Solving Cosine Similarity Underestimation between High Frequency Words by ℓ_2 Norm Discounting

Saeth Wannasuphoprasit[♠] **Yi Zhou**[◇] **Danushka Bollegala**[♣,♠]
University of Liverpool[♠], Cardiff University[◇], Amazon[♣]
{s.wannasuphoprasit,danushka}@liverpool.ac.uk
zhouy131@cardiff.ac.uk

Abstract

Cosine similarity between two words, computed using their contextualised token embeddings obtained from masked language models (MLMs) such as BERT has shown to underestimate the actual similarity between those words (Zhou et al., 2022). This similarity underestimation problem is particularly severe for highly frequent words. Although this problem has been noted in prior work, no solution has been proposed thus far. We observe that the ℓ_2 norm of contextualised embeddings of a word correlates with its log-frequency in the pretraining corpus. Consequently, the larger ℓ_2 norms associated with the highly frequent words reduce the cosine similarity values measured between them, thus underestimating the similarity scores. To solve this issue, we propose a method to *discount* the ℓ_2 norm of a contextualised word embedding by the frequency of that word in a corpus when measuring the cosine similarities between words. We show that the so called *stop* words behave differently from the rest of the words, which require special consideration during their discounting process. Experimental results on a contextualised word similarity dataset show that our proposed discounting method accurately solves the similarity underestimation problem.

1 Introduction

Cosine similarity is arguably the most popular word similarity measure used in numerous natural language processing (NLP) tasks, such as question answering (QA), information retrieval (IR) and machine translation (MT) (Echizen-ya et al., 2019; Oniani and Wang, 2020; Kim et al., 2022; Hanifi et al., 2022). First, a word is represented by a vector (aka *embedding*) and then the similarity between two words is computed as the cosine of the angle between the corresponding vectors (Rahutomo et al., 2012). Despite the good performance of cosine similarity as a similarity measure in various downstream tasks, Zhou et al. (2022) showed that it systematically underestimates the true similarity between highly frequent words, when computed using contextualised word embeddings obtained from MLMs such as BERT (Devlin et al., 2018).

Compared to the problem of estimating similarity between highly frequent words, the opposite problem of estimating the similarity between (or involving) rare (low frequency) words has received greater attention, especially in the scope of static word embeddings (Levy and Goldberg, 2014; Hellrich and Hahn, 2016; Mimno and Thompson, 2017; Wendlandt et al., 2018). If a word is rare in a corpus, we might not have a sufficiently large number of contexts containing that word to learn an accurate embedding for it. This often leads to unreliable similarity estimations between words and has undesirable implications in downstream tasks such as the detection of analogies and social biases (Ethayarajh et al., 2019a,b).

On the other hand, Zhou et al. (2022) studied the impact of frequency on contextualised word embeddings and showed that the cosine similarity between highly frequent words are systematically underestimated. Unlike in the previously discussed low frequency word scenario, we do have adequate contexts to learn an accurate semantic representation for highly frequent words. Therefore, it might appear surprising at first that cosine similarity cannot be correctly estimated even for the highly frequent words. Zhou et al. (2021) show that the diversity (measured by the volume of the bounding hypersphere) of the contextualised embeddings of a target word, computed from multiple contexts containing the word, increases with the frequency of that word. They provide an explanation that holds true only for 2-dimensional embeddings, which relates diversity to the underestimation of cosine similarity. Unfortunately, this explanation does not extend to the high dimensional embeddings used in practice by the NLP community (e.g. BERT token embeddings are typically more than 768 di-

Findings of the Association for Computational Linguistics: ACL 2023, pages 8644–8652

