] – predict the co-occurrent context words for a word of interest using a sliding window of one or more (n-) words that continually moves along the corpus with each new word of interest. The co-occurring context words captured by this sliding window fill the cells of the vector for a given word of interest as we move through the corpus, in contrast to count-based methods that count all cases of co-occurrence with a word of interest across the entire corpus, often resulting in a huge and sparse vector space that typically grows with the quadratic size of the vocabulary.

An even simpler use of these methods was presented in [16], where a combination of an input and an output layer – each with a length corresponding to the size of the vocabulary – and one hidden layer with approximately 60 neurons are trained to estimate the probability distribution of the next word in a text given a previous word from the vocabulary, as in [3]. This work was later extended with two new models – the continuous bag of words (CBOW) and Skip-gram – to further simplify the neural probabilistic language models [15]. These two models represent a shallow use of neural networks – the CBOW model introduces a shared projection layer for all words, a weighted matrix between the first two layers, and a sliding co-occurrent context window for the training of the current word of interest; while the Skip-gram model creates a standalone vector representing the combination of contextual words and then predicts the context vector closest to the current word vector.

These new methods – designed to leverage maximum information from large data sets at minimum computational costs – began to be evaluated in a semantic

textual similarity task. One of the comprehensive test sets for measuring both syntactic and semantic regularities using analogies that was made available [15] is the same test set we use for comparison in this paper[1]. As discussed in [12], no qualitative difference can be pinpointed to one or the other of the different approaches to word vectors, count or prediction-based ones. Any difference to be found among particular models is instead due to the optimization of various hyperparameters, which can yield better results using one approach or the other. Notwithstanding, both methods still resort to different computational means to create the shared semantic models and in a range of different sets of tasks the Skip-gram prediction-based model has been shown to be better, on average.

Both the CBOW and Skip-gram models are available with the standalone implementation 'word2vec'[2]. For the creation of the Portuguese vectors we used Gensim [19], a Python-based Skip-gram implementation. Gensim is a good choice for our work because it allows for different distributional semantic methods to be deployed within the same framework, models which can be ported to word2vec later if convenient.

Regarding related work concerning distributional semantics of Portuguese, Portuguese corpora are used for the creation of distributional semantic models in [10,20], the latter using the CharWNN deep neural network for boosting named entity recognition and the former applying a novel method that uses parallel data for document classification. Using long short-term memory (LSTM) neural networks, [14] constructs vector representations of words with a Portuguese model yielding state-of-the-art results in language modeling and part-of-speech (POS) tagging. [21] also uses distributed word representation in POS tagging for Portuguese by using a neural language model. They resort to word2vec and to the Portuguese Wikipedia, CETENFolha and CETEMPublico corpora, obtaining state-of-the-art results for POS tagging. Another article using Portuguese word embeddings is the Polyglot project [1], which uses a Portuguese Wikipedia corpus to support POS tagging. Finally, [7] also seeks to improve on the POS tagging of Portuguese using word embeddings.

Although all of these works used models for distributional semantics of Portuguese, none of them report an evaluation of parameter optimization or an assessment against current test sets. In this article, we seek to overcome this shortcoming by performing a comparison with state-of-the-art evaluation methods. The models trained in the related works above are also not available except the one from the Polyglot project, which has a unique model. A major contribution of this article is making these trained and tuned models of Portuguese word embeddings available as a freely available resource.

3 Implementation

For the creation of the Portuguese word embeddings we chose Skip-gram as the training algorithm, since it obtains the best accuracy, on average, from a range

[1] For a more complete description of the evaluation methods, see [22].
[2] http://code.google.com/p/word2vec/.

of test sets in the distributional semantics domain [12]. We installed the Gensim framework and developed the necessary scripts for the training and evaluation, which are made available at http://github.com/nlx-group.

The first step in the implementation process was the gathering of corpora, described below in Subsect. 3.1. For a reasonable comparison with the original (English) evaluation of Skip-gram, our evaluation should be performed with a similar test set. Since the original test set is in English, it was necessary to translate it – this process is described in Subsect. 3.2.

With Portuguese corpora and a test set in place, we could then design and undertake a set of experiments encompassing the training of different models with the objective of maximizing the accuracy of the models obtained. These experiments are described in Subsect. 3.3.

