9^{th} International Conference on Computational Social Science IC^{2}S^2
July 17-20, 2023, Copenhagen, Denmark

Sentence-Level Explanations of Word Embeddings Associations

Keywords: Gender bias, Word2Vec, GloVe, Influence functions, News corpora

Extended Abstract

Humans interact through language and tend to intentionally or unintentionally introduce associations between concepts when they communicate [1]: these associations may be referred to as *biases* when they are stronger than what would be normally expected. Word embedding algorithms, such as Word2Vec [1] and GloVe [2], are sensitive to language associations and thus inherit (and can even amplify) biases from the training corpora. One of the most common approaches used to measure implicit associations learned by embedding models is the Word Embedding Association Test (WEAT) effect size [2]: its definition on the word sets related to the concepts of *science*, *art*, *man* and *woman* will be the reference for this work.

Given an association learned by a word embedding model M trained on a corpus C (i.e. a set of documents $d_1, d_2, ..., d_n$) and thus measured in the embedding space $E(M,C)$, an interesting yet less explored research task is to verify the existence of a group of documents $S \subset C$ of arbitrary size k that, if removed, alleviates the association measured in the $E(M,C)$. Recent research has explored the use of influence functions [3] to identify S [4, 5] but, to the best of our knowledge, there are no approaches that examine how the choice of p for identifying S affects its composition. The purpose of the proposed work is to evaluate the use of some influence functions-based procedures $p_1, p_2, ..., p_m$, based on potentially different embedding models $M_1, M_2, ..., M_m$ for the identification of S given an association measured on C and to explore possible links between the definition of S and the intrinsic features of the procedures under investigation.
The corpus C of reference is a subset of Wikipedia dumps [3] (WIKI) and the explored procedures correspond to the approaches proposed in [5] and in [4], which will be referred to as CLIF and BRUNET respectively. Both CLIF and BRUNET, applied to find the set S which alleviates the WEAT effect size measured on C, are designed to express the influence of a document being aware of the optimization methods of the embedding models M_1, M_2 of reference (Word2Vec for CLIF and GloVe for BRUNET). Taking into account the stochasticity of word embeddings, we consider multiple model instances $m_1, m_2, \ldots, m_l$ by varying the initialization seed: in our experiments, we set $l = 10$. Each model instance allows us to explore the selection of sentences resulting from the application of p_j to $M_j, j = 1, ..., m$.

The first step we follow in the experiments is the evaluation of the statistical significance of the WEAT effect size measured on C, on which the two models M_1, M_2 of reference are trained: both Word2Vec and GloVe show a significant bias (as defined and verified by [2]) (see Figure 1). The next step is related to the definition of an influence-based ranking of documents in C: generally speaking, an influence function-based procedure p takes as input the reference model M and the corpus C and returns a ranking of the documents $d_1, d_2, \ldots, d_n$ in C according to their impact on the computed WEAT effect size (note that the word sets used in WEAT are also used

[1] `https://arxiv.org/abs/1301.3781`
[2] `https://aclanthology.org/D14-1162.pdf`
[3] `https://github.com/GermanT5/wikipedia2corpus`

in the computation of influence functions). Following the claims of the original papers, top positions of the ranking identify those documents that increase the observed association if part of C during training (*strenghtening sentences*) while bottom positions refer to those documents that invert the direction of association if part of the training set (*weakening sentences*). Elaborating on this, we can say that S exists and is defined by the top-k documents in the ranking. Without loss of generality, we focus on sentences (i.e. we consider $d_i, i = 1, ..., n$ to be a sentence) and we develop an average ranking for both CLIF-based and BRUNET-based experiments by averaging the scores of sentences' influence on the WEAT effect size computed for each model instance, aiming at reducing the stochasticity resulting from the use of embedding models. We select and remove from the corpus C a fraction k of the top-positioned sentences in the average ranking; we then train the word embedding models M_1, M_2 on this reduced version of the corpus; lastly, we compare the WEAT effect size computed on the embeddings obtained from the original and the reduced corpora (see Figure 2) and we draw some observations. In particular:

- After the removal of the top-k association strengthening sentences, WEAT effect size decreases for both CLIF-based and BRUNET-based experiments.
- As a control experiment, we remove k random sentences and compute the WEAT effect size. We verify that the values of WEAT effect size lie in the original distribution.

We then consider the average sentences ranking and evaluate the number of tokens in the sentences to find a relationship with positions in the ranking. In Figure 3, sentences are grouped by contiguous ranking positions (1-100, 101-200, and so on). The observed relationship between the average number of tokens per sentence and positions in the ranking depends on the embedding model M and consequently on the procedure p used to build the ranking: using the CLIF-based approach, the most impacting sentences are shorter that non-impacting ones, using the BRUNET-based approach the most impacting sentences are longer than non-impacting ones. This fact might be related to the intrinsically different training processes of Word2Vec and GloVe.

