SemantiGal: An online visualizer of vector representations for Galician[⋆]

Marcos Garcia, Ignacio Rodríguez and Pablo Gamallo

Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS)
Universidade de Santiago de Compostela
{marcos.garcia.gonzalez,i.rodriguez.perez,pablo.gamallo}@usc.gal

Abstract. This paper presents SemantiGal, an online tool to explore vector models for Galician. It includes a masked language modeling demonstrator using two monolingual BERT-based models, and three demos using static word embeddings: similarity between words, algebraic operations using word vectors, and searching for similar words in both monolingual and cross-lingual scenarios.

Keywords: Galician · Neural Language Models · Word Embeddings

1 Introduction

Low dimensional vector representations of words have gained popularity in several research areas, including Computational Linguistics and Cognitive Sciences, and they have become crucial for developing downstream NLP applications, such as PoS-taggers or dependency parsers [11]. There are a variety of models for languages with large textual resources, but word vectors of linguistic varieties with fewer data available are scarce. Furthermore, there is a lack of online tools and visualizers of vector space models for these low-resource languages, which are useful for several domains, such as language learning or lexicography [2, 12].

This paper presents SemantiGal, an online service that includes demonstrators that exemplify some of the possibilities of using vector models for Galician (in its ILG/RAG spelling).[1] It includes both static vectors, which represent words at type-level, and contextualized models, which generate different (token-level) representations for a word depending on its context. We also include cross-lingual models between Galician and other linguistic varieties for the former.

[⋆] This research was funded by a *Ramón y Cajal* grant (RYC2019-028473-I), by the project "Nós: Galician in the society and economy of artificial intelligence" (agreement between Xunta de Galicia and University of Santiago de Compostela), and by grants ED431G2019/04 and ED431F 2021/01 of the Galician Government.

[1] `https://tec.citius.usc.es/demos-lingua/`

2 Models

All Galician models included in SemantiGal were trained using a Galician corpus of about 600M tokens (details on [5]). 300 dimensions static word embeddings were trained using the Gensim implementation [10] of fastText [3], while the contextualized models have a Transformer-based BERT architecture [4]. We trained 'small' (6 layers) and 'base' (12 layers) models using the HuggingFace library.[2] All models can be freely downloaded from SemantiGal and from the HuggingFace hub.[3] Additionally, we trained fastText models for Portuguese, English, and Spanish (using Wikipedia), which were then mapped to the same vector space as the Galician static model with VecMap [1].

Evaluation. Our static models obtained better results than the official fastText vectors for Galician [7], and than several models trained with other algorithms (e.g., word2vec, GloVe). They were evaluated on several tasks: On the one hand, on word analogy and concept categorization,[4] and using an in-context homonymy and synonymy identification dataset [5], on the other.[5] Regarding the BERT models, both the 'small' and the 'base' ones achieved better performance on the former dataset than the other available models (monolingual and multilingual). The unsupervised approach used to train the cross-lingual models obtained competitive performance (the best values in various settings) in several language pairs at the *Translation Inference Across Dictionaries* shared task [6].

3 SemantiGal

SemantiGal includes the following demonstrators aimed at showing some possibilities of word vector representations:

Context prediction: it uses our two BERT models (together with the Google's multilingual mBERT) to predict the most likely words in a given context. For instance, given the context "O peixe e a carne * no frigorífico" ('The fish and the meat * in the fridge', where * is the masked position), the model gives the highest probability to the verb form "están" ('are').

Similar words: with static models, this tool shows the forms whose vectors are close to the input one. It has two types of searches: monolingual, using the Galician model; and cross-lingual, which allows for searching similar words between Galician, and Portuguese, Spanish, or English. As an example of the cross-lingual search, the most similar words to the Galician "xanela" ('window') in Spanish are "ventana", "ventanal", or "ventanuco" (which refer to windows of different sizes).

Similarity between words: it computes the similarity between two input words using the Galician fastText model, and shows a 2d plot (where dimensionality reduction is done with the JavaScript implementation of t-SNE [8]).[6]

[2] `https://github.com/huggingface/transformers`

[3] `https://huggingface.co/marcosgg`

[4] `https://github.com/marcospln/vector_models_evaluation`

[5] `https://github.com/marcospln/homonymy_acl21`

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

Vector operations: it allows to perform algebraic operations with word vectors, such as the semantic and morphological analogies proposed by Mikolov *et al.* [9]. Given an operation (e.g., 'king -man +woman'), the words whose vectors are closest to the vector resulting from that operation are shown ('queen', in the given example). For Galician, we include some examples of morphological (e.g., 'facían -facer +deseñar≈deseñaban', where the morphological features of the verb "facían" are kept with the verb "deseñar"), semantic ('maior -grande +pequeno≈menor', where the comparative relations are identified), and word knowledge ('Lisboa -Portugal +Xapón≈Toquio', retaining the capital-country relation) analogies to show the capabilities of the model.

References

1. Artetxe, M., Labaka, G., Agirre, E.: A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. pp. 789–798 (2018)
2. Barzegar, S., Sales, J.E., Freitas, A., Handschuh, S., Davis, B.: DINFRA: A One Stop Shop for Computing Multilingual Semantic Relatedness. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1027–1028 (2015)
3. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics **5**, 135–146 (2017)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the ACL. pp. 4171–4186 (2019)
5. Garcia, M.: Exploring the representation of word meanings in context: A case study on homonymy and synonymy. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL-IJCNLP). pp. 3625–3640 (2021)
6. Garcia, M., García-Salido, M., Alonso, M.A.: Exploring cross-lingual word embeddings for the inference of bilingual dictionaries. In: Proceedings of TIAD-2019 Shared Task – Translation Inference Across Dictionaries. pp. 32–41 (2019)
7. Grave, E., Bojanowski, P., Gupta, P., Joulin, A., Mikolov, T.: Learning word vectors for 157 languages. In: Proceedings of LREC 2018 (2018)
8. van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research **9**(86), 2579–2605 (2008)
9. Mikolov, T., Yih, W.t., Zweig, G.: Linguistic regularities in continuous space word representations. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics. pp. 746–751 (2013)
10. 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)
11. Schnabel, T., Labutov, I., Mimno, D., Joachims, T.: Evaluation methods for unsupervised word embeddings. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 298–307 (2015)
12. Silva, J., Garcia, M., Rodrigues, J., Branco, A.: LX-SemanticSimilarity. In: 13th International Conference on the Computational Processing of the Portuguese Language (PROPOR 2018). Demo papers (2018)