Finely Tuned, 2 Billion Token Based Word Embeddings for Portuguese

João Rodrigues and António Branco
University of Lisbon
NLX-Natural Language and Speech Group, Department of Informatics
Faculdade de Ciências
Campo Grande, 1749-016 Lisboa, Portugal
{joao.rodrigues, antonio.branco}@di.fc.ul.pt

Abstract

A distributional semantics model — also known as word embeddings — is a major asset for any language as the research results reported in the literature have consistently shown that it is instrumental to improve the performance of a wide range of applications and processing tasks for that language. In this paper, we describe the development of an advanced distributional model for Portuguese, with the largest vocabulary and the best evaluation scores published so far. This model was made possible by resorting to new languages resources we recently developed: to a much larger training corpus than before and to a more sophisticated evaluation supported by new and more fine-grained evaluation tasks and data sets. We also indicate how the new language resource reported on here is being distributed and where it can be obtained for free under a most permissive license.

Keywords: distributional semantics, word embeddings, word analogy, lexical similarity, conceptual categorization, Portuguese.

1. Introduction

Distributional semantics assumes that the frequency of contexts in which expressions occur helps to capture important syntactic and semantic properties of these expressions (Garvin, 1962). Exploring distributional models has led to advances in a range of natural language processing tasks, from dialog systems (Chen et al., 2014) to question answering (Bordes et al., 2014), among many others.

A distributional model is a major language resource for any language as it is instrumental to enhance the performance of many applications and processing tasks for that language. Recent work concerned with distributional models for Portuguese contributed with the creation and public release of a free distributional model for this language that resorts to a considerably large corpus, for the training, and to a standard analogy test set, for the evaluation and tuning of this model (Rodrigues et al., 2016).

Since then that corpus has been greatly expanded by our group and is now a very large data set, reaching more than 2 billion tokens.

More recently, finer tuning of distributional models for Portuguese became possible with the creation and distribution of a wide range of data sets by our group that support the evaluation of lexical similarity and conceptual categorization tasks, and that are comparable in size and domain to mainstream datasets with the same purpose for other languages (Querido et al., 2017).

The aim of the present paper is to report on the development, and its free distribution, of a more advanced distributional model for Portuguese, with the largest vocabulary and the best evaluation scores published so far, made possible by a much larger training corpus than before and by a more sophisticated evaluation supported by new and more fine-grained evaluation tasks and data sets.

In Section 2., we discuss previous work in the literature related to the construction and application of distributional models of Portuguese. The methods that were used for the development of the new distributional model are presented in Section 3.. Section 4. describes the data sets used for this purpose, and Section 5. the experiments undertaken to fine-tune the model. We present and discuss the outcome of this development in Section 6.. The paper closes with Section 7., indicating how the resources developed are distributed and can be obtained, and with Section 8. with final remarks.

2. Related Work

Recently, there has been an increased interest in the distributional semantics for Portuguese to support a number of natural language processing tasks, where distributional models tend to be seen as an additional instrumental resource that may help to improve the performance of these tasks. But taking the distributional as our core subject of research, this can alternatively be envisaged under a different perspective. As the development of distributional models is a research topic in itself, applying different models — developed under different design options — in different tasks offers the test bed for the evaluation of these models and of their level of fitness in terms of capturing the essential linguistic information they are meant to encode.

Extrinsic evaluation of these models consists in testing how differences in the design of distributional models induce possible improvements in complex applications or systems where they are embedded. Depending on the level of contribution to the final performance of these systems, it may thus happen that improvements on the models may induce limited to none improvement to those systems. That is why it is important to have also *intrinsic* evaluation, which is meant to give an indication of the appropriateness of the models irrespectively of their eventual impact to larger systems where they may happen to become embedded.

