SWEAT-XS: A method to analyze implicit associations in word embeddings trained on small corpora

Keywords: Implicit Association Test, Word Embedding Association Test, Small Corpora

Extended Abstract

Language is a fundamental aspect of human interaction and is known to transmit some cultural, historical, or educational associations present in the speakers. The transmission of associations can be intentional or unintentional [1] and associations are known as *biases*. In 2016, Caliskan et al [5] demonstrated that some biases are present in Word Embeddings obtained by common word embedding algorithms, such as Word2Vec [2, 3] and GloVe [4]. The presence of biases can be measured with a metric introduced in the same article, named Word Embedding Association Test (WEAT) [5].

WEAT is a measure inspired by Implicit Association Tests (IAT). IATs are used by psychologists to measure implicit bias in individuals by measuring the difference in response times in the evaluation of associations between concepts, represented as *target concepts* (also known as *categories*) and *attribute concepts*. As an example, measures were taken to evaluate the associations of *flowers* and *insects*, respectively, with the concepts of *pleasant* and *unpleasant*. Caliskan et al. demonstrated how WEAT can be used to find some common biases and that the WEAT values are positively correlated with IAT measures.

SWEAT [6] is a measure proposed by Bianchi et al. to analyze a slightly different scenario w.r.t WEAT: instead of analyzing the associations between two target concepts and two attributes inside the same community (represented by the same corpus), SWEAT is used to analyze the association between the same target concept and two attributes inside different communities (represented by different corpus). The idea is to extract which concepts are represented by the same words but with different usage and consequently meaning in the two communities. SWEAT enables also for temporal analysis of the same community.

Both WEAT and SWEAT represent the target and attribute concepts using set of words, called wordsets. It is known that the selection of words while building wordsets is crucial to obtain evidences; moreover, it is also known that the occurrences of words are highly related with the quality of the vectors obtained by training Word2Vec, i.e., the more the occurrences of a word, the more its most common meaning is represented in the word embedding space. It follows that WEAT and SWEAT evidence extracted from a Word2Vec space built over a small corpus is difficult to interpret due to the expected low quality and reliability of embeddings.

The presented work is focused on the definition and development of SWEAT-XS, a version of SWEAT [6] that is agnostic to the size of the corpora involved in the analysis of implicit associations. To do this, we introduced an additional corpus considered neutral from which the embeddings of the words present in the axis wordsets are extracted.

We validated our approach by replicating some of the experiments proposed in the original SWEAT paper [6], confirming all the evidences on large corpora. On reduced version of the same corpora, we demonstrated how SWEAT-XS is able to obtain the same evidences obtained on large corpora, with a significant p-value, evidence that otherwise would not be obtainable if SWEAT were used on the same reduced corpora.

References

[1] Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. "Measuring individual differences in implicit cognition: the implicit association test." In: *Journal of personality and social psychology* 74.6 (1998), p. 1464.

[2] Tomas Mikolov et al. "Distributed representations of words and phrases and their compositionality". In: *Advances in neural information processing systems* 26 (2013).

[3] Tomas Mikolov et al. "Efficient estimation of word representations in vector space". In: *arXiv preprint arXiv:1301.3781* (2013).

[4] Jeffrey Pennington, Richard Socher, and Christopher D Manning. "Glove: Global vectors for word representation". In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, pp. 1532–1543.

[5] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. "Semantics derived automatically from language corpora contain human-like biases". In: *Science* 356.6334 (2017), pp. 183–186.

[6] Federico Bianchi et al. "SWEAT: Scoring Polarization of Topics across Different Corpora". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 10065–10072. DOI: `10.18653/v1/2021.emnlp-main.788`. URL: `https://aclanthology.org/2021.emnlp-main.788`.