Using Diachronic Distributional Models to Study Semantic Variation in Spanish Idioms

Research on word variation over time is fundamental for the development of language models, as well as for the understanding of language change. Idioms are the result of language change, thus are subjected to a high degree of variation (Jimenez et al, 2018). Because technology progresses in such a fast pace, multiword expressions, such as idioms, appear, disappear and change in meaning continuously. At the same time, technology allows us to track semantic change employing word embeddings as a diachronic tool (Rosenfeld & Erk (2018), Hamilton et al. (2018)). While word embeddings show promise in the study of language variation over time, diachronic distributional models in languages such as Spanish need to be investigated further. In this paper, I study diachronic variation in Spanish idioms utilizing word embedding models from large corpora.

Idioms are a type of multiword expression (MWE) whose meaning is non-compositional. Thus, the meaning of *"ya salió el peine"* (to uncover the truth) cannot be derived by breaking it down and examining the meaning of its constituent parts (Peng and Feldman, 2016). It is well established within the field of linguistics that the properties attributed to idiomatic expressions represent great challenges for several linguistic applications (e.g. idioms present different degrees of statistical idiomacity; idioms can be syntactically ill-formed or semantically ambiguous without context). Despite that, the development of word embedding models, as well as the rise of big data facilitate the study of word variation over time. I describe a method using word embeddings for tracking variation in idiomatic expressions based on two hypotheses: 1) the proposed diachronic distributional model can detect semantic variation by comparing vectors for a given idiom across different time points, 2) an idiom should be treated as a single token the models Idiom Principle (see Peng & Feldman, 2016).

Diachronic distributional models use words embedded in vector spaces correspondent to their co-ocurrence relationships where the vector changes over time (Hamilton et. al, 2018). These vectors are then compared across time utilizing metrics to quantify semantic change. The method used to construct word embeddings is SkipGram with Negative Sampling (SGNS) (i.e. Word2vec). Then, the distance between words (semantic similarity) is calculated at different points in time using the cosine similarity (Turney and Pantel, 2010) to measure change overt time. A nearest neighbor analysis then is performed in order to understand in which way a word has changed (Rosenfeld & Erk, 2018). For example, with this method we can observe the idiom 'ser/parecer un bodrío' has changed its meaning over time. We observe in sentence (1) the meaning of *'bodrio'* is closer to the one described in Diccionario de Las Lenguas Española e Inglesa, printed in 1837: a mixture of things put together without any order. Differently, in sentence (2) *'bodrio'* refers to something very ugly, in bad taste or boring, which aligns with a more modern use of the expression (see García-Robles, 2012).

1. …**es bodrio** confusísimo del que pueden hacerse mil esplicaciones… [1]

   ...it is a very confusing mess of which a thousand explanations can be made...

2. No **es** más que un **bodrio** vergonzoso que demuestra que ni siquiera sabemos nada de ella. [2]

It's nothing more than an embarrassing creature that shows we don't even know about her.

References

Fernández, J., and Garrigós, H. 2020. *Changes in Meaning and Function: Studies in Historical Linguistics with a Focus on Spanish.* edited by Jaén. John Benjamins Publishing Company.

García-Robles., J. 2016. Diccionario de Modismos Mexicanos. Published by Editorial Porrua. México. Primera Edición.

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096*.

Jimenez, S., Dueñas, G., Gelbukh, A., Rodriguez-Diaz, C.A., Mancera, S. 2018. Automatic Detection of Regional Words for Pan-Hispanic Spanish on Twitter. In: Simar, G., Fermé, E., Gutiérrez Segura, F., Rodríguez Melquiades, J. (eds) Advances in Articial Intelligence - IBERAMIA 2018. IBERAMIA 2018. Lecture Notes in Computer Science (), vol 11238.

Peng, J., Feldman, A., Jazmati, H. 2015. Classifying idiomatic and literal expressions using vector space representations. In: RANLP, pp. 507–511

Peng, J., Feldman, A. 2016. Experiments in idiom recognition. In: COLING, pp. 2752– 2762.

Rosenfeld, A., & Erk, K. (2018). Deep neural models of semantic shift. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 474-484).

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, *37*, 141-188.

[1] Un Final para Rachel by Jesse Andrews, 2015. Penguin Random House Grupo Editorial

[2] Tratado del verdadero origen de la religión y sus principales épocas, en que se impugna La Obra de Dupuis. José de Jesus Muñoz. Vol I. 1828. p14