Recent research on the distributional semantics of Portuguese has focused mainly on its eventual instrumental contribution to a number of language processing tasks, and thus directed more to the extrinsic evaluation of distributional models. Examples of this increased interest include named entity recognition (Hermann and Blunsom, 2014), document classification (dos Santos et al., 2015), language modeling (Ling et al., 2015), topic mining (Machado et al., 2017) and paraphrase detection (Hartmann et al., 2017). Part-of-speech tagging represent the largest share of this interest, with the works of (Al-Rfou et al., 2013), dos Santos and Zadrozny (2014), Fonseca et al. (2015), Fonseca and Aluísio (2016) and Hartmann et al. (2017).

In what concerns the intrinsic evaluation of distributional models, the *word analogy* task is the test bed of choice. The goal is to complete an analogy with a fourth word, for which only the other three words are given. Having, for example, the following analogy, ”Berlin is for Germany as Lisbon is for Portugal”, and hiding one of the four key terms of the analogy, for example ”Portugal”, the successful completion of the task will consist of correctly indicating that ”Portugal” is the missing term.

This task was one of the first used for the intrinsic evaluation of distributional semantics spaces as its resolution is considered to be tapping in some crucial way on the underlying syntactic and semantic knowledge encoded in the semantic vectors.

An evaluation data set for the word analogy task in Portuguese, LX-4WAnalogies, was first made available by our group, namely Rodrigues et al. (2016), who also developed and distributed the first publicly available word embeddings for Portuguese, LX-DSemVectors, developed on a 1.7 billion token training corpus and tunned on this LX-4WAnalogies data set. This distributional model LX-DSemVectors was obtained with the Skip-Gram neural network from word2vec (Mikolov et al., 2013) and achieves 37.7% accuracy in the word analogy task.

The evaluation data set LX-4WAnalogies was subsequently used by Hartmann et al. (2017) to fine tune another distributional model for Portuguese. This model was developed with Glove (Pennington et al., 2014) on a 1.4 billion token corpus, achieving 46.2% accuracy in the word analogy task.

The intrinsic evaluation of distributional models has been further pursued in the literature, and further tasks have been proposed to complement the assessment supported by the word analogy task. The respective evaluation data sets have been development, most prominently for the English language.

In our group, we have recently pioneered also the development of these types of data sets for the intrinsic evaluation of distributional models (Querido et al., 2017), which we now briefly introduce, together with a description of the respective task.

The task of *lexical similarity* consists of indicating a score, in some predefined scale, that represents the level of semantic similarity between two words that are entered as input. For example, the word ”ice-cream” has a weak to none semantic similarity with the word ”whale”, but a strong similarity to words such as ”gelato” or ”sorbet”, and should be annotated a higher score than the first pair.

The LX-WordSim-353, LX-SimLex-999, and LX-Rare Word Similarity data sets were developed in our group. they contain pairs of Portuguese words together with scores representing their similarity, assigned by human annotators.

The task of *conceptual categorization* consists in clustering a set of words into categories taking into account the semantic relations across those words. For example, given the words ”apple”, ”orange”, ”sun” and ”moon”, the task would be to cluster the first two words in one group, corresponding to a fruit related category, and the last two words in a second group, corresponding to an astronomy related category.

The LX-ESSLLI 2008, LX-Battig and the LX-AP datasets were also developed in our group for Portuguese. They permit to evaluate the performance of a resolver for conceptual categorization that consists in clustering the semantic vectors of the words in a predefined number of categories.

Some of these evaluation data sets were already put to use by Oliveira (2017), who evaluated the word embeddings also developed in our group, LX-DSemVectors, under these new tasks by running them over the the LX-SimLex-999, LX-Rare Words Similarity and LX-WordSim-353 datasets. Saleiro et al. (2017) have also used data sets for word similarity to evaluate word embeddings they trained over a corpus of Tweets. This author also used the Portuguese word embeddings for Portuguese from the Facebook team (Bojanowski et al., 2016), with no evaluation however reported in the respective paper, and evaluated them on the LX-SimLex-999, LX-Rare Words Similarity and LX-WordSim-353 datasets, obtaining 0.34ρ, 0.34ρ and 0.43ρ, respectively — these scores are worse than our best model reported in the present paper, in Section 6..

