LX-LR4DistSemEval: a collection of language resources for the evaluation of distributional semantic models of Portuguese

Andreia Querido, Rita de Carvalho, João Rodrigues, Marcos Garcia†, João Silva, Catarina Correia, Nuno Rendeiro, Rita Valadas Pereira, Marisa Campos, António Branco

Faculty of Sciences of the University of Lisbon

†Faculty of Philology, University of Coruña

Abstract:

In this paper we describe a collection of publicly available data sets for Portuguese that are suitable for the evaluation of distributional semantics models in lexical similarity tasks and in conceptual categorization tasks. These data sets were adapted from English gold-standard test sets, allowing any Portuguese distributional semantics model to be evaluated and also to be compared to mainstream results that have been obtained for this language. We also present an online service that showcases some functionalities of the distributional semantics models.

Keywords: distributional semantics, data sets, evaluation, Portuguese.

Palavras-chave: semântica distribucional, conjuntos de dados, avaliação, português.

1. Introduction

Distributional semantics explores the principle that expressions with similar syntactic and/or semantic properties are used in similar contexts (Firth, 1957; Harris, 1954). A distributional semantic model associates each expression to a vector of real numbers where in its components is encoded the information of the frequency in which the expression co-occurs with other expressions. With a distributional semantic space defined for a set of expressions, syntactic and semantic properties between input expressions can be determined, by the distance between the vectors of those expressions, using the cosine distance.

With the increasing volume of digital texts and larger computational power available, distributional semantics has received renewed attention as a way of enriching the resources and tools used in natural language processing, contributing to a better performance of tasks carried out in this area (Collobert and Weston, 2008). That is the case, for instance, of speech recognition tasks (Mikolov *et al*., 2009), morphosyntactic annotation, sentiment analysis (Li

and Jurafsky, 2015), named entities recognition, similarity relations between words (Mikolov *et al*., 2013), formal semantics (Baroni *et al*., 2014), etc.

A distributional semantic model trained for the computational processing of Portuguese is available as well as a set of instructions on how to make it work (Rodrigues *et al*., 2016). The creation of this distributional vectors for the Portuguese language was a first step to improve the resources and tools for the Portuguese language in this area.

In the current paper, we present the first Portuguese publicly available data sets that have been prepared to evaluate the suitability of Portuguese distributional models. These data sets not only enable the evaluation of a Portuguese distributional model, but also help to bring distributional semantic of Portuguese in line with the work that has been done in other languages in this respect.

For the evaluation of distributional semantics models, there have been three tasks used in the literature: a) the analogy task, b) the lexical similarity task and c) the conceptual categorization task.

In the **analogy task** the goal is to find a missing word in a relation between two pairs of words: for instance, for the pairs "Berlin is to Germany as *x* is to Portugal", the task is to determinate the expression that should instantiate *x*.

On the other hand, when there are two words and a score should be given to them in some predefined scale, which expresses the similarity relation between them, we are faced with a **lexical similarity task**.

Lastly, in a **conceptual categorization task,** the goal is to cluster a set of words in a predefined number of categories, according to the relation between those words.

An accurate distributional semantic space should have encoded in its vectors the syntactic and semantic properties necessary to solve the tasks mentioned above as successfully as possible.

For the evaluation of Portuguese distributional vectors, there is only one data set available, suitable for the evaluation on analogy tasks, namely LX-4WAnalogies (Rodrigues *et al*., 2016). This set is the result of the translation to Portuguese of an English test set (Mikolov *et al*., 2013). As far as we know, at the time of writing, there was not any suitable and publicly

available data set for Portuguese that could be used to evaluate a distributional semantic model on a lexical similarity task or on a conceptual categorization task.

In the Section 2, we will describe the methodology adopted in the creation of the data sets for Portuguese that are suitable for the evaluation of distributional semantics models. In the Section 3, we will present a detailed description of the data sets for the evaluation of lexical similarity tasks and in the Section 4, we will present a detailed description of the data sets for the evaluation of conceptual categorization tasks. After that, in the Section 5, we will present an online service that showcases a distributional semantics model and its use in lexical similarity tasks. Finally, in the Section 6, concluding remarks will be presented.

2. Methodology

The test sets described in the next sections result from a translation from English to Portuguese of data sets widely used in the literature of English. All the translators involved in this undertaking were skilled, native Portuguese-speaking language experts.