3.1 Acquisition of Corpora

For Portuguese (both Brazilian and European variants), a total of 1,723,693,241 tokens from 121,706,288 sentences were gathered. To the best of our knowledge, this is the largest raw text data set whose gathering was ever reported for the Portuguese language. Table 1 lists each of the gathered corpora used along with their respective token and sentence volumes (obtained after tokenization). We used a web crawler to gather news articles from Jornal Digital[3] and Observador[4]. The crawl gathered all public news articles available on November 20, 2015, including their titles, headlines and the articles themselves.

After the search, extraction and cleaning of the corpora, a tokenization process took place. No lowercasing was performed over the source texts and the original surface form of the word was used. For the tokenization, the LX-Tokenizer [5] was used, which has a reported f-score of 99.72 %.

3.2 Test Set

The test set described in [15] – a collection of word analogies – was used as the basis for the assessment of word embeddings. An example entry in this data set would read: 'Berlin Germany Lisbon Portugal'. With these four words relations – as in this example – one can test semantic analogies by using any of the possible combinations of three of the four word vectors in one entry and testing whether or not the resulting vector is similar to the (fourth) word vector missing from the combination being tested. In the example above, the completed analogy should read: 'Berlin is to Germany as Lisbon is to Portugal'.

The test set contains five types of semantic analogy: common capitals and countries, all capitals and countries, currency, cities and states, and family relations. Nine types of syntactic analogy are also represented: adjective to adverb, opposite, comparative, superlative, present participle, nationality (adjective),

[3] www.jornaldigital.com.
[4] www.observador.pt.

Table 1. Portuguese corpora used for training

Corpus	Tokens	Sentences	Description	Ref.
TCC	61,979	642	scientific corpus	[18]
QTLeap	56,255	4,000	q/a pairs in the IT domain	[9]
CRPC	133,497	5,061	oral communication of direct inquiries	[17]
Tanzil	178,225	9,377	Quran translation to Portuguese	[24]
CINTIL	707,444	30,344	International corpus of Portuguese	[2]
JDigital	3,891,407	110,227	news articles from Jornal Digital	
Ted2013	3,173,357	156,033	corpus from the TED talks	[6]
KDE4	3,123,310	230,178	KDE4 localization files	[23]
Observador	34,900,297	732,240	news articles from Observador	
EMEA	19,083,444	1,213,566	documentation from EMA	[23]
ECB	71,387,581	2,162,343	documentation from the ECB	[23]
Europarl	67,506,802	2,171,029	European Parliament sessions	[11]
DGT	73,788,835	3,153,654	translation memories from the Acquis	[23]
Stackoverflow	36,200,297	3,767,771	posts from Stackoverflow	
EUBookshop	203,762,634	7,310,336	documentation from the EU bookshop	[23]
Wikipedia	246,550,786	7,460,428	PT Wikipedia dump of 01/09/2015	
CETEMPublico	225,906,693	8,065,830	news articles from the Público	
OpenSubtitles	442,182,528	54,415,635	Portuguese subtitles until 2013	[23]
CETENFolha	291,097,870	30,707,594	news articles from Folha de S. Paulo	
Total	1,723,693,241	121,706,288		

past tense, plural nouns and plural verbs. The test set contains a total of 8869 semantic and 10675 syntactic entries.

For the evaluation of the Portuguese word embeddings, the original English test set was translated into Portuguese by skilled, native Portuguese-speaking language experts. The resulting translations, LX-4WAnalogies, and corresponding English terms are available at http://github.com/nlx-group.

There were some English words that could not be accurately translated into a unique Portuguese word. Given that the original evaluation does not support vector composition, if a single word from the original four words in an analogy could not be translated as a single word, then the analogy in question had to be dropped. Therefore, the resulting Portuguese test set kept only 17558 analogies from the original 19544 English analogies. The groups of analogies affected were:

- **family**: From the original 506 analogies, 462 were retained in the translation (506, 462, -44). For example, the single word *copwoman* is translated to the two word expression *mulher polícia*.
- **gram1-adjective-to-adverb** (992, 930, -62): For example, *most* is translated to *a maioria*.
- **gram2-opposite** (812, 756, -56): For example, *uncompetitive* is translated to *não competitivo*.
- **gram3-comparative** (1332, 30, -1302): The Portuguese language needs, in most of the cases, a separate word to mark the comparison. This is accomplished with adverbs that quantify the adjective. For example *brighter* is translatable to *mais brilhante*.
- **gram4-superlative** (1122, 600, -522): This group is affected by the same linguistic phenomena as in the comparative group. For example, *tastiest* is translated to *o mais saboroso*.

