An exploration of the semantic knowledge in vector models: polysemy, synonymy and idiomaticity

Exploración del conocimiento semántico en modelos vectoriales: polisemia, sinonimia e idiomaticidad

Marcos Garcia[1], Pablo Gamallo[1], Martín Pereira-Fariña[2] and Iria de-Dios-Flores[1]

[1]*Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela*
[2]*Departamento de Filosofía e Antropoloxía, Universidade de Santiago de Compostela*

Abstract

In this paper, we present the project *An exploration of the semantic knowledge in vector models: polysemy, synonymy and idiomaticity*, funded by the Xunta de Galicia within the program "Consolidación e estruturación de unidades de investigación competitivas e outras accións de fomento: Proxectos de Excelencia", with a duration of 5 years (2021-2026). The main objective of the project is the analysis of the most recent language models regarding the representation of several aspects of lexical semantics: polysemy and homonymy, synonymy and idiomaticity. The languages in which we are working are Galician-Portuguese (in its Galician and Portuguese varieties, fundamentally), Spanish and English.

Keywords

lexical semantics, distributional semantics, language models.

1. Introduction and objectives

The use of architectures based on artificial neural networks has become the most dominant approach to natural language processing (NLP) in recent years [1], producing significantly better results in numerous areas than supervised models designed by selecting individual features of the target tasks [2]. This paradigm shift has promoted the popularization of vector models inspired by the distributional hypothesis [3, 4], which until then were mainly used in research in cognitive science and computational linguistics [5, 6, 7]. In this field, the implementation of computationally more efficient architectures, with drastic reductions in dimensionality [8], has sparked great interest in distributional semantics studies, boosted also by the findings about the various linguistic regularities encoded by these models [9]. This area, previously dominated by linguistically informed and more interpretable methodologies (e.g., using vectors built through syntactic dependencies [10]), has become one of the most productive in NLP research [11].

In this regard, the emergence of deep learning techniques using multilayer deep neural networks with millions of hyperparameters (which require large computational infrastructures) has led to the proliferation of language models that perform NLP tasks more accurately. Among various others, we can highlight the public models ELMo (Embeddings from Language Models [12]), or the different variants of BERT (Bidirectional Encoder Representations from Transformers [13]).

The project presented in this paper fits into this new line of research and focuses on the analysis of the ability of these models to solve various types of lexical ambiguity:[1]

1. Polysemy and homonymy, i.e., a single orthographic form that has different meanings (or senses) depending on the context. For example, *school* as a building, as an organization, or as a group of people (polysemy), or *bank* as a financial institution, or as a sloping raised land (homonymy).
2. Synonymy, i.e., different words expressing the same meaning in certain contexts (e.g., *coach* or *bus* to refer to a long motor vehicle).
3. Idiomaticity, i.e., multiword expressions (MWEs) whose meaning does not correspond to the one of its constituent elements (e.g., *glass ceiling* as a social barrier for women).

[1]We broadly follow [14] for the definition of the phenomena mentioned here.

SEPLN-PD 2022. Annual Conference of the Spanish Association for Natural Language Processing 2022: Projects and Demonstrations, September 21-23, 2022, A Coruña, Spain
marcos.garcia.gonzalez@usc.gal (M. Garcia); pablo.gamallo@usc.gal (P. Gamallo); martin.pereira@usc.gal (M. Pereira-Fariña); iria.dedios@usc.gal (I. de-Dios-Flores)
0000-0002-6557-0210 (M. Garcia); 0000-0002-5819-2469 (P. Gamallo); 0000-0002-1982-2472 (M. Pereira-Fariña); 0000-0002-5941-1707 (I. de-Dios-Flores)

Taking the above into account, our research aims to fill a particularly important gap in the evaluation of these computational models by investigating the presence of various types of knowledge related to lexical semantics in several languages. Thus, the main goal of the project is to explore the most recent language models concerning the representation of polysemy and homonymy, synonymy and semantic compositionality, as well as to compare them with more interpretable distributional and compositional methods.