We considered different stages to create each Portuguese gold-standard data set: a) the double-blind translation of the words; b) the adjudication of the translations performed; c) and the attachment of scores in the data sets for lexical similarity tasks.

Each one of the data sets was translated by two translators, working independently of each other, under a double-blind scheme. To ensure a reliable and not biased data set, the translators were free to translate each word in the way they considered the best. They did not have access to the scores in English and they did not know what would be the goal of the translation. The only context they had was the other word of the pair (in the data sets for the lexical similarity task) or the categories to which the words belonged (in the data sets for the conceptual categorization task).

In the second stage, the translations put forward by each one of the two translators were compared and a third expert adjudicated those cases in which the translators had provided different translations for the same expressions. The adjudicator could see the English words and some guidelines were created for the adjudication task. For the nouns and adjectives, it was decided to use the lemma, or the dictionary form, i.e., they were translated to the masculine singular whenever the English form was ambiguous. When the adjudicator could choose between two correct possible translations, one being a multiword, the adjudicator was instructed

to choose the single word translation. For example, the word "sweater" was translated to *camisola* and to *camisola de lã* by different translators. The adjudicator chose the one that is not a multiword (*camisola*). More information on the guidelines will be given in each data set description below, because some of them are specific to each set.

Regarding the test sets used in lexical similarity tasks, a score that measured the similarity between the words was always provided. The score was attached to the pair at stake taking into account the words in Portuguese. Further on, while we describe each one of the test sets, we will explain how the annotation proceeded in each case.

3. Data sets for evaluation in lexical similarity tasks

The data sets used in lexical similarity tasks have a similar structure: they are a list of word pairs, and to each word pair a score is given and this value expresses the semantic similarity degree between the two words of the pair. On average, more than two native-speakers assigned a score to each pair. The final score is the average of the annotators' scores.

The data sets created for Portuguese are: the LX-WordSim-353, the LX-SimLex-999 and the LX-Rare Word Similarity Data set. They are publicly available in LX-Center site[1]. Each one will be presented in turn below.

3.1. LX-SimLex-999

The LX-SimLex-999 was created from SimLex-999 (Hill *et al*., 2015) which, in turn, was based in the University of South Florida Free Association Database (USF) (Nelson *et al.*, 2014).

There were strict guidelines to create SimLex-999. Both words in each pair have the same morphosyntactic category and the multiword expressions and named entities were excluded from that data set. Besides the morphosyntactic category criteria, the level of concreteness of each word was important. The word pairs in the USF data set had been tagged with a concreteness level that was provided by human annotators, on a scale of 1-7. In the

[1]http://metashare.metanet4u.eu/go2/lx-lr4distsemeval

creation of SimLex-999, this classification was taken into account and the pairs in which one of the concepts was more concrete than the other were not included.

The result was 999 word pairs organized in the following way: 666 pairs of noun-noun, 222 pairs of verb-verb and 111 pairs of adjective-adjective. Each pair received a score on a scale from 0 (totally unrelated) to 6 (very similar).

To create the LX-SimLex-999, two translators translated the 999 English word pairs to Portuguese. After that, the comparison between the translations showed that the translators agreed in 67.3% of the pairs (if we count the words, they agreed in 80.2% of the 1998 words of the data set). In the majority of the cases in which the translators disagreed, that happened due to the several possibilities of translation into Portuguese of an English term. Faced with these situations, the adjudicator should pick just one translation possibility in Portuguese. Below (Table 1), we show examples of word pairs in which the translators disagreed but the two possible translations are correct.

English word 1	**English word 2**	**POS**	**Translation 1**		**Translation 2**		**Adjudication**	
			Word1	Word2	Word1	Word2		
insane	crazy	A	insano	maluco	insano	louco	insano	maluco
teacher	helper	N	professor	auxiliar	professor	ajudante	professor	ajudante
communicate	pray	V	comunicar	orar	comunicar	rezar	comunicar	rezar

Table 1: Examples of word pairs translations in LX-SimLex-999

Concerning score attachment, we have followed the same methodology used for SimLex-999: The annotators assign scores in a scale from 0 (totally unrelated words) to 6 (very similar words). The final score of each pair is the average of the scores given by each annotator,

mapped into a scale from 0 to 10. Chart 1 shows the score distribution of LX-SimLex-999 and Chart 2 shows the score distribution of SimLex-999.