Against this background, our objectives are twofold: on the one hand, to take advantage of the larger training data set that we developed (described in detail in Section 4.) to develop an enhanced version of our distributional model, which expectedly would ensure a better capture of the linguistic properties of Portuguese words given it can be trained on a substantially larger corpus; on the other hand, and in confluence with the first objective, to take advantage of the additional data sets for intrinsic evaluation we have just distributed in order to support the development a finer tuned distributional model that more appropriately capture the linguistic properties of the Portuguese words.

3. Methods

For the training of the word embeddings for Portuguese, we resorted to the Skip-Gram model, consisting of a shallow neural network that we implemented using the Gensim framework (Řehůřek and Sojka, 2010).

The training of the Skip-Gram model consists on the learning of a distributional semantics space by predicting the words in the neighborhood for the target word of interest, using a sliding window over the words in a given corpus.

With the model eventually trained, each word is mapped to a specific vector in the resulting semantic space. By scoring the distance between the vectors corresponding to two given words along with some metric, typically the cosine distance, syntactic and semantic properties and relations of these words may be captured and support natural processing tasks and applications.

A Skip-Gram model typically supports tasks with better performance when it is trained with larger data sets. For the gathering of the largest possible data set, we used the largest raw corpora previously gathered for Portuguese and publicly described in (Rodrigues et al., 2016), and expanded it with new data massively web-crawled mostly from newspapers.

For the evaluation and tuning of the resulting distributional semantics space, we resorted to a range of evaluation data sets. These include the analogy test set LX-4WAnalogies already developed and used by (Rodrigues et al., 2016), but also several more, namely the datasets for Portuguese recently made available by Querido et al. (2017): LX-WordSim-353, LX-SimLex-999, LX-Rare Word Similarity, LX-ESSLLI 2008, LX-Battig and the LX-AP. These evaluation datasets were obtained by translating into Portuguese similar ones existing for English, which have been translated also into other languages (Hassan and Mihalcea, 2009) (Joubarne and Inkpen, 2011) (Camacho-Collados et al., 2015) (Freitas et al., 2016) (Cinková, 2016). Given the latter are mainstream datasets, and that these datasets for Portuguese are the most prominent ones existing for this language, this allows thus to develop word embeddings that are mainstream for Portuguese and are comparable with those in English provided large enough data is available for their training.

In the next Section 4., we present in more detail the datasets we used.

4. Datasets

To the best of our knowledge, the largest raw text corpus ever gathered for the Portuguese language, with 1.7 billion tokens, and publicly described in a publication is presented in (Rodrigues et al., 2016). To this data set, we added now over 500 million tokens, corresponding to almost 25 million new sentences, increasing the vocabulary from 873,909 to 1,172,295 items.

These new sentences were mainly gathered in a year time span resorting to the continued crawling of newspaper articles, and the outcome of this crawling was subject to cleaning and tokenization. Table 1 describes the sources of these texts and their respective size.

Once a distributional semantics model is trained with the method described in the previous Section 3., its evaluation can then be accomplished with the support of the evaluation data sets. Next, we provide more information about these datasets, summarized in Table 2.

These corpora can be grouped into three major categories, as summarized in Table 2, namely the ones containing the so-called word analogies, those with pairs with lexical similarity scores and the ones with sets of words supporting conceptual categorization. They are intended to support the evaluation of an equal number of types of natural language processing tasks.

As described above, in the word analogy task, the goal is to complete an analogy with a fourth word, for which only the other three words are given, the accuracy is measured by counting the percentage of given correct answers. The LX-4WAnalogies data set is used to evaluate the performance of resolvers of this word analogy tasks, scored in terms of accuracy. The LX-4WAnalogies contains 17,558 entries of word analogies, each analogy entry containing a set of four words, making a total of 70,232 tokens.