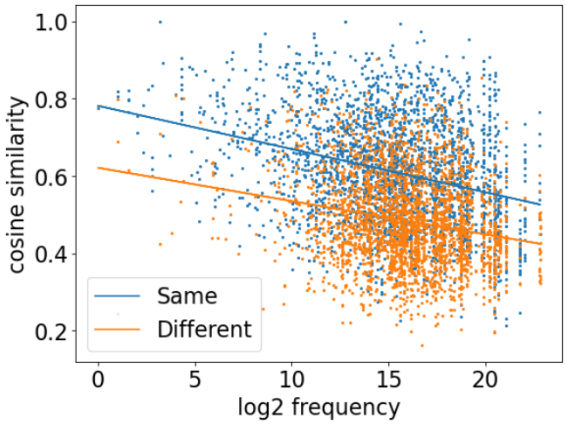

Figure 1: Cosine similarity between two instances of the same word w in two contexts in the WiC train dataset. When the log-frequency of w in the corpus increases, cosine similarities computed for both contexts that express the same meaning of w as well as its different meanings decreases.

mensional). More importantly, to the best of our knowledge, no solution has been proposed in the literature to address the cosine similarity underestimation problem associated with the highly frequent words.

In prior work, the ℓ_2 norm of a static word embedding has been shown to linearly correlate with the log-frequency of that word (Arora et al., 2016; Bollegala et al., 2018). On the other hand, we empirically study the ℓ_2 norm of the contextualised embedding of a word w averaged over all of its contexts, and find that it too approximately linearly correlates with the log-frequency of w in the corpus used to pretrain the MLM. Recall that the cosine similarity is defined as the inner-product between two embeddings, divided by the ℓ_2 norm of those embeddings. Therefore, we suspect that the underestimation of cosine similarity between highly frequent words is due to the larger ℓ_2 norms associated with those words.

To correct for this bias associated with the ℓ_2 norms of highly frequent words, we propose a linearly parameterised discounting scheme in the log-frequency space. Specifically, we use Monte-Carlo Bayesian Optimisation (Balandat et al., 2019) to find the optimal discounting parameters. Our proposed discounting method is shown to accurately correct the underestimation of cosine similarities between highly frequent words on the Word-in-Context (WiC) (Pilehvar and Camacho-Collados, 2019) dataset where human similarity ratings are available for the same word in two different contexts. Source code for reproducing the experiments reported in this is paper is publicly available.[1]

2 Underestimation of Cosine Similarity

Let us denote the d-dimensional contextualised word embedding produced by an MLM f for a target word w appearing in a context c by $\boldsymbol{f}(w, c) (\in \mathbb{R}^d)$. Moreover, let the set of contexts where w occurs in a given corpus be $\mathcal{S}(w)$. We refer to $\{\boldsymbol{f}(w, c) | w \in \mathcal{S}(w)\}$ as the set of *sibling embeddings* of w. To study the relationship between the cosine similarity scores and the frequency of words, we use the 768-dimensional `bert-base-uncased`[2] as the contextualised embedding model. We use the token embedding of w from the final hidden layer of BERT as $\boldsymbol{f}(w, c)$. We approximate the word frequencies in BERT pretraining corpus using the BookCorpus (Zhu et al., 2015). Let ψ_w be the frequency of w in this corpus.

We use the WiC dataset, which contains 5428 pairs of words appearing in various contexts with annotated human similarity judgements. WiC dataset is split into official training and development sets, while a separate hidden test set is used by the leaderboard for ranking Word Sense Disambiguation systems.[3] WiC dataset contains pairs of contexts labelled as having the **same meaning** (e.g. "to *drive* sheep out of a field" vs. "to *drive* the cows into the barn") and **different meaning** (e.g. "the *play* lasted two hours" vs. "they made a futile *play* for power").

We compute the cosine similarity between the two contextualised embeddings of a target word in two of its contexts to predict a similarity score. Figure 1 shows the predicted similarity scores for both contexts in which a target word has been used in the same or different meanings for all words in the WiC dataset against $\log(\psi_w)$. As seen from Figure 3, ψ_w has a power-law distribution. Therefore, we plot its log instead of raw frequency counts in Figure 1.