As it was done for English, we computed the inter-rater agreement using Spearman ρ correlation coefficient between the annotators. The inter-rater agreement in Portuguese was ρ = 0.688, which is in line with the value of the SimLex-999 inter-rater agreement (ρ = 0.67).

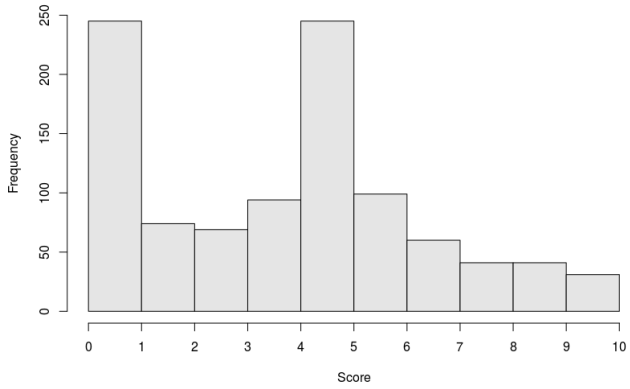

Chart 1: Scores distribution in LX-SimLex-999

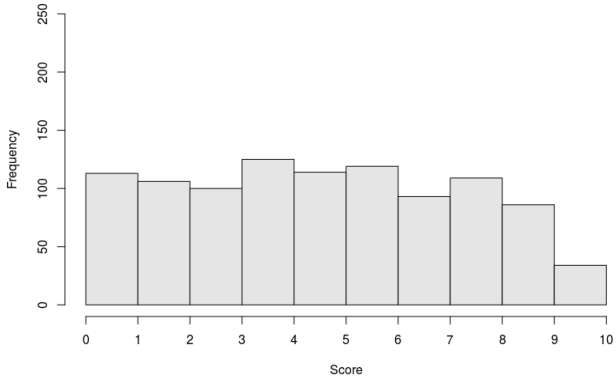

Chart 2: Scores Distribution in SimLex-999

In LX-SimLex-999, two native-speakers assigned a score to each pair, on the other hand, in SimLex-999, there were 500 annotators giving a score to each pair. The disparity between the distribution of the Portuguese and the English results can be understood by this difference in the number of annotators. In English there is a larger number of annotations to average, thus the final scores are more homogeneous. Despite the disparity of the distribution of data, the

median for the Portuguese data set (4.167) is almost the same as the median for the English data set (4.670).

In the near future we aim to increase the number of annotations of LX-SimLex-999. Nevertheless, as it is, this gold-standard test set already enables a reliable evaluation of distributional semantic models in a lexical similarity task, allowing the comparison between the work done in Portuguese and in the other languages.

3.2. LX-Rare Word Similarity Data set

The LX-Rare Word Similarity Data set was created from Stanford Rare Word (RW) Similarity data set (Luong et al., 2013). This list contains 2 034 words (1 017 pairs of words). All the words were extracted from Wikipedia and from WordNet (Miller, 1995), a lexical database where the concepts are grouped into sets of synonyms.

The construction of this list followed this procedure: a) firstly, a list of rare words was selected from Wikipedia, b) after that, each rare word was paired with a related word picked from WordNet. Rare words are those words that have between 5 000 to 10 000 occurrences in Wikipedia.

In the end, the result was a set of word pairs in which one of the words is rare and the other one, which can be rare or not, is related to the first word by some WordNet relation - it can be an hyponym, hyperonym, meronym, holonym or attribute of the former.

The comparison between the translations of this list into Portuguese showed that the translators disagreed on 58.5% of the words translated. We think that this high number of disagreements is due to the rare words in the data set. When the words are rare, it is more probable that the translators have difficulties to translate the word without context and, thus, it is more likely that they disagree in the translation of it. In Table 2, we show examples of word pairs in which the translators disagreed.

English word 1	**English word 2**	**Translation 1**		**Translation 2**		**Adjudication**		**score**
		Word1	Word2	Word1	Word2			
soulless	insensitive	sem alma	insensível	desalmado	insensível	desalmado	insensível	7,5
preordained	predetermine	predes-tinado	predeter-minar	preordenado	predeter-minado	predestinado	predeter-minado	7
preservers	worker	presser-vadores	trabalhador	defensores	trabalhador	defensores	traba-lhador	1

Table 2: Examples of word pairs translations in LX-Rare Word Similarity Data set

Ten annotators attached the scores in LX-Rare Word Similarity data set the scores were given by four annotators. Each pair of words received a human judgment on a scale from 0 (totally unrelated words) to 10 (very similar words), as in Stanford Rare Word (RW) Similarity data set.