Corpus	**Tokens**	**Sentences**	**Cat.**
Libreoffice	1,456	995	Term.
MSTerminology	38,820	13,030	Term.
Semanario	45,686	2,148	News
JmMadeira	229,197	10,280	News
WikipediaIT	673,932	24,723	Wiki
Abola	838,439	29,260	News
TSF	1,432,332	63,609	News
RegiaoDeLeiria	1,528,711	97,802	News
TMadeira	3,015,973	153,949	News
JornalDoFundao	3,285,362	131,274	News
Sabado	4,527,134	209,848	News
ORegional	4,652,416	203,243	News
OJogo	4,822,878	702,578	News
Euronews	5,355,977	130,672	News
SicNoticias	6,088,051	188,562	News
Lux	7,499,680	358,975	News
IOnline	8,037,427	332,573	News
Sol	8,288,925	323,939	News
Visão	8,987,410	274,193	News
Resistir.info	13,576,530	297,266	News
DN	14,664,275	565,424	News
TVI24	15,332,649	609,173	News
OMirante	17,683,489	705,868	News
CorreioManha	19,171,780	930,494	News
Ocasiao	23,105,092	1,158,281	Sales
OInterior	24,775,869	793,259	News
UniLeipzig	24,933,538	1,000,000	News
JNegócios	27,705,401	791,931	News
Destak	28,137,748	1,148,997	News
DinheiroVivo	37,108,769	1,975,316	News
JornalDiario	43,838,564	1,924,182	News
Expresso	47,272,332	1,893,705	News
Zwame	54,238,235	3,449,424	Forum
JRC-acquis	75,911,681	3,684,145	Law
Total	536,805,758	24,179,118	

Table 1: Novel Portuguese corpora used for the training of word embeddings

The task of lexical similarity consists of indicating a score representing the level of semantic similarity between two words that are entered as input. The LX-WordSim-353, LX-SimLex-999, and LX-Rare Word Similarity data sets contain pairs of words together with scores representing their similarity, assigned by human annotators.

The LX-WordSim-353 is the smallest of the three similarity data sets, it contains 352 entries, each with two tokens. It follows in size the LX-SimLex-999 with 999 entries and the largest similarity data set, the LX-Rare Word Sim. containing 2,034 entries, both data sets have approximately two tokens for each entry.

They are used to evaluate the resolvers of the lexical similarity task, where the cosine distance of the vectors was measured and compared with the original human score, and the Spearman's rank correlation coefficient metric is calculated over those individual scores.

The Spearman's rank correlation coefficient metric measures the correlation between two rank-ordered scales, in this case, between the resolvers evaluation and the original human score for semantic similarity.

Data set	Type	Evaluation Metric	# Entries	Tokens (avg.)
LX-4WAnalogies	Analogy	Accuracy	17,558	70,232 (4)
LX-WordSim-353	Similarity	Spearman's rank	352	715 (2)
LX-SimLex-999			999	2,051 (2.05)
LX-Rare Word Sim.			2,034	4,342 (2.13)
LX-ESSLLI 2008	Categorization	Purity	44 (6 clusters)	44 (1)
LX-Battig			82 (10 clusters)	87 (1.06)
LX-AP			402 (21 clusters)	442 (1.09)

Table 2: Evaluation datasets by type with corresponding evaluation metric

The task of conceptual categorization consists in clustering words into categories. Resorting to the LX-ESSLLI 2008, LX-Battig and the LX-AP datasets, it is possible to evaluate the performance of a resolver for conceptual categorization. The LX-ESSLI 2008 is the smallest of the three conceptual categorization data sets, it contains 44 entries for 6 categories which results in approximately 7 entries per category, each entry is a single word/token. The LX-Battig contains 82 entries for 10 categories, approximately 8 entries per category and 1 word/token per entry. The largest of the categorization data sets, the LX-AP, contains 402 entries for 21 categories, approximately 19 entries per category.