From Figure 1, we see that for both same as well as different meaning contexts, the predicted cosine similarities drop with the word frequencies. Moreover, the gradient of the drop for same meaning pairs (Pearson's $r = -0.3001$) is larger than that

[1] https://github.com/LivNLP/cosine-discounting

[2] https://huggingface.co/bert-base-uncased

[3] https://pilehvar.github.io/wic/

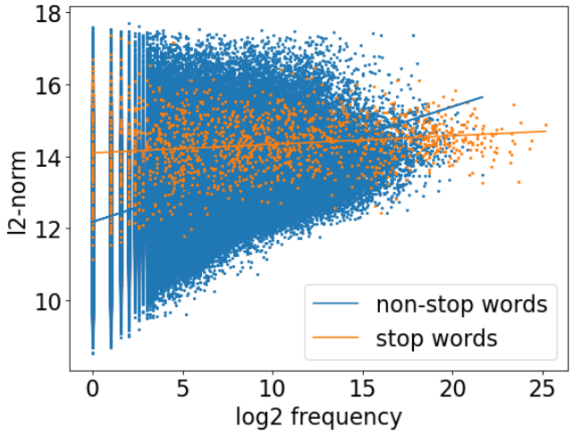

Figure 2: ℓ_2 norm of the averaged contextualised word embedding of a word against its log-frequency in the pretrain corpus. Stop words and non-stop words are shown respectively in orange and blue dots. Lines of best fits for each category are superimposed.

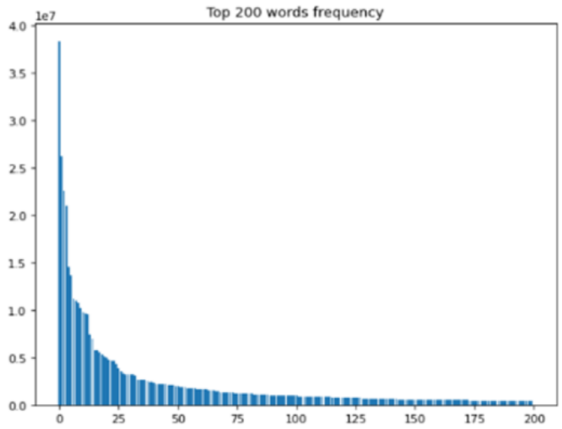

Figure 3: Histogram of word frequencies in the BERT pretrain corpus. We see a Zipfian (power-law) distribution, which turns out to be approximately liner in the log-frequency space.

for the different meaning pairs ($r = -0.2125$), indicating that the underestimation of cosine similarity is more sever for the similar contexts of highly frequent words.

3 ℓ_2 norm Discounting

To understand the possible reasons behind the cosine similarity underestimation for highly frequent words discussed in § 2, for each word w we compute its mean sibling embedding, $\hat{\boldsymbol{w}}$, given by (1).

$$\hat{\boldsymbol{w}} = \frac{1}{|\mathcal{S}(w)|} \sum_{c \in \mathcal{S}(w)} \boldsymbol{f}(w, c) \tag{1}$$

We plot $||\hat{\boldsymbol{w}}||$ against $\log(\psi(w))$ in Figure 2 separately for a predefined set of stop words and all other words (i.e. non-stop words). For this purpose, we use the default 1466 stop words from NLTK and randomly selected 997,425 non-stop words from the BookCorpus. Pearson r values of stop words and non-stop words are respectively 0.1697 and 0.3754, while the lines of best fits for each class of words are superimposed. From Figure 2, we see that overall, $||\hat{\boldsymbol{w}}||$ increases with $\log(\psi_w)$ for both stop and non-stop words, while the linear correlation is stronger in the latter class. Considering that stop words cover function words such as determiners and conjunctions that co-occur with a large number of words in diverse contexts, we believe that the ℓ_2 norm of stop words mostly remains independent of their frequency. Recall that the cosine similarity between two words is defined as the fraction of the inner-product of the corresponding embeddings, divided by the product of the ℓ_2 norms of the embeddings. Therefore, even if the inner-product between two words remain relatively stable, it will be divided by increasingly larger ℓ_2 norms in the case of highly frequent words. Moreover, this bias is further amplified when both words are high frequent due to the *product* of ℓ_2 norms in the denominator.