Taking into account that for the creation of Stanford Rare Word (RW) Similarity Dataset the words of the pairs should be related, mandatorily, the English scores distribution (Chart 3) reflects this criterion.

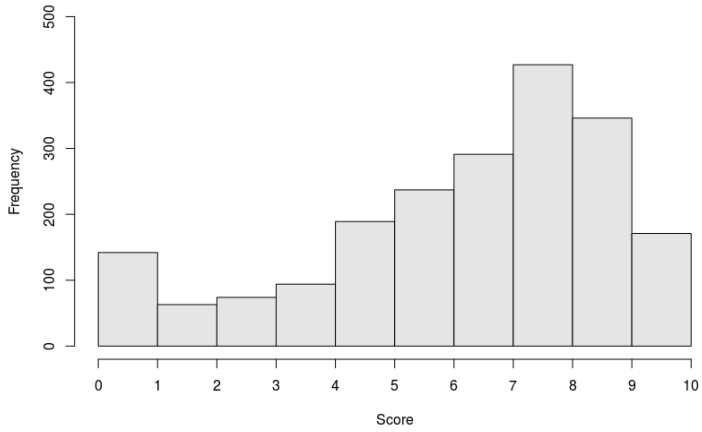

Chart 3: Scores distribution in Stanford RW Similarity Data set

Observing the chart of the Portuguese scores (Chart 4), we can see that the distribution of scores is different, since there are fewer pairs with high scores. Due to the existence of at least one rare word in each English pair, we can speculate that the translation challenges in this data set are bigger than the challenges found in data sets with pairs of frequent words. Therefore,

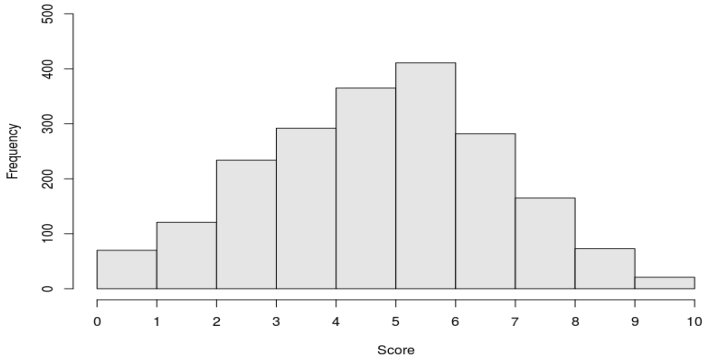

Chart 4: Scores distribution in LX-Rare Word Similarity Data set

it is possible that the outcome of the translation - the LX-Rare Word Similarity data set - may have distanced itself from the original data set and that the level of similarity in each pair has changed.

When more sophisticated lexicographic instruments for the Portuguese language become available, for example with a more complete Portuguese WordNet, it would be interesting to analyse the relations between the words in each pair and their level of rareness. Only then would it be possible to draw definitive conclusions about the differences between the English and the Portuguese data sets.

3.3. LX-WordSim-353

The LX-WordSim-353 was created from WordSim-353 (Agirre *et al*., 2009). As the name suggests, this data set contains 353 pairs of words. Both words in each pair can have different morphosyntactic categories. The data set is made of nouns, adjectives, verbs and named entities, and has no multiwords.

Originally (Finkelstein, *et al.*, 2002), each pair of words received a human judgement on a scale from 0 (totally unrelated words) to 10 (very much related or identical words).

Agirre *et al.* (2009) observed that the numeric annotation did not distinguish between similar and related pairs. In an attempt to know which was the true relation between the words of each pair, they advanced with a different approach in the annotation of this data set. Thus, the annotators should classify all pairs as being synonyms, antonyms, identical, hyperonym-hyponym, sibling terms (terms with a common hyperonymy), meronym-holonym or none-of-the-above. With this annotation, they could determine which pairs had a relation of similarity among the two words and which pairs had related words. At the end, they distinguished between the pairs with related words and the pairs with similar words. In the word pairs categorized as synonyms, antonyms, identical and hyperonym-hyponym, there was a relation of similarity between both words. In the word pairs categorized as sibling terms, holonym-meronym or none-of-the-above, which had on average a similarity greater than 5, there was a relation of relatedness between both words.