Categorization consists in clustering the semantic vectors of the words in a predefined number of categories. The k-means clustering method is applied and resolvers are scored with the purity metric. Purity measures the accuracy obtained by assigning each cluster to the class which is most frequent in the cluster (Christopher et al., 2008).

5. Experiments

In order to develop the word embeddings for Portuguese, we experimented with different arrangements in terms of the training data set used, taking into account the best practices that have been reported in the literature. We undertook four major experiments where we resorted to the best parameterization setup of the Skip-gram models for Portuguese empirically determined in (Rodrigues et al., 2016), which optimizes the performance of the word analogy task. This corresponds to a Skip-gram parameterization that consists of a vector dimension of 400 units, a 10-word window, a learning-rate of 0.025, using 15 negative samples and a total of 5 epochs.

These four experiments (exp1 to exp4) are described below. Their results are presented in the next Section 3.

- First (exp1), we trained a model using the novel 500+ million token data set we collected, described in section 4.. This experience aims at getting an initial assessment as to whether these new data are cleaned and fit enough to train the word embeddings, by observing whether the performance results on this task are in line with the respective results previously obtained by Rodrigues et al. (2016), even tough with a data set of a different size and composition.

- Second (exp2) we trained a model using the 1+ billion token sub-data set of the whole 1.7 million token dataset that was previously gathered in Rodrigues et al. (2016). This is a replication of a previous experiment in (Rodrigues et al., 2016), only with one difference from it, namely the use of a new and improved version of Gensim (from 0.13.1 to 2.1). The goal here is to make sure that this change is not induce any degradation of the previous level of performance of the best model developed in that previous work.

- Third (exp3), we trained a model on the data set that results from the merging of the two datasets used in exp 1 and exp 2. The aim is to assess whether more data support better performance of the analogy resolver.

- Fourth (exp4), a model was trained with the largest possible data set, which contains the whole 1.7 billion token dataset from Rodrigues et al. (2016) plus the 0.5 billion novel dataset we gathered and describe in the present paper. The goal is to make use of the largest amount of data in order to observe if the models gain from it in accuracy.

Given the large training data, the training of the distributional semantics space is an intensive computational and time-consuming task. The time required for the most intensive task was approximately 220 hours (around 9 days). This value was obtained using 15 dedicated CPUs[1].

In addition to the above four experiments, we carried out three more experiments, which use the same training data as the one used in exp3.

- In a fifth experiment (exp5), the number of training epochs was increased from 5 to 30 in order to assess its impact on the accuracy, given that with a higher number of training epochs the distributional semantic space will be fitter and impact the resolver.

- A sixth experiment (exp6) was undertaken by increasing the vector dimension from 400 to 500 units in order to assess if the vector dimension is enough for the amount of data or a higher data representation is needed.

- In a seventh experiment (exp7), we increased the number of training epochs from 5 to 15 and the negative samples from 15 to 30 in order to assess if repeated negative samples may improve the tasks not related to word analogy.

[1] Intel Core Processor (Haswell, no TSX) @ 2.50GHz model.

Data set	exp1	exp2	exp3	exp4	exp5	exp6	exp7
LX-4WAnalogies	33.5%	43.4%	45.8%	**47.1%**	46.5%	45.9%	45.8%
LX-WordSim-353	0.4614ρ	0.5089ρ	0.5002ρ	**0.5146**ρ	0.4923ρ	0.4937ρ	0.4869ρ
LX-SimLex-999	0.3190ρ	0.3265ρ	0.3341ρ	**0.3502**ρ	0.3239ρ	0.3351ρ	0.3303ρ
LX-Rare Word Sim.	0.3196ρ	0.3325ρ	0.3520ρ	**0.3618**ρ	0.3457ρ	0.3526ρ	0.3501ρ
LX-ESSLLI 2008	0.7045	0.6364	0.6818	0.5909	0.6364	0.5455	**0.7045**
LX-Battig	0.6235	0.6941	**0.8589**	0.8000	0.7294	0.7059	0.7294
LX-AP	0.5297	0.5845	**0.6575**	0.6438	0.5982	0.6301	0.6187

Table 3: Results from the seven experiments (columns) evaluated over the word analogy, semantic similarity and conceptual categorization tasks (rows), with best score in bold.