To address this problem, we propose to discount the ℓ_2 norm of a word w by a discounting term, $\alpha(\psi_w)$, and propose a discounted version of the cosine similarity given by (2).

$$\cos_\alpha(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x}^\top \boldsymbol{y}}{||\boldsymbol{x}||\,\alpha(\psi_x)\,||\boldsymbol{y}||\,\alpha(\psi_y)} \tag{2}$$

Following Figure 2, we linearly parameterise $\alpha(\psi_w)$ separately for stop vs. non-stop words as in (3).

$$\alpha(\psi_w) = \begin{cases} 1 + m_s(b_s - \log(\psi_w)) & \text{w is a stop word} \\ 1 + m_n(b_n - \log(\psi_w)) & \text{w is a non-stop word} \end{cases} \tag{3}$$

The scalar parameters m_s, m_n, b_s and b_n are estimated as follows. First, we randomly initialise all parameters uniformly in $[0, 1]$ and use (2) to predict cosine similarity between two contexts in which a target word w occurs in the WiC train instances. We then make a binary similarity judgement (i.e. **same** or **different** meaning) for the pair of contexts in an instance depending on whether the predicted cosine similarity is greater than a threshold θ. Next, we compute the overall binary classification accuracy for the similarity predictions made on the entire WiC training dataset,

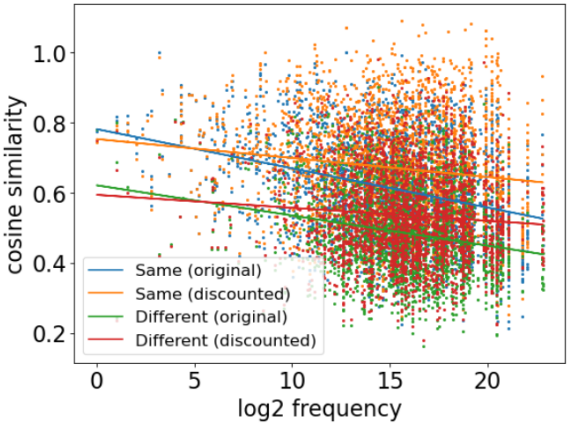

Figure 4: Cosine similarity between two instances of the same word w in two contexts in the WiC train dataset, computed using the original (non-discounted) cosine similarity (shown in blue and green respectively for the same and different meaning pairs) and using the proposed ℓ_2 norm discounted ((2)) (shown in orange and red respectively for the same and different meaning pairs). We see that the gradients of the drops have *decreased* for both same and different meaning pairs *after* applying the discounting.

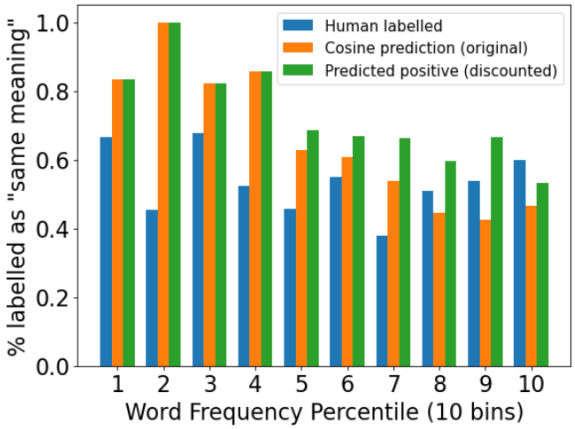

Figure 5: Percentage of examples labelled as having the "same meaning". In high frequency words, we see that the cosine similarity-based predictions (orange/middle) are systematically **underestimate** the human similarity judgements (blue/left). However, after the proposed discounting method has been applied (green/right) the underestimation has reduced.

and use Bayesian Optimisation to find the optimal values: $\theta = 0.545$, $m_s = 0.00422$, $b_s = 0.643$, $m_n = 0.00427$ and $b_n = 4.821$. Specifically we used the Adaptive Experimentation Platform[4] for learning those optimal values. We found this is more efficient than conducting a linear search over the parameter space. We repeat the estimation five times and use the averaged parameter values in the remainder of the experiments. Note that $m_n > m_s$ above, which indicates that non-stop words must be discounted slightly more heavily than the stop words. This makes sense since the impact of word frequency of non-stop words on their ℓ_2-norm is stronger than that for the stop words as indicated by the slopes of the lines of best fit in Figure 2.