The LX-WordSim-353 was the outcome of a) the translation of WordSim-353 into Portuguese and b) the annotation of that list with the classification established by Agirre, *et al.* (2009). The translation process followed the same procedures as the translation of the data sets

in the sections above: two translators translated the same data and a third expert adjudicated when there were mismatches. The annotators agreed in the translation of 67,1% of the pairs. When the total number of words in the data set is taken into account, the agreement rises to 81,2%, since the disagreements were not always in both words of the pair.

After the adjudication of the translation, six annotators classified each pair with one of the following categories:

i = identical tokens
s = synonym (at least in one meaning of each)
a = antonyms (at least in one meaning of each)
h = first is hyponym of second (at least in one meaning of each)
H = first is hyperonym of second (at least in one meaning of each)
S = sibling terms (terms with a common hyperonymy)
m = first is part of the second one (at least in one meaning of each)
M = second is part of the first one (at least in one meaning of each)
t = topically related, but none of the above

Table 3 below contains some examples of this data set.

English word1	**English word2**	**Translation**		**Classification**
		Word1	Word2	
bird	cock	ave	galo	H = first is hyperonym of second
coast	shore	costa	litoral	s = synonym
drink	mouth	bebida	boca	t = topically related

Table 3: Examples of word pairs translations in LX-WordSim-353

4. Data sets for evaluation in conceptual categorization tasks

A conceptual categorization task consists in clustering a set of words into a predefined number of categories. The goal is to obtain sets of words which are similar among themselves and that correspond to a semantic category: transports, fruit or tools, for example. Test sets to evaluate this type of task are usually lists of words grouped into categories. The data sets created

for Portuguese are three: LX-ESSLLI 2008, LX-Battig and LX-AP. They are publicly available in LX-Center site[2]. They will be presented in turn.

4.1. LX-ESSLLI 2008

The LX-ESSLLI 2008 data set was created from the ESSLLI 2008 Distributional Semantic Workshop shared-task set[3], made of 44 concrete nouns grouped in 6 semantic categories (4 animate and 2 inanimate). The grouping is done in an hierarchical way following the top 10 properties from the McRae (2005) norms: bird-animal-natural; ground animal-animal-natural; fruit tree-vegetable-natural; green-vegetable-natural; tool-artifact-artifact; vehicle-artifact-artifact.

We kept the organization into the same categories, resulting in a list with the same size as the original data set. Table 4 shows one example of an animate category where it is clear that the context provided by the semantic category was relevant for the translation. For example, "cherry" could also have been translated to *cereja* (the fruit of the cherry tree) but in this case the translation should be *cerejeira* (the cherry tree).

ESSLLI 2008	LX-ESSLLI 2008
fruitTree-vegetable-natural	**árvore de fruto-vegetal-natural**
cherry	cerejeira
banana	bananeira
pear	pereira
pineapple	ananaseiro

Table 4: Example of one category in ESSLLI 2008 and LX-ESSLLI 2008

The translation of the data set was done by two translators and adjudicated by a third one with an agreement among translators of 80%.

4.2. LX-Battig

The LX-Battig was created from Battig test.set (Baroni *et al.*, 2010). This data set has 83 concrete concepts of the following 10 categories: mammals, birds, fish, vegetables, fruit, trees,

[2] http://metashare.metanet4u.eu/go2/lx-lr4distsemeval

[3] http://wordspace.collocations.de/doku.php/data:esslli2008:concrete_noun_categorization

vehicles, clothes, tools and kitchenware. The categories names and the concepts were translated by two translators and adjudicated by a third one. The translators agreed in the translation of 80,7% of the words. Table 5 shows some examples of the categories and concepts.

Battig test.set	LX-Battig
mammals	**mamíferos**
dog	cão
elefant	elefante
cat	gato
fruit	**fruta**
apple	maçã
orange	laranja
grape	uva

Table 5: Examples from Battig test set and LX-Battig test set

4.3. LX-AP

LX-AP was created from the translation of Almuhareb-Poesio (ap) benchmark (Almuhareb and Poesio, 2005). The original data set was created considering three aspects: POS, frequency and ambiguity.