3.3 Experiments

A total of five experiments were ran, with the objective of narrowing the choice of corpora and parameters to arrive at the most accurate word embeddings. The evaluation was performed in two ways: with the original restriction, where only analogies in which the frequencies of all four words are in the top 30000 most frequent words overall, and a second evaluation where this restriction is not applied. The unrestricted evaluation is useful for grasping the achievable generalization of a model.

- **First experiment** – Firstly, we use the vanilla parameters of Gensim to evaluate each of the gathered corpora separately (both with and without restriction). Secondly, we evaluate each of the gathered corpora incrementally – that is, one of the other gathered corpora is added to the whole in each incrementation step.
- **Second experiment** – We take the largest resulting data sets from the incremental phase of the first experiment and evaluate them with larger vector dimensions. The reason for performing this experiment is to test whether or not the proportionality of data and vector dimensions influences the result, as expected.
- **Third experiment** – We use only the data sets that in the first experiment yielded an improvement in accuracy when they were incremented with other corpora (using the vanilla parameters).

- **Fourth experiment** – We compare the best model obtained in the third experiment with the model obtained by assembling together: (a) corpora that improved over each other during the incremental phase of experiment 1; (b) Europarl (which improved over the previous two increments although they had not improved over the highest incremental score at that point); and (c) CETEMPublico and CETENFolha (which were shown to yield the best scores without restriction in experiment 2). Both models were evaluated along a range of vector dimensions.
- **Fifth experiment** – We evaluated the effect of additional parameterization (besides vector dimensions) for the most accurate model obtained in the fourth experiment, including: (a) sliding window size (value 5 or 10); (b) initial learning rate, which linearly drops to zero as the training progresses (0.025 or 0.05); (c) the threshold for configuring which higher-frequency words are randomly down sampled (0 or 0.00005); (d) hierarchical sampling (0-off or 1-on); and (e) negative sampling (5 or 15).

4 Evaluation

4.1 Experiments with Portuguese Embeddings

The first experiment (see Table 2) shows that – as expected – better accuracy is obtained with larger data sets. Wikipedia excels here both because of its higher quality and its relevance to the analogies test set. When incrementing the data set in a step wise fashion, a large increase in accuracy can be seen when Wikipedia is added to all of the previous corpora incremented up to that point (incr_15, Table 3). Beyond this (incr_16 to 18) the accuracy drops, possibly due to the fixed vector dimension.

The second experiment (Table 4) tests for the vector dimension with the four largest corpora (incr_15 to 18 in Table 3). Increasing the vector dimension here leads to increased accuracy, with the strongest results appearing around vectors with dimension 400. The OpenSubtitles corpus seems to introduce some noise, which reduces the accuracy. Although the incrementing with CETEMPublico and CETENFolha did not improve over the top accuracy score with the typical restriction, an increase can be seen without such restriction that reaches the top value in that setting.

The third experiment (Table 5) reveals that by using only those corpora that induced improved accuracy better generalization is obtained as indicated in the score 28.5 for accuracy without restriction (against 26.3 from the first experiment).

The fourth experiment (Table 6) reveals that although the assembled corpora in the third experiment (chosen_incr_9 in Table 5, labeled as model 2 in Table 6) permit to obtain the best result in the restricted evaluation, that accuracy can be surpassed by a non-restricted evaluation with the selected larger corpus as the vector dimension is increased to 400.

Because model 1 in fourth experiment yielded higher accuracy than model 2 in almost all non-restricted evaluations (Table 6), model 1 was chosen – with a

Table 2. First experiment – accuracy (with/without restriction) obtained by training on the different corpora

Corpus	Accuracy %
TCC	0.0/0.0
QTLeap	0.0/0.0
CRPC	0.0/0.0
Tanzil	5.0/5.0
CINTIL	0.8/0.8
JDigital	2.4/2.4
Ted2013	3.2/3.2
KDE4	3.7/3.7
Observador	9.8/5.9
EMEA	4.0/3.6
ECB	6.5/2.6
Europarl	12.7/6.4
DGT	8.4/3.8
Stackoverflow	6.9/3.1
EUBookshop	17.3/5.9
Wikipedia	**36.3/26.1**
CETEMPublico	27.6/20.1
OpenSubtitles	25.4/18.5
CETENFolha	19.0/13.8