6. Results and Discussion

In this Section, we discuss the results obtained with the different experiments described in Section 5., which are summarized in Table 3.

Given that the new additional data set originates mainly from the crawling of newspapers, we proceeded with the first experiment (exp 1) in order to assess if this data set (0.5 billion tokens) is suitable for the training of a distributional semantics space.

That happens to be the case as the score obtained for exp1 (33.5% accuracy) is in-line with the best score of the experiments reported in (Rodrigues et al., 2016) (37.7%), for the word analogy task, the only evaluation data set available then.

In the second experiment (exp 2), this state-of-the-art model from Rodrigues et al. (2016) is reproduced with that data set used by then (1 billion tokens) but with the improved version of the Gensim framework.

No degradation of the previous performance was observed, with exp 2 bringing even a better result (43.4%) than the one obtained before (37.7%), thus indicating that, when used, this new version of Gensim is responsible for the improvement of the output model.

In the third experiment (exp 3), a new model was trained on the data set used for the state-of-the-art model from (Rodrigues et al., 2016) merged with the newly collected data set.

Comparing that new model against the models trained with each one of these two data sets separately (exp1 and exp2), a higher accuracy is obtained for almost all evaluation data sets except one, namely LX-WordSim-353 and even in this case with an inferior score by a very thin margin (0.5002ρ against 0.5089ρ). This confirms our hypothesis that the merged data sets (1.5 billion tokens) would support the development of a more appropriate distributional model.

The fourth experiment (exp 4) also confirms the hypothesis that by using the largest amount of data (2.2 billion tokens), one obtains the best scores in the word analogies task and in the semantic similarity tasks. This model scored a 47.1% accuracy with the LX-4WAnalogies evaluation data set, and 0.5146ρ, 0.3502ρ and 0.3618ρ, respectively with LX-WordSim-353, LX-SimLex-999 and LX-Rare Word Sim data sets.

It is worth noting that the scores obtained for the conceptual categorization tasks are lower than the ones obtained with some models trained on smaller data sets.

The fifth, sixth and seventh experiment (exp 5, 6 and 7) used the same training data set as exp 3 (1.5 billion).

By increasing the number of training epochs, in exp 5 only one task got an improvement over exp3, namely the word analogy with 46.5% accuracy, against 45.8% in exp 3. This appears to indicate that the increase of the number of epochs is not bringing a clear advantage to the output model.

In exp 6, by increasing the vector dimension, there was a residual improvement over exp 3 and only on the LX-4WAnalogies (delta of 0.1%), the LX-SimLex-999 (delta of 0.001ρ) and LX-Rare Word Sim (delta of 0.0006ρ). No clear improvements were thus gained by increasing the vector dimension.

Regarding exp 7, where the negative sampling was increased, the results are identical to the word analogy task and worst for all of the data sets in the semantic similarity tasks, when compared to exp 3. The model obtains an improvement in the conceptual categorization task for the LX-ESSLLI data set, matching the best results obtained in exp 1 with a 0.7045 score. The two other conceptual categorization data sets got a worse result than exp 3 though.

The attempts of re-tuning the parameters with exp 5, 6 and 7 yielded no substantive improvements. This seems to indicate that the parameters tuned in (Rodrigues et al., 2016) when using 1 billion tokens by then still deliver pretty good results.

Distributional models for Portuguese previously published in the literature were evaluated only against the word analogy task, with the LX-4WAnalogies data set, scoring 37.7% and 46.2%, respectively by Rodrigues et al. (2016) and Hartmann et al. (2017). Our initial results with the new 2.2 billion token data set under this task, with 47.1% in exp 4, indicate that the model described here is the one with the best score for this task.