It contains 402 names from 21 categories of WordNet, with 13 to 21 names from each one of those categories. Examples of some categories: feeling, game, time, tree, vehicle, chemical element or motivation (more examples are shown in Table 6).

To estimate the word frequency it was used the British National Corpus. Concerning frequency, ⅓ of the words of the corpus has high frequency (1 000 occurrences or more), ⅓ has medium frequency (between 100 to 1 000 occurrences) and ⅓ has low frequency (5 to 100 occurrences).

The evaluation of the degree of ambiguity of each word was calculated taking into account the amount of senses of each word found in the WordNet. With four or more senses, the word was considered very ambiguous; with two or three meanings, the word would have medium ambiguity; and with one meaning, the word was considered not ambiguous. Each level of frequency and ambiguity is equally represented in the set.

We are aware that a word that is frequent in English can be less frequent in Portuguese and that a word that is ambiguous in English can be less ambiguous in Portuguese. More than translating the original data set, it would be interesting to build a data set that, in Portuguese, would also be balanced in terms of frequency and ambiguity of words. As a possible future

work, an analysis of the frequency of the words using a large Portuguese data set as a reference, and an analysis of the ambiguity of the words using the Portuguese Wordnet would improve this data set. However, because the lexicographic resources required to fulfil those tasks are not available yet, the LX-AP is made of the translation from the English words, resulting in a test set with the same size as the original.

The translation process of this data set from English to Portuguese involved two annotators and a third adjudicator. The translators agreed in the translation of 75,9% of the words.

Almuhareb-Poesio benchmark	LX-AP
state (illness)	**estado (doença)**
asthma	asma
cancer	cancro
cholera	cólera
event (social occasion)	**evento (evento social)**
ball	baile
celebration	celebração
ceremony	cerimónia

Table 6: Examples from Almuhareb-Poesio benchmark and LX-AP

5. LX-Semantic Similarity online service

The datasets described in this paper are used to evaluate models for distributional semantic. To showcase these models and their use in lexical similarity tasks, we created the LX Semantic Similarity online service.[4]

When accessing this service, the user is shown a form presenting two options regarding the mode of functioning of the service. The user can either: a) enter a word and get some of the words most similar to it; or b) enter two words, obtain the cosine distance between them, and interact with a 2-dimensional visualization of the vector-space around those two words.

The following sections present these two modes of functioning, as well as some additional technical details.

[4] http://lxsemsimil.di.fc.ul.pt

5.1 Find the most similar words

Given a word, the 15 words in the model that are closest to it in terms of cosine distance are picked. These words — together with their cosine distance to the input one — are shown to the user as a table ranked by cosine similarity.

The service also displays a word cloud with the 100 words in the model that are closest to the input word. In the word cloud, the font size is proportional to the cosine distance (the font size decreases as the distance increases).

Figure 1 shows a screenshot of the service running in this mode.

Figure 1: Screenshot of the Web demo after entering the word *Aveiro*

5.2 Vector-space visualization

Given two words, the service can also displays the cosine distance between them as well as a 2-dimensional visualization of their neighboring vector-space.

To obtain this visualization, for each word, the set containing the top-100 words most similar to it is selected. These two sets are merged, resulting in a set with at most 200 words, at most (since a word occurring in both sets only appears once). The vectors for these words have very high dimensionality and cannot be directly visualized. To enable their visual representation, the vectors are transformed into a 2-dimensional vector using t-SNE (t-Distributed Stochastic Neighbor Embedding).

t-SNE (van der Maaten and Hinton, 2008) is a dimensionality reduction technique often used to visualize high-dimensional data, which works by assigning a 2 (or 3) dimensional vector to each original vector, while trying to ensure that, in the 2-dimensional visualization, each vector is represented by a point in such a way that similar vectors are nearby points, while dissimilar vectors are represented as distant points.

LX Semantic Similarity makes use of the resulting 2-dimensional representation of the space around the two input words as follows: The data are shown in a plot where the two input words are highlighted in different colors. As t-SNE runs, it tries to find a good placing for the words in the 2-dimensional space, settling in a stable positioning after a few iterations.

The plot is interactive, allowing the user to zoom in and out and to move around the vector-space, and thus better analyze the vector space around each input word, as well as the relative positioning between them and their neighboring words.

Figure 2 shows a screenshot of the service running in this model.