Table 3. First experiment – accuracy (with/without restriction) obtained from incrementally adding each corpus to the previous (for example, incre_ 3 consists of the TCC+QTLeap+CRPC+Tanzil corpora)

Corpus	Accuracy %
incr_0 (TCC)	0.0/0.0
incr_1 (+QTLeap)	0.0/0.0
incr_2 (+CRPC)	0.0/0.0
incr_3 (+Tanzil)	1.8/1.8
incr_4 (+CINTIL)	2.9/2.9
incr_5 (+JDigital)	2.0/1.9
incr_6 (+Ted2013)	6.3/4.9
incr_7 (+KDE4)	6.9/4.9
incr_8 (+Observador)	13.2/8.3
incr_9 (+EMEA)	5.9/3.4
incr_10 (+ECB)	3.9/2.2
incr_11 (+Europarl)	8.4/4.5
incr_12 (+DGT)	8.3/3.5
incr_13 (+Stackoverflow)	7.2/3.5
incr_14 (+EUBookshop)	16.9/6.6
incr_15 (+Wikipedia)	**38.2/26.3**
incr_16 (+CETEMPublico)	32.8/23.9
incr_17 (+OpenSubtitles)	30.3/20.5
incr_18 (+CETENFolha)	30.5/21.4

vector dimension of 400 – to be used in the fifth and final experiment (Table 7). In the fifth experiment model 1 of the fourth experiment is evaluated against different settings of parameters. The results of this experiment clearly shows that all models using hierarchical sampling induce a reduced accuracy, while increasing the negative sampling from 5 to 15 units increases accuracy. The best obtained score was in p_17 with an accuracy of 52.8 % in the restricted evaluation and 37.7 % without restriction.

4.2 Comparison with English Models

In the original evaluation of the English word embeddings [15] – trained with a vector dimensionality of 300 and a corpus of 783 million tokens – accuracy of 50.4 % was obtained with restriction. In a second experiment – where the word embeddings were trained with a vector dimensionality of 1000 and a corpus of 6 billion tokens – accuracy of 65.6 % was obtained without restriction.

Our best trained word embeddings for Portuguese – using a corpus of approximately 1 billion tokens – obtained an accuracy of 52.8 % with restriction (compared to 50.4 % in English, with 783 million tokens) and 37.7 % without restriction (compared to 65 % in English, with 6 billion tokens).

Table 4. Second experiment – different ranges of vector dimensions evaluated on the incremental corpora with the highest accuracy (incr_15 to 18) from experiment 1 (with/without restriction)

	vector dimension					
corpus	100	200	300	400	500	600
incr_15	38.2/26.3	43.4/29.2	44.7/29.6	**44.9**/30.6	43.4/29.5	40.4/26.8
incr_16	32.8/23.9	40.0/29.8	43.6/**32.8**	43.6/32.1	42.4/32.2	43.5/32.0
incr_17	30.3/20.5	35.2/26.2	35.0/25.4	35.3/25.8	37.2/28.3	36.7/27.2
incr_18	30.5/21.4	36.3/27.0	37.3/27.2	35.3/27.0	35.2/27.1	35.3/26.1

The results obtained for the Portuguese embeddings seem to be in line with those obtained for English when using data sets of similar size. When analyzing this kind of comparison, the variety of different training conditions – including differences in parameterization, vector dimensionalities and the larger English corpora – must be taken into account. We believe that our work reported here suggests room for further improvements in the Portuguese models, and that the work described in this paper paves the way for further exploration, specially if larger data sets are used.

Table 5. Third experiment – selected incremental corpora, accuracy with/without restriction (each new corpus is added to the existing with each incrementation – for example, incr_3 consists of the TCC+QTLeap+CRPC+Tanzil corpora)

Corpus	Accuracy %
chosen_incr_0 (TCC)	0.0/0.0
chosen_incr_1 (+QTLeap)	0.0/0.0
chosen_incr_2 (+CRPC)	0.0/0.0
chosen_incr_3 (+Tanzil)	1.8/1.8
chosen_incr_4 (+CINTIL)	2.8/2.8
chosen_incr_5 (+Ted2013)	4.1/3.9
chosen_incr_6 (+KDE4)	5.1/4.1
chosen_incr_7 (+Observador)	13.8/9.1
chosen_incr_8 (+EUBookshop)	15.5/6.4
chosen_incr_9 (+Wikipedia)	**37.3/28.5**
chosen_incr_10 (+CETEMPublico)	31.4/25.2
chosen_incr_11 (+CETENFolha)	33.9/26.4