The study proposes, to the best of our knowledge, the first comparison of approaches to identify rankings of influential documents that can alleviate an implicit association in textual data, and future research will focus on evaluating the rankings through human-based metrics.

References

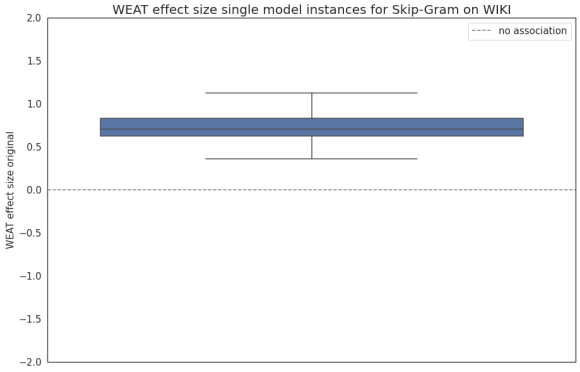
[1] Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. "Measuring individual differences in implicit cognition: the implicit association test." In: *Journal of personality and social psychology* 74.6 (1998), p. 1464.

[2] 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.

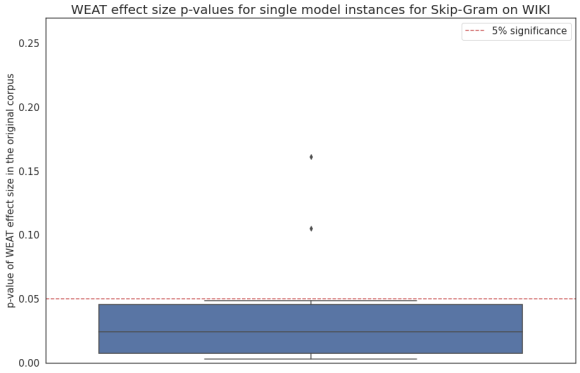
[3] Pang Wei Koh and Percy Liang. "Understanding black-box predictions via influence functions". In: *International conference on machine learning*. PMLR. 2017, pp. 1885–1894.

[4] Marc-Etienne Brunet et al. "Understanding the origins of bias in word embeddings". In: *International conference on machine learning*. PMLR. 2019, pp. 803–811.

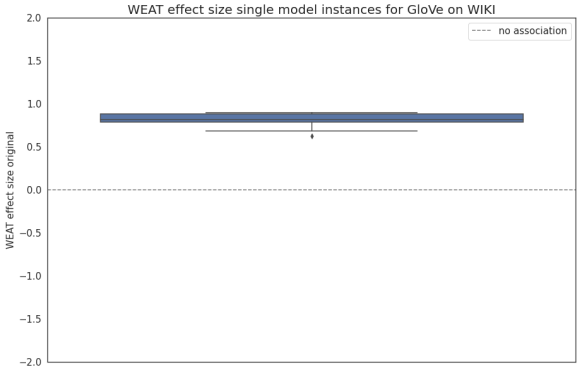
[5] Andrew Silva, Rohit Chopra, and Matthew Gombolay. "Cross-Loss Influence Functions to Explain Deep Network Representations". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2022, pp. 1–17.