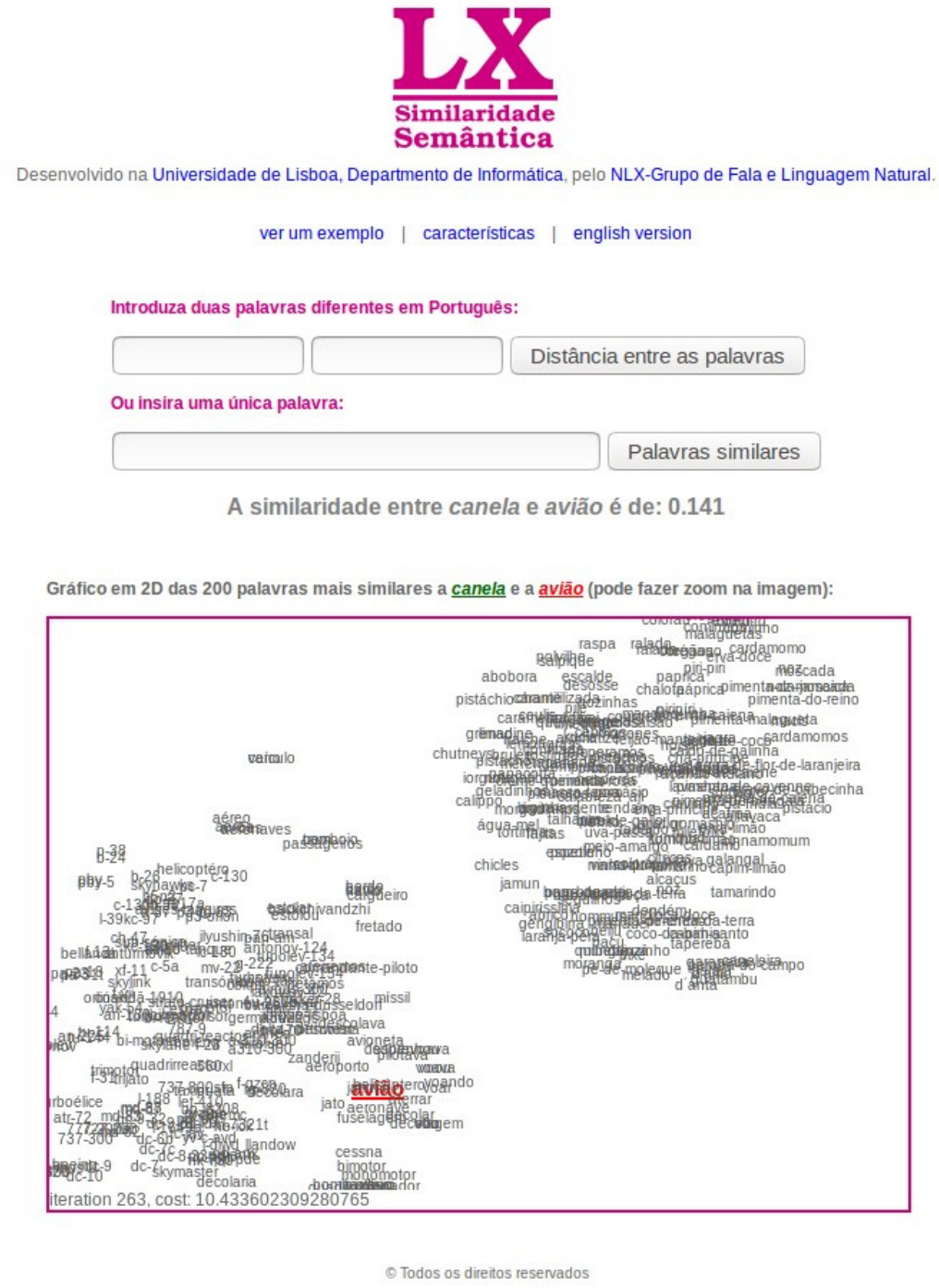

Figure 2: Screenshot of the Web demo after entering the words *canela* and *avião*

5.3 Architecture and technical details

The distributional semantic model supporting this online service was developed in line with the procedure described in Rodrigues *et al*., (2016). The model comprises over 1.3 million lowercased words with vectors of 500 components. When loaded, it requires nearly 6GB of RAM.

A Python script, through the genism package[4], is used to load and interact with the model. This script works as a server that wraps and exposes part of the gensim API through the XML-RPC protocol.