The use of larger training data clearly indicates that, as expected, it is a most important factor for the improvement of the scores of the analogy and semantic similarity data sets.

7. Resources distributed

All the distributional semantics models whose development is reported in the present paper are distributed at http://github.com/nlx-group under a Attribution 4.0 International (CC BY 4.0) license. These models are termed as **LX-DSemVectors 2.2b** and users of these language resources should refer to them by citing both the present paper and (Rodrigues et al., 2016).

8. Final remarks

In this paper, we presented an advanced distributional model (aka word embeddings) for Portuguese, LX-DSemVectors 2.2b, with the largest vocabulary and the best intrinsic evaluation scores published so far.

This model permits to score 47.1% accuracy in the word analogy task, with the LX-4WAnalogies data set; 0.5146ρ, 0.3502ρ and 0.3618ρ in the lexical similarity task, respectively with the evaluation data sets LX-WordSim-353, LX-SimLex-999, and LX-Rare Word Similarity; and a purity of 0.5909, 0.8000 and 0.6438 in the conceptual categorization task, respectively with the LX-ESSLLI 2008, LX-Battig and the LX-AP evaluation data sets.

This model is distributed for free under a most permissive license.

The reproducibility of the results reported in this paper can be verified by using this model and obtaining the evaluation data sets, also available for free and under a most permissive license.

Acknowledgment

The present research was partly supported by the Infrastructure for the Science and Technology of the Portuguese Language (CLARIN Língua Portuguesa), by the National Infrastructure for Distributed Computing (INCD) of Portugal, and by the ANI/3279/2016 grant.

9. Bibliographical References

Bordes, A., Chopra, S., and Weston, J. (2014). Question answering with subgraph embeddings. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Chen, Y.-N., Wang, W. Y., and Rudnicky, A. I. (2014). Leveraging frame semantics and distributional semantics for unsupervised semantic slot induction in spoken dialogue systems. In *Proceedings of the Spoken Language Technology Workshop (SLT), 2014 IEEE*, pages 584–589. IEEE.

Christopher, D. M., Prabhakar, R., and Hinrich, S. (2008). Introduction to information retrieval. *An Introduction To Information Retrieval*, 151:177.

dos Santos, C. and Zadrozny, B. (2014). Learning character-level representations for part-of-speech tagging. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1818–1826.

dos Santos, C., Guimaraes, V., Niterói, R., and de Janeiro, R. (2015). Boosting named entity recognition with neural character embeddings. In *Proceedings of the NEWS 2015 The Fifth Named Entities Workshop*, page 25.

Fonseca, E. R. and Aluísio, S. M. (2016). Improving POS tagging across portuguese variants with word embeddings. In *Proceedings of the International Conference on Computational Processing of the Portuguese Language*, pages 227–232. Springer.

Fonseca, E. R., Rosa, J. L. G., and Aluísio, S. M. (2015). Evaluating word embeddings and a revised corpus for part-of-speech tagging in portuguese. *Journal of the Brazilian Computer Society*, 21(1):2.

Garvin, P. L. (1962). Computer participation in linguistic research. *Language*, 38(4):385–389.

Hartmann, N., Fonseca, E., Shulby, C., Treviso, M., Rodrigues, J., and Aluisio, S. (2017). Portuguese word embeddings: Evaluating on word analogies and natural language tasks. *arXiv preprint arXiv:1708.06025*.

Hermann, K. and Blunsom, P. (2014). Multilingual distributed representations without word alignment. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Ling, W., Luís, T., Marujo, L., Astudillo, R. F., Amir, S., Dyer, C., Black, A. W., and Trancoso, I. (2015). Finding function in form: Compositional character models for open vocabulary word representation. *arXiv preprint arXiv:1508.02096*.