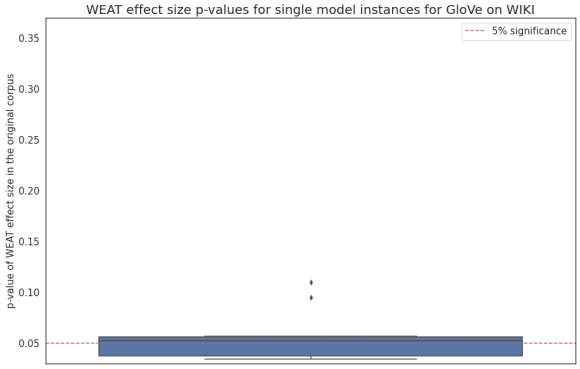
(a) WEAT effect size measure Word2Vec



(b) WEAT effect size p-value Word2Vec

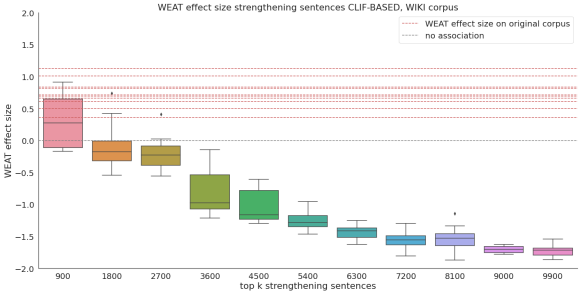


(c) WEAT effect size measure GloVe

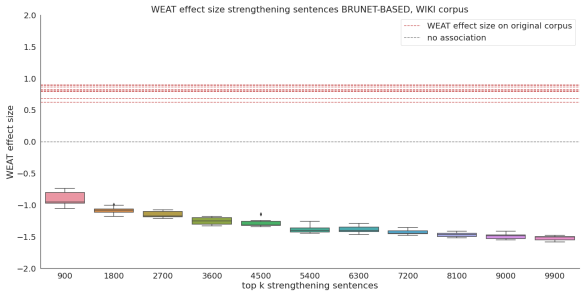


(d) WEAT effect size p-value GloVe

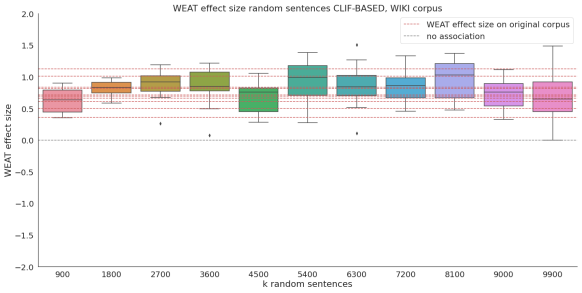
Figure 1: WEAT effect sizes and p-values for the 10 Word2Vec and 10 GloVe model instances (built by considering different initialization seeds), trained on the Wikipedia corpus.



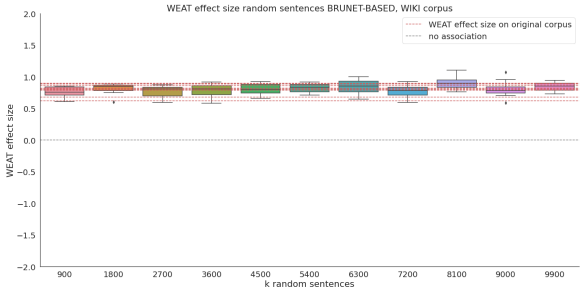
(a) strengthening sentences CLIF-based experiment



(b) strengthening sentences BRUNET-based experiment



(c) random sentences CLIF-based experiment



(d) random sentences BRUNET-based experiment

Figure 2: WEAT effect size via model training (CLIF-based and BRUNET-based experiments) on reduced corpora, grouped by number of sentences that are removed, for each model instance. Red dashed lines are the baseline effect sizes for each instance and the gray dashed line represents a null effect size.

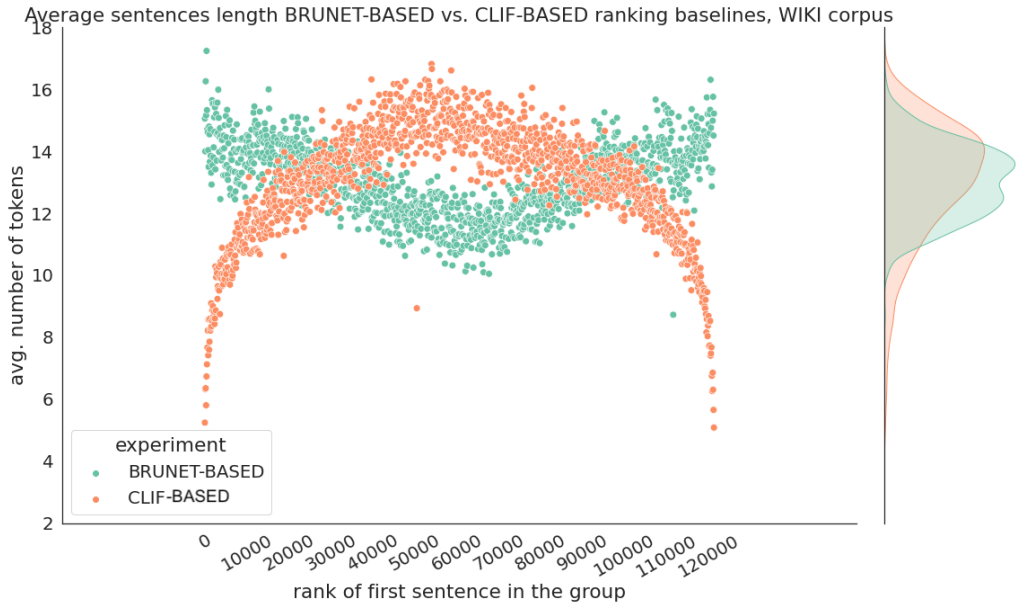


Figure 3: Comparison of the average length of sentences, grouped by 100, between BRUNET-BASED and CLIF-BASED average rankings.