A corpus-based variational measure of the order of the words

Vector representations of words and texts extracted from the weights of neuronal networks trained with large corpus (neural embeddings) provide models where vectors represent semantic relationships (Mikolov et al. 2013) and grammatical dependencies (Hewitt & Manning, 2019). Although these semantic and grammatical patterns are effective, they do not have a theoretical justification, they are not yet well understood, and it is not possible to dissociate semantics and syntax.

Alternatively, we propose an information-theoretical measure of the order of the words for variational change. We use the method of Montemurro & Zanette (2011) to measure the amount of information in a large corpus due to the order of the words. The idea is to compare the sizes of the corpus compressed file (using bzip2) and a version of the same where the order of words within sentences is randomized. The difference between them approximates the amount of information due to the order of the words in the corpus. We extend this to compare two corpora A and B by intervening a sentence of each corpus at a time (Gouws & Søgaard, 2015). Next, 1) the compressed size of the combined corpus is measured; 2) also the compressed size of the combined corpus, but randomizing only the sentences from A; 3) then randomizing only those of B, 4) and finally randomizing all sentences. The final measure is obtained by combining these four measures with the Jaccard coefficient.

We tested this concept using 8 corpora of 15 million tweets each corresponding to Spain (2 corpora), Mexico (2 corpora), Colombia, Venezuela, Argentina, and Chile. The results have a good degree of agreement with the current linguistic knowledge associated with these variants of the Spanish language. We conclude that the proposed measure is a sound approach, conceptually simple, theoretically grounded, and independent of the language.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).

Hewitt, J., & Manning, C. D. (2019, June). A structural probe for finding syntax in word representations. 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) (pp. 4129-4138).

Montemurro, M. A., & Zanette, D. H. (2011). Universal entropy of word ordering across linguistic families. PLoS One, 6(5), e19875.

Gouws, S., & Søgaard, A. (2015). Simple task-specific bilingual word embeddings. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 1386-1390).