Table 6. Fourth experiment – comparing model 1 (corpora: TCC, QTLeap, CRPC, Tanzil, CINTIL, Ted2013, KDE4, Observador, Europarl, DGT, EUBookshop, Wikipedia, CETEMPublico and CETENfolha) with model 2 (corpora: chosen_incr_9 from experiment 3, accuracy with/without restriction)

	vector dimension					
corpus	100	200	300	400	500	600
model_1	34.4/26.3	40.3/30.7	41.4/32.3	42.6/**33.1**	43.0/32.2	41.8/32.0
model_2	37.3/28.5	40.9/30.0	42.4/31.5	**43.1**/31.6	42.6/30.4	42.4/30.7

Table 7. Fifth experiment – accuracy (acc. with/without restriction) of model 1 from experiment 4 with additional parameterization including: sliding window size (win), learning rate (lrate), threshold for configuring which higher-frequency words are randomly downsampled (hf), hierarchical sampling (hs), and negative sampling (ns). The training of these models took a week and a half using a server consisting of 30 processors (Intel(R) Xeon® 8C CPU E5-2640 V2 @ 2.00 GHz, 20 M Cache, RAM 16x 16 GB RDIMM, 1600 MHz)

	win	lrate	hf	hs	ns	acc. %		win	lrate	hf	hs	ns	acc. %
p_0	5	0.025	0	0	5	51.6/35.4	p_16	10	0.025	0	0	5	52.0/37.0
p_1	5	0.025	0	0	15	49.3/34.9	p_17	10	0.025	0	0	15	**52.8/37.7**
p_2	5	0.025	0	1	5	45.4/36.3	p_18	10	0.025	0	1	5	48.0/36.5
p_3	5	0.025	0	1	15	47.2/36.1	p_19	10	0.025	0	1	15	48.6/36.6
p_4	5	0.025	1e-05	0	5	50.7/31.4	p_20	10	0.025	1e-05	0	5	50.0/30.5
p_5	5	0.025	1e-05	0	15	52.1/32.6	p_21	10	0.025	1e-05	0	15	51.3/32.0
p_6	5	0.025	1e-05	1	5	45.2/33.7	p_22	10	0.025	1e-05	1	5	44.4/33.0
p_7	5	0.025	1e-05	1	15	47.1/35.0	p_23	10	0.025	1e-05	1	15	44.4/32.3
p_8	5	0.05	0	0	5	50.2/36.4	p_24	10	0.05	0	0	5	50.7/36.4
p_9	5	0.05	0	0	15	50.5/36.7	p_25	10	0.05	0	0	15	51.0/36.8
p_10	5	0.05	0	1	5	45.8/34.7	p_26	10	0.05	0	1	5	44.6/32.1
p_11	5	0.05	0	1	15	44.8/34.6	p_27	10	0.05	0	1	15	46.1/33.2
p_12	5	0.05	1e-05	0	5	50.6/30.5	p_28	10	0.05	1e-05	0	5	46.4/28.0
p_13	5	0.05	1e-05	0	15	52.5/34.3	p_29	10	0.05	1e-05	0	15	49.7/31.8
p_14	5	0.05	1e-05	1	5	43.9/30.9	p_30	10	0.05	1e-05	1	5	41.2/28.5
p_15	5	0.05	1e-05	1	15	44.3/31.6	p_31	10	0.05	1e-05	1	15	40.5/28.6

5 Conclusion

In this paper we described the creation, parameterization and evaluation of the first publicly available distributional semantic models for Portuguese, which perform in line with the original state-of-the-art models for English. All the models from the fifth experiment are made available from http://github.com/nlx-group.

In future work we plan to account for missing analogies in our test set by using phrases instead of words. While introducing lowercasing and lemmatization

steps and making use of richer linguistic knowledge are also promising directions, acquiring a larger Portuguese corpora to train on remains the most important step as we seek to improve the accuracy of our models.

Acknowledgements. The results reported in this paper were partially supported by the Portuguese Government's P2020 program under the grant 08/SI/2015/3279: ASSET-Intelligent Assistance for Everyone Everywhere, and by the EC's FP7 program under the grant number 610516: QTLeap-Quality Translation by Deep Language Engineering Approaches.