4 Results

To evaluate the effect of the proposed ℓ_2 norm discounting when computing cosine similarity, we repeat the analysis presented in Figure 1 using (2) to predict the similarity between contextualised word embeddings. Comparing the lines of best fit for the original (blue, $r = -0.3006$) vs. discounted (orange, $r = -0.1366$) for the same meaning contexts, we see that the gradient of the drop has decreased by 51.65%. Likewise, comparing the lines of best fit for the original (green, $r = -0.2125$) vs. discounted (red, $r = -0.0843$) for the different meaning contexts, we see the gradient of the drop has decreased by 57.04%. This result clearly shows that the proposed ℓ_2 norm discounting method is able to reduce the underestimation of cosine similarities for the highly frequent words.

Given that the discounting parameters in (3) are learned from the WiC train data, it remains an open question as to how well the proposed discounting method generalises when predicting similarity between contextualised embeddings of unseen words. To evaluate this generalisability of the proposed method, we use (3) with its learned parameters from WiC train data, to predict the similarity between contextualised word embeddings in WiC dev data.[5] Specifically, we predict binary (same vs. different meaning) similarity labels according to the similarity threshold θ learnt in § 3 and compare against the human judgements using binary classification accuracy.

The maximum accuracy on WiC dev split obtained using the original (non-discounted) cosine similarities is 0.6667, which indicates that the cosine similarity is somewhat predictive of the human binary judgements. The overall F1 is improved by 2.4% (0.68 with original cosine vs. 0.71 with the proposed discounting method) and recall is improved by 12% (0.75 with original cosine vs. 0.84 with the proposed). On the other hand, the drop

[4] https://ax.dev/

[5] Note that the test set of WiC is publicly unavailable due to being used in a leaderboard.

in precision is 4.7% (from 0.64 to 0.61). Therefore, the proposed method solves the cosine similarity underestimation problem associated with high-frequent words, without significantly affecting the similarity scores for low-frequent ones

Figure 5 shows the average proportion of instances predicted to be the same meaning as a function of frequency, grouped into ten bins, each with the same number of examples. From Figure 5, we see that in high frequency bins (i.e. bins 8, 9 and 10), the percentage of predicted instances as having the same meaning is consistently lower than that compared to the human judgements. This shows an underestimation of the true (human judged) similarity between contextualised word embeddings.

On the other hand, when we use the proposed ℓ_2 norm discounted cosine similarity (defined in (2)), in the highest frequent bin (i.e. 10) we see that the gap between human judgements vs. predicted similarities has reduced. Moreover, in the low frequency bins (i.e. 1–4), we see that the proposed discounting method does not affect the predictions made using cosine similarities. We see an overestimation of the cosine similarities in the low frequency bins as reported by Zhou et al. (2021). As discussed already in § 1, the word embeddings learnt for low frequency words tend to be unreliable due to data sparseness. Therefore, we believe it is important to focus on the problem of learning accurate word embeddings rather than to adjust cosine similarities between low-frequency words in a post-processing step.

We see that in bins 5, 6 and 7 the similarity scores are slightly increased by the proposed discounting method, which is a drawback that needs to be addressed in future work. More importantly however, the overall percentage recall across all bins for retrieving same meaning instances improves significantly from 74.7% to 83.7% compared to using respectively the original cosine similarity vs. the discounted cosine similarity. Overall, this result confirms the validity of the proposed discounting method for addressing the underestimation of cosine similarity involving highly frequent words.

