LX-SemanticSimilarity*

João Silva[1], Marcos Garcia[2], João Rodrigues[1], and António Branco[1]

[1] University of Lisbon
NLX—Natural Language and Speech Group, Department of Informatics
Faculdade de Ciências, Campo Grande, 1749-016 Lisboa, Portugal
{jsilva,joao.rodrigues,antonio.branco}@di.fc.ul.pt
[2] University of Coruña, Faculty of Philology
marcos.garcia.gonzalez@udc.gal

Abstract. We present the LX-SemanticSimilarity web service and the respective demo, offered as an online service for human users. The web service provides an API to common operations over the LX-DSemVectors word embeddings for Portuguese without requiring the embeddings to be loaded locally.

Keywords: Web service · Online service · Distributional semantics · Word embeddings · Portuguese.

1 Introduction

Distributional semantic models, also known as word embeddings, represent the meaning of an expression as a high-dimension vector of real numbers. This vectorial representation of meaning allows, among other possibilities, to reify semantic similarity in terms of distance in a vector space. Having a way to quantitatively measure semantic similarity has opened up many avenues of research that explore how the integration of distributional features can improve a variety of natural language processing tasks, such as determining similarity between words [4], formal semantics [1], sentiment analysis [2], etc.

High-quality embeddings are hard to obtain due to the amount of data and computational effort required. LX-DSemVectors [7] are publicly available word embeddings for Portuguese and their existence helps in this regard, though they still require a great deal of RAM to operate and some technical skills, which may pose problems for some researchers, including from the Digital Humanities. In this paper, we present the LX-SemanticSimilarity web service, which provides access to the LX-DSemVectors through an API with several operations commonly used on such semantic representations.

* The research presented here was partly supported by the ANI/3279/2016 grant, by the Infrastructure for the Science and Technology of the Portuguese Language (PORTULAN / CLARIN), and by a *Juan de la Cierva* grant (IJCI-2016-29598).

2 LX-DSemVectors embbedings and LX-LR4DistSemEval evaluation datasets

LX-DSemVectors [7] are the first publicly available word embeddings for Portuguese. Trained over a corpus of 1.7 billion words, these embeddings were evaluated over the LX-4WAnalogies dataset [7], a translation of the *de facto* standard English dataset for analogies [4], and were found to have a performance at the level of the state-of-the-art.

LX-LR4DistSemEval [5] is a collection of datasets adapted via translation from various English gold standard datasets for different mainstream evaluation tasks for embeddings, namely the analogy task, the conceptual categorization task and the semantic similarity task. These datasets provide a standard way to intrinsically evaluate and compare distributional semantic models for Portuguese.

3 LX-SemanticSimilarity

The embeddings in LX-DSemVectors require nearly 6 GB of RAM when loaded, making them unfeasible to use on many desktop computers. We have found that loading them on a server and accessing them through a web service help to neatly circumvent this issue.

3.1 Web service

The LX-SemanticSimilarity web service exposes an API with operations commonly used on word embeddings, namely getting the (cosine) similarity between two words, and also between two sets of words; finding the top-n most similar words, allowing to specify words that contribute positively and words that contribute negatively; and getting the n words closest to a given word.

The server works as a XML-RPC wrapper around the gensim [6] library. Having a standard protocol like XML-RPC makes it easy to use any of a variety of programming languages on the client side, as seen in the following example in Python that queries the service to get the similarity between the words “batata” (potato) and “banana”:

```
import xmlrpc.client
lxsemsim = xmlrpc.client.ServerProxy(url)
result = lxsemsim.similarity("batata", "banana")
print(result)
```

3.2 Online service and demo

The LX-SemanticSimilarity online service/demo (http://lxsemsimil.di.fc.ul.pt/) is built on top of the web service and showcases some simple examples of possible applications of embeddings. The users are presented with two modes of

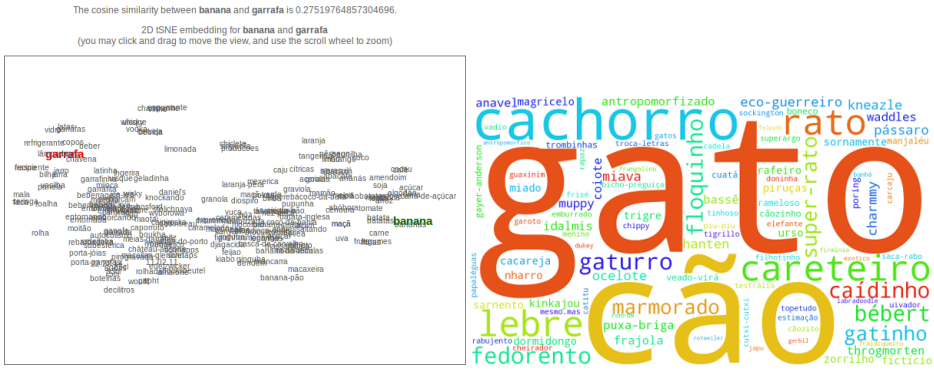

Fig. 1. Examples of output by LX-SemanticSimilarity online service and demo

operation: They can either (i) provide two words to see their distance and an interactive visualization of their surrounding vector-space; or (ii) provide a single word to see a list of the most similar words to it, in a tabular format and as a word cloud. The outputs of these two modes are exemplified in Figure 1.

The first mode is supported by the t-SNEJS JavaScript library,[3] an implementation of the t-SNE [3] dimensionality reduction technique; while the word-cloud image is generated by resorting to the wordcloud[4] Python package.

References

1. Baroni, M., Bernardi, R., Zamparelli, R.: Frege in space: A program for compositional distributional semantics. Linguistic Issues in Language Technology **9**, 241–346 (2014)
2. Li, J., Jurafsky, D.: Do multi-sense embeddings improve natural language understanding? arXiv preprint arXiv:1506.01070 (2015)
3. van der Maaten, L., Hinton, G.: Visualizing high-dimensional data using t-SNE. Journal of Machine Learning Research **9**, 2579–2605 (2008)
4. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
5. Querido, A., de Carvalho, R., Rodrigues, J., Garcia, M., Correia, C., Rendeiro, N., Pereira, R.V., Campos, M., Silva, J., Branco, A.: LX-LR4DistSemEval: A collection of language resources for the evaluation of distributional semantic models of Portuguese. Revista da Associação Portuguesa de Linguística **3** (2017)
6. Řehůřek, R., Sojka, P.: Software framework for topic modelling with large corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. pp. 45–50 (2010)
7. Rodrigues, J., Branco, A., Neale, S., Silva, J.: LX-DSemVectors: Distributional semantics models for the Portuguese language. In: Proceedings of the 12th International Conference on the Computational Processing of Portuguese (PROPOR'16). pp. 259–270 (2016)

[3] https://github.com/karpathy/tsnejs
[4] https://github.com/amueller/word_cloud