The results of the present project will be useful, on the one hand, to advance the understanding of semantic information encoded both in static distributional representations and in large language models trained with deep neural networks. In addition, and although the project is mainly focused on the exploration of models, both the datasets and the results of manual annotation will be an important contribution regarding the semantic interpretation of polysemy and homonymy, synonymy and idiomaticity by native speakers of various languages.

2. Methodology and work plan

To develop this project, we will use the following methodology and instrumental techniques, which in general correspond to the state-of-the-art research in NLP and computational linguistics.

Regarding the experimental design and the data collection, we will use standard methodologies from studies in semantics [14] and in psycholinguistics [15, 16], aimed at generating controlled stimuli. Likewise, to collect annotations from human informants, we will use crowdsourcing methods which will allow us to obtain data from native speakers quickly and efficiently, with quality control of the annotations [17].

Regarding the computational models, those based on Transformer architectures will be implemented using the *transformers* library, which includes the latest models based on deep learning. We will eventually use other open source libraries that may incorporate additional models. To train and run static embeddings, we will use *gensim*[2] and the official tools released by the authors of other distributional methods based on interpretable syntactic dependencies (e.g., [18]).

Finally, to compare the representations of the computational models with the values obtained from the human annotations, we will use three methods:

1. Precision scores, in evaluations with discrete values (e.g. homonymy or synonymy, and in the results of linear classifiers).
2. Correlation values, in graded evaluations (polysemy or idiomaticity).
3. Representation Similarity Analysis, to see if the models predict relative differences between examples of the same type (e.g., a word or MWE with the same meaning in different contexts) in a similar way to humans.

It should be noted that these methods have already been used in previous works, which we briefly mention below.

2.1. First results

Although we are at an early stage, we already have some published results, both from previous research directly related to this proposal and from work carried out since the beginning of the project. Thus, we have already presented various datasets with semantic idiomaticity annotation at token and type levels in English and Portuguese, and used them to evaluate several language models [19, 20]. In addition, we have created a new dataset in Galician-Portuguese, English and Spanish that includes examples of homonymy and synonymy in context, also used to compare various contextualization models and strategies [21].

More recently, we have compared Transformers models and distributional strategies based on syntactic dependencies in semantic compositionality tasks [18, 22]. Finally, we have participated in the co-organization of the task *Multilingual Idiomaticity Detection and Sentence Embedding* (SemEval 2022), in which we have presented new resources with annotation of semantic idiomaticity in context in Galician-Portuguese and English [23].

3. Work team

The project presented in this paper is carried out at the Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS) of the Universidade de Santiago de Compostela, and belongs to its scientific program in Natural Language Technologies. In this sense, members of the center collaborate on different tasks of our work plan, that are part of their respective areas of expertise.

Besides the principal investigator, the project has research and work teams formed by three PhDs with specializations in Computational Linguistics, Psycholinguistics, Logic and Computer Science. In collaboration with a pre-doctoral researcher and

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

technical staff that will be hired with the project funds, these teams actively participate in the different stages of the project. Finally, we also rely on the collaboration of researchers from other universities, both Galician and international, with whom we have already participated in joint initiatives and projects with similar themes to the one presented in this paper.

Acknowledgments

Project funded by the Galician Government (*Consolidación e estruturación de unidades de investigación competitivas e outras accións de fomento: Proxectos de Excelencia*, ED431F 2021/01) and by a *Ramón y Cajal* grant (RYC2019-028473-I).

References

[1] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, P. Kuksa, Natural language processing (almost) from scratch, Journal of Machine Learning Research 12 (2011) 2493–2537.

[2] T. Schnabel, I. Labutov, D. Mimno, T. Joachims, Evaluation methods for unsupervised word embeddings, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Lisbon, Portugal, 2015, pp. 298–307. URL: https://aclanthology.org/D15-1036. doi:10.18653/v1/D15-1036.

[3] Z. S. Harris, Distributional structure, Word 10 (1954) 146–162.