The online service is hosted in an Apache Web. Using PHP, it functions as a client to the server script that wraps the gensim API. The client-server architecture is used since loading the Word2Vec model for each user request would be too time-consuming.

The wordcloud image is created using the Python wordcloud package.[5] The t-SNE algorithm is implemented in Javascript, using the t-SNEJS library[6] which, in turn, relies heavily on the well-known jquery and d3 Javascript libraries.

6. Conclusion

In this paper we presented publicly available data sets that have been developed for the Portuguese language in order to evaluate distributional semantics models. With these data sets, it is possible to intrinsically evaluate a distributional semantic space by exploring its syntactic and semantic relations.

Since these vectors are important for a wide range of natural language processing tasks, this work represents a important contribuition to evaluate these models and enable the integration of distributional semantics in natural language processing tasks for the computational processing of Portuguese.

As an example of such tasks, we have also presented the LX Semantic Similarity online service, which showcases two options that make use of semantic similarity scores between words, as provided by a distributional semantics model.

[4] https://radimrehurek.com/gensim/

[5] https://github.com/amueller/word_cloud

[6] https://github.com/karpathy/tsnejs

Currently, we are making use of these new data sets to improve over the existing distributional semantic models for Portuguese described in Rodrigues et al., (2016).

Acknowledgments

This work was partially supported by the Portuguese Government's P2020 program under the grant 08/SI/2015/3279: ASSET-Intelligent Assistance for Everyone Everywhere.

References

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

Almuhareb, Abdulrahman and Massimo Poesio (2005) Concept learning and categorization from the web. In *Proceedings of the Cognitive Science Society*. Vol. 27. No. 27, pp. 103-108.

Baroni, Marco, Raffaela Bernardi and Roberto Zamparelli (2014) Frege in space: A program of compositional distributional semantics. *LiLT (Linguistic Issues in Language Technology)*, 9, pp. 241-346.

Baroni, Marco, Brian Murphy, Eduard Barbu and Massimo Poesio (2010) Strudel: A Corpus□ Based Semantic Model Based on Properties and Types. *Cognitive Science*, 34 (2), pp. 222-254.

Collobert, Ronan and Jason Weston (2008) A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*. Association for Computational Linguistics, pp. 160-167.

Finkelstein, Lev, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin (2002) Placing Search in Context: The Concept Revisited. In *ACM Transactions on Information Systems,* 20(1), pp. 116-131.

Firth, John Rupert (1957) A synopsis of linguistic theory 1930-1955. In *Studies in Linguistic Analysis*. Philological Society.

Harris, Zellig S. (1954) Distributional structure. *Word*, *10*(2-3), pp. 146-162.

Hill, Felix, Roi Reichart and Anna Korhonen (2015) Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, Vol. 41, No. 4, pp. 665-695.

Li, Jiwei and Dan Jurafsky (2015) Do multi-sense embeddings improve natural language understanding?. *arXiv preprint arXiv:1506.01070*.

Luong, Thang, Richard Socher and Christopher D. Manning (2013) Better Word Representations with Recursive Neural Networks for Morphology. In *CoNLL,* pp. 104-113.

McRae, Ken, George S. Cree, Mark S. Seidenberg and Chris McNorgan (2005) Semantic feature production norms for a large set of living and nonliving things. *Behavioral Research Methods, Instruments, and Computers*, 37(4), pp. 547-559.

Mikolov, Tomas, Kai Chen, Greg Corrado and Jeffrey Dean (2013) Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, Tomas, Jiri Kopecký, Lukas Burget, Ondrej Glembek and Jan Černocký (2009) Neural network based language models for highly inflective languages. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, IEEE, pp. 4725-4728.

Nelson, Douglas, Cathy L. McEvoy, and Thomas A. Schreiber (2004) The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers,* 36(3), pp. 402-407.

Rodrigues, João, António Branco, Steven Neale and João Silva (2016) LX-DSemVectors: Distributional Semantics Models for the Portuguese Language, *Lecture Notes in Artificial Intelligence*, 9727, Berlin: Springer, pp. 259-270.

van der Maaten, Laurens and Geoffrey Hinton (2008) Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research,* 9, pp. 2579-2605.

Miller, George A. (1995) WordNet: a lexical database for English. *Communications of the ACM* 38(11), pp.39-41.