5 Conclusion

We proposed a method to solve the cosine similarity underestimation problem in highly frequent words. Specifically, we observed that the ℓ_2 norm of a contextualised word embedding increases with its frequency in the pretrain corpus and proposed a discounting scheme. Experimental results on WiC dataset confirmed the validity of the proposed method.

6 Limitations

We proposed a solution to the cosine similarity underestimation problem associated with contextualised word embeddings of highly frequent words. Our evaluations used only a single contextualised embedding model (i.e. BERT) with a single dimensionality (i.e. 768). Therefore, we believe that our proposed method must be evaluated with other (more recent) MLMs to test for its generalisability. Moreover, our evaluations were conducted only on the English language, which is known to be morphologically limited. Although in our preliminary experiments we considered discounting schemes based on the part-of-speech of words (instead of considering stop words vs. non-stop words), we did not find any significant improvements despite the extra complexity. However, these outcomes might be different for more morphologically richer languages. In order to evaluate similarity predictions in other languages, we must also have datasets similar to WiC annotated in those languages, which are difficult to construct. Although having stated that using a single MLM and single language as limitations of this work, we would like to point out that these are the same conditions under which Zhou et al. (2022) studied the cosine similarity underestimation problem.

We used only a single dataset (i.e. WiC) in our experiments in this short paper due to space constraints. Other contextual similarity datasets (e.g. Stanford Contextualised Word Similarity (SCWS) (Huang et al., 2012)) could be easily used to further validate the proposed discounting method in an extended version.

7 Ethical Considerations

In this paper, we do not annotate novel datasets nor release any fine-tuned MLMs. Therefore, we do not see any direct ethical issues arising from our work. However, we are proposing a method to address the underestimation of cosine similarity scores computed using contextualised word embeddings obtained from (possibly socially biased) pretrained MLMs. We would therefore discuss the ethical implication of this aspect of our work in this section.

Cosine similarity has been used in various social bias evaluation measures such as the WEAT (Caliskan et al., 2017), SemBias (Zhao et al., 2018), WAT (Du et al., 2019), etc. These methods measure the cosine similarity between a gender and a set of pleasant or unpleasant set of attributes to compute a social bias evaluation score. Although originally these methods were developed for evaluating the social biases in static word embeddings, they have been later extended to contextualised word embeddings (Kaneko and Bollegala, 2022; Kaneko et al., 2022) and sentence embeddings (May et al., 2019), where cosine similarity still remains the main underlying metric. However, Ethayarajh et al. (2019c) showed that inner-products to be superior over cosine similarity for social bias evaluation purposes. It remains unclear as to how the underestimation in cosine similarities discussed in our work would influence the social bias evaluations. In particular, the effect of the proposed ℓ_2 norm discounting scheme on social bias evaluation must be carefully studied in the future work.

Acknowledgements

Danushka Bollegala holds concurrent appointments as a Professor at University of Liverpool and as an Amazon Scholar. This paper describes work performed at the University of Liverpool and is not associated with Amazon.

References

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2016. A latent variable model approach to PMI-based word embeddings. *Transactions of the Association for Computational Linguistics* 4:385–399. https://aclanthology.org/Q16-1028.

Maximilian Balandat, Brian Karrer, Daniel R. Jiang, Samuel Daulton, Benjamin Letham, Andrew Gordon Wilson, and Eytan Bakshy. 2019. BoTorch: A Framework for Efficient Monte-Carlo Bayesian Optimization. Advances in Neural Information Processing Systems 33, 2020.

Danushka Bollegala, Yuichi Yoshida, and Ken-ichi Kawarabayashi. 2018. Using k-way Co-occurrences for Learning Word Embeddings. In *Proc. of AAAI*. pages 5037–5044.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356:183–186.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. https://doi.org/10.48550/ARXIV.1810.04805.

Yupei Du, Yuanbin Wu, and Man Lan. 2019. Exploring human gender stereotypes with word association test. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, pages 6132–6142. https://doi.org/10.18653/v1/D19-1635.