[4] J. R. Firth, A synopsis of linguistic theory 1930-1955, Studies in Linguistic Analysis (1957) 1–32. Reprinted in F.R. Palmer (Ed.), Selected Papers of J.R. Firth 1952–1959, London: Longman (1968).

[5] G. A. Miller, Empirical methods in the study of semantics, in: D. D. Steinberg, L. A. Jakobovits (Eds.), Semantics: An Interdisciplinary Reader in Philosophy, Linguistics and Psychology, 1971, pp. 569–585.

[6] T. K. Landauer, S. T. Dumais, A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge, Psychological Review 104 (1997) 211.

[7] J. Mitchell, M. Lapata, Composition in distributional models of semantics, Cognitive science 34 (2010) 1388–1429.

[8] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, in: Workshop Proceedings of the International Conference on Learning Representations, 2013.

[9] T. Mikolov, W.-t. Yih, G. Zweig, Linguistic regularities in continuous space word representations, in: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Atlanta, Georgia, 2013, pp. 746–751. URL: https://aclanthology.org/N13-1090.

[10] S. Padó, M. Lapata, Dependency-based construction of semantic space models, Computational Linguistics 33 (2007) 161–199. URL: https://aclanthology.org/J07-2002. doi:10.1162/coli.2007.33.2.161.

[11] G. Boleda, Distributional semantics and linguistic theory, Annual Review of Linguistics 6 (2020) 213–234.

[12] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 2227–2237. URL: https://aclanthology.org/N18-1202. doi:10.18653/v1/N18-1202.

[13] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[14] D. A. Cruse, Lexical semantics, Cambridge University Press, 1986.

[15] R. L. Goldstone, Influences of categorization on perceptual discrimination., Journal of Experimental Psychology: General 123 (1994) 178.

[16] R. Richie, B. White, S. Bhatia, M. C. Hout, The spatial arrangement method of measuring similarity can capture high-dimensional semantic structures, Behavior Research Methods 52 (2020) 1906–1928.

[17] R. Munro, S. Bethard, V. Kuperman, V. T. Lai, R. Melnick, C. Potts, T. Schnoebelen, H. Tily, Crowdsourcing and language studies: the new generation of linguistic data, in: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, Association for Computational Linguistics, Los Angeles, 2010, pp. 122–130. URL: https://aclanthology.org/W10-0719.

[18] P. Gamallo, M. de Prada Corral, M. Garcia, Comparing Dependency-based Compositional Models with Contextualized Word Embeddings, in: Proceedings of the 13th International Conference on Agents and Artificial Intelligence (ICAART 2021), Volume 2, 2021, pp. 1258–1265.

[19] M. Garcia, T. Kramer Vieira, C. Scarton, M. Idiart, A. Villavicencio, Assessing the representations of idiomaticity in vector models with a noun compound dataset labeled at type and token levels, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL-IJCNLP), ACL, 2021, pp. 2730–2741. URL: https://aclanthology.org/2021.acl-long.212. doi:10.18653/v1/2021.acl-long.212.

[20] M. Garcia, T. Kramer Vieira, C. Scarton, M. Idiart, A. Villavicencio, Probing for idiomaticity in vector space models, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, Online, 2021, pp. 3551–3564. URL: https://aclanthology.org/2021.eacl-main.310. doi:10.18653/v1/2021.eacl-main.310.

[21] M. Garcia, 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 and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 3625–3640. URL: https://aclanthology.org/2021.acl-long.281. doi:10.18653/v1/2021.acl-long.281.

[22] P. Gamallo, M. Garcia, I. de-Dios-Flores, Evaluating Contextualized Vectors from Large Language Models and Compositional Strategies, Procesamiento del Lenguaje Natural 69 (2022).

[23] H. Tayyar Madabushi, E. Gow-Smith, M. Garcia, C. Scarton, M. Idiart, A. Villavicencio, SemEval-2022 task 2: Multilingual idiomaticity detection and sentence embedding, in: Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), Association for Computational Linguistics, Seattle, United States, 2022, pp. 107–121. URL: https://aclanthology.org/2022.semeval-1.13.