References

1. Al-Rfou, R., Perozzi, B., Skiena, S.: Polyglot: distributed word representations for multilingual NLP. In: Proceedings of the Seventeenth Conference on Computational Natural Language Learning. pp. 183–192. Association for Computational Linguistics, Sofia, August 2013
2. Barreto, F., Branco, A., Ferreira, E., Mendes, A., Nascimento, M.F., Nunes, F., Silva, J.: Open resources and tools for the shallow processing of portuguese: the tagshare project. In: Proceedings of LREC 2006. Citeseer (2006)
3. Bengio, Y., Ducharme, R., Vincent, P., Janvin, C.: A neural probabilistic language model. J. Mach. Learn. Res. **3**, 1137–1155 (2003)
4. Bowman, S.R., Potts, C., Manning, C.D.: Recursive neural networks can learn logical semantics. In: ACL-IJCNLP, p. 12 (2015)
5. Branco, A., Silva, J.: Evaluating solutions for the rapid development of state-of-the-art pos taggers for portuguese. In: LREC (2004)
6. Cettolo, M., Girardi, C., Federico, M.: Wit3: Web inventory of transcribed and translated talks. In: Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT), pp. 261–268 (2012)
7. Fonseca, E.R., Rosa, J.L.G., Aluísio, S.M.: Evaluating word embeddings and a revised corpus for part-of-speech tagging in portuguese. J. Braz. Comput. Soc. **21**(1), 1–14 (2015)
8. Garvin, P.L.: Computer participation in linguistic research. Language **38**, 385–389 (1962)
9. Gaudio, R.D., Burchardt, A., Branco, A.: Evaluating machine translation in a usage scenario. In: Proceedings of LREC (to appear in print, 2016)
10. Hermann, K.M., Blunsom, P.: Multilingual models for compositional distributed semantics. arXiv preprint arXiv:1404.4641 (2014)
11. Koehn, P.: Europarl: a parallel corpus for statistical machine translation. In: MT Summit, vol. 5, pp. 79–86. Citeseer (2005)
12. Levy, O., Goldberg, Y., Dagan, I.: Improving distributional similarity with lessons learned from word embeddings. Trans. Assoc. Comput. Linguist. **3**, 211–225 (2015)
13. Li, J., Jurafsky, D.: Do multi-sense embeddings improve natural language understanding? arXiv preprint arXiv:1506.01070 (2015)
14. Ling, W., Luís, T., Marujo, L., Astudillo, R.F., Amir, S., Dyer, C., Black, A.W., Trancoso, I.: Finding function in form: Compositional character models for open vocabulary word representation. arXiv preprint arXiv:1508.02096 (2015)
15. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)

16. Mikolov, T., Kopeckỳ, J., Burget, L., Glembek, O., Černockỳ, J.H.: Neural network based language models for highly inflective languages. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4725–4728. IEEE (2009)
17. do Nascimento, M.F.B., Pereira, L., Saramago, J.: Portuguese corpora at CLUL. PRAXIS **2**(2.1/759), 95 (2000)
18. Pardo, T.A.S., Nunes, M.d.G.V.: A construção de um corpus de textos científicos em português do brasil e sua marcação retórica. Tech. rep. (2003)
19. Rehurek, R., Sojka, P.: Gensim–python framework for vector space modelling. NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic (2011)
20. dos Santos, C., Guimaraes, V., Niterói, R., de Janeiro, R.: Boosting named entity recognition with neural character embeddings. In: Proceedings of NEWS 2015 The Fifth Named Entities Workshop, p. 25 (2015)
21. Santos, C.D., Zadrozny, B.: Learning character-level representations for part-of-speech tagging. In: Proceedings of the 31st International Conference on Machine Learning (ICML), pp. 1818–1826 (2014)
22. Schnabel, T., Labutov, I., Mimno, D., Joachims, T.: Evaluation methods for unsupervised word embeddings. In: Proceedings of EMNLP (2015)
23. Tiedemann, J.: News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In: Nicolov, N., Bontcheva, K., Angelova, G., Mitkov, R. (eds.) Recent Advances in Natural Language Processing, vol. V, pp. 237–248. John Benjamins, Amsterdam/Philadelphia (2009)
24. Tiedemann, J.: Parallel data, tools and interfaces in OPUS. In: Chair, N.C.C., Choukri, K., Declerck, T., Dogan, M.U., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S. (eds.) Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC). European Language Resources Association (ELRA), Istanbul, May 2012