Hiroshi Echizen-ya, Kenji Araki, and Eduard Hovy. 2019. Word embedding-based automatic mt evaluation metric using word position information. 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)*. pages 1874–1883.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019a. Towards understanding linear word analogies. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, pages 3253–3262. https://doi.org/10.18653/v1/P19-1315.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019b. Understanding undesirable word embedding associations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. pages 1696–1705.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019c. Understanding undesirable word embedding associations. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, pages 1696–1705.

Masih Hanifi, Hicham Chibane, Remy Houssin, and Denis Cavallucci. 2022. Problem formulation in inventive design using doc2vec and cosine similarity as artificial intelligence methods and scientific papers. *Engineering Applications of Artificial Intelligence* 109:104661.

Johannes Hellrich and Udo Hahn. 2016. Bad Company—Neighborhoods in neural embedding spaces considered harmful. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, Osaka, Japan, pages 2785–2796. https://aclanthology.org/C16-1262.

Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. 2012. Improving word representations via global context and multiple word prototypes. In *ACL'12*. pages 873 – 882.

Masahiro Kaneko and Danushka Bollegala. 2022. Unmasking the mask – evaluating social biases in masked language models. In *Proc. of the 36th AAAI Conference on Artificial Intelligence*.

Masahiro Kaneko, Aizhan Imankulova, Danushka Bollegala, and Naoaki Okazaki. 2022. Gender bias in masked language models for multiple languages. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Seattle, United States, pages 2740–2750. https://aclanthology.org/2022.naacl-main.197.pdf.

Suyoun Kim, Duc Le, Weiyi Zheng, Tarun Singh, Abhinav Arora, Xiaoyu Zhai, Christian Fuegen, Ozlem Kalinli, and Michael L. Seltzer. 2022. Evaluating User Perception of Speech Recognition System Quality with Semantic Distance Metric. In *Proc. of INTERSPEECH*.

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. *Advances in neural information processing systems* 27.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. 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, pages 622–628. https://www.aclweb.org/anthology/N19-1063.

David Mimno and Laure Thompson. 2017. The strange geometry of skip-gram with negative sampling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, pages 2873–2878. https://doi.org/10.18653/v1/D17-1308.

David Oniani and Yanshan Wang. 2020. A qualitative evaluation of language models on automatic question-answering for covid-19. In *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. pages 1–9.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning 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)*. Association for Computational Linguistics, Minneapolis, Minnesota, pages 1267–1273.

Faisal Rahutomo, Teruaki Kitasuka, and Masayoshi Aritsugi. 2012. Semantic cosine similarity. In *The 7th international student conference on advanced science and technology ICAST*. 1, page 1.

Laura Wendlandt, Jonathan K. Kummerfeld, and Rada Mihalcea. 2018. Factors influencing the surprising instability of word embeddings. 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, pages 2092–2102. https://doi.org/10.18653/v1/N18-1190.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Association for Computational Linguistics, pages 15–20. http://aclweb.org/anthology/N18-2003.

Kaitlyn Zhou, Kawin Ethayarajh, Dallas Card, and Dan Jurafsky. 2022. Problems with cosine as a measure of embedding similarity for high frequency words. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Dublin, Ireland, pages 401–423. https://doi.org/10.18653/v1/2022.acl-short.45.

Kaitlyn Zhou, Kawin Ethayarajh, and Dan Jurafsky. 2021. Frequency-based distortions in contextualized word embeddings. *arXiv preprint arXiv:2104.08465* .

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *The IEEE International Conference on Computer Vision (ICCV)*.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
section 6

☑ A2. Did you discuss any potential risks of your work?
Section 7

☑ A3. Do the abstract and introduction summarize the paper's main claims?
abstract and section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

sections 2 and 3

☑ B1. Did you cite the creators of artifacts you used?
section 2

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
sections 2 and 3

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
sections 3 and 4

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
sections 3 and 4

C ☑ Did you run computational experiments?

sections 3 and 4

☒ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
This is a methodology paper and not related to computation cost.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Sections 3 and 4

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
section 4

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
section 3 and 4

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.