Machado, M. T., Temporal, J. C. A. N., Pardo, T. A. S., and Ruiz, E. E. S. (2017). Mineração de tópicos e aspectos em microblogs sobre dengue, chikungunya, zika e microcefalia. In *Proceedings of the XXXVII Congresso da Sociedade Brasileira de Computação - 17º WIM - Workshop de Informática Médica*.

Oliveira, H. G. (2017). Unsupervised approaches for computing word similarity in portuguese. In *Proceedings of the Portuguese Conference on Artificial Intelligence*. Springer.

Querido, A., Carvalho, R., Rodrigues, J., Garcia, M., Silva, J., Correia, C., Rendeiro, N., Pereira, R., Campos, M., and Branco, A. (2017). LX-LR4DistSemEval: a collection of language resources for the evaluation of distributional semantic models of portuguese. In *Proceedings of the XXXII Encontro Nacional da Associação Portuguesa de Linguística (ENAPL)*.

Rodrigues, J., Branco, A., Neale, S., and Silva, J. (2016). LX-DSemVectors: Distributional semantics models for portuguese. In *Proceedings of the International Conference on Computational Processing of the Portuguese Language*, pages 259–270. Springer.

Saleiro, P., Sarmento, L., Rodrigues, E. M., Soares, C., and Oliveira, E. (2017). Learning word embeddings from the portuguese twitter stream: A study of some practical aspects. In *Proceedings of the Portuguese Conference on Artificial Intelligence*. Springer.

10. Language Resource References

Al-Rfou, R., Perozzi, B., and Skiena, S. (2013). Polyglot: Distributed word representations for multilingual nlp. *CoNLL-2013*, page 183.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Camacho-Collados, J., Pilehvar, M. T., and Navigli, R. (2015). A framework for the construction of monolingual and cross-lingual word similarity datasets. In *Proceedings of the ACL Workshop on Evaluating Vector Space Representations for NLP*, pages 1–7.

Cinková, S. (2016). Wordsim353 for czech. In *Proceedings of the International Conference on Text, Speech, and Dialogue*, pages 190–197. Springer.

Freitas, A., Barzegar, S., Sales, J. E., Handschuh, S., and Davis, B. (2016). Semantic relatedness for all (languages): A comparative analysis of multilingual semantic relatedness using machine translation. In *Proceedings of the Knowledge Engineering and Knowledge Management: 20th International Conference, EKAW 2016*, pages 212–222. Springer.

Hartmann, N., Fonseca, E., Shulby, C., Treviso, M., Rodrigues, J., and Aluisio, S. (2017). Portuguese word embeddings: Evaluating on word analogies and natural language tasks. *arXiv preprint arXiv:1708.06025*.

Hassan, S. and Mihalcea, R. (2009). Cross-lingual semantic relatedness using encyclopedic knowledge. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1201. Association for Computational Linguistics.

Joubarne, C. and Inkpen, D. (2011). Comparison of semantic similarity for different languages using the google n-gram corpus and second-order co-occurrence measures. In *Proceedings of the Canadian Conference on Artificial Intelligence*, pages 216–221. Springer.

Mikolov, T., Yih, W., and Zweig, G. (2013). Linguistic regularities in continuous space word representations. In Lucy Vanderwende, et al., editors, *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pages 746–751. The Association for Computational Linguistics.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Querido, A., Carvalho, R., Rodrigues, J., Garcia, M., Silva, J., Correia, C., Rendeiro, N., Pereira, R., Campos, M., and Branco, A. (2017). LX-LR4DistSemEval: a collection of language resources for the evaluation of distributional semantic models of portuguese. In *Proceedings of the XXXII Encontro Nacional da Associação Portuguesa de Linguística (ENAPL)*.

Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA.

Rodrigues, J., Branco, A., Neale, S., and Silva, J. (2016). LX-DSemVectors: Distributional semantics models for portuguese. In *Proceedings of the International Conference on Computational Processing of the Portuguese Language*, pages 259–270. Springer.