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

SensePOLAR: Word sense aware interpretability for pre-trained contextual word embeddings

Keywords: interpretability, contextual word embeddings, pre-trained language model, semantic differential, explainable AI

Word embeddings have proven essential to many cutting-edge applications in natural language processing (NLP) and have been widely adopted in various domains, for example, to study stereotypes [1], social class [2] or public health [3]. As language models become more complex and continue to exhibit biases [4], their lack of *interpretability* becomes a growing concern [5].

To improve interpretability, recent approaches such as SemAxis [6], POLAR [7], and BiImp [8] have explored the potential of embedding words via polar dimensions (e.g. good ↔ bad, correct ↔ wrong). However, these approaches are limited to interpreting traditional (*static*) word embedding models like Word2Vec [9] or Glove [10] and have not been designed to deal with polysemy, i.e. multiple senses of words. Currently, static embedding models are being replaced by *contextual* word embedding models like BERT [11] which have achieved competitive performance in various NLP benchmarks. Addressing polysemy, in this paper we aim to enable *word-sense aware* interpretability for pre-trained *contextual* word embeddings.

We base our approach on the original POLAR framework [7] and extend it to contextual word embeddings. The key idea is to transform pre-trained word embeddings into an interpretable, *sense aware* space. In this space, each dimension represents a scale on which words are rated, inspired by the idea of semantic differentials [12], which are psychometric scales between two antonym words. In contrast to existing approaches, we define opposite *senses* for the poles of these scales (e.g. "left direction" ↔ "right direction"), as opposed to opposite words (e.g. "left" ↔ "right"), as used in [7].

The individual steps of our approach are illustrated in Figure 1 and outlined below. 1) Given a contextual word embedding model $\mathcal{M}$, we use $\mathcal{M}$ to obtain the (non-interpretable) contextual embedding space. 2) We obtain polar senses with contextual information from an oracle. 3) We proceed with generating representative sense embeddings from which we 4) construct the interpretable polar sense space. 5) The original embedding is transformed into the polar sense space, which enables interpretation with regard to opposite sense pairs.

For our experiments, we use BERT [11] as embedding model and WordNet [13] as oracle with 1763 polar sense pairs and corresponding contexts. We show that our interpretable dimensions align with human judgements, with the top-k dimensions selected by SensePOLAR matching those that were selected by human annotators much more often than might be expected from chance alone. Figure 2 illustrates that SensePOLAR allows for interpretation along different senses of a word and can uncover the connotative meaning of words.

Remarkably, we find that using interpretable embeddings provided by SensePOLAR does not decrease performance in downstream tasks. In particular, we test SensePOLAR by using it to create input features for a separate model (feature-based approach) as well as directly integrated in the model itself (fine-tuning approach) on several NLP benchmarks (e.g. GLUE [14] and SQuAD [15]). SensePOLAR shows competitive performance on all tasks in comparison to the original BERT model [11].

By adding interpretability to state-of-the-art word embedding models, SensePOLAR allows for a number of applications. We demonstrate this by employing SensePOLAR for bias analysis. By comparing embeddings related to different ethnicities, we find that the most discriminative interpretable dimensions relate to harmful stereotypes such as employment status and legal status. SensePOLAR's capacity to provide interpretability offers valuable insights into the biases inherent in pre-trained word embeddings and can guide future efforts to mitigate these biases.

References

[1] N. Garg, L. Schiebinger, D. Jurafsky, and J. Zou, "Word embeddings quantify 100 years of gender and ethnic stereotypes," *Proceedings of the National Academy of Sciences*, vol. 115, no. 16, pp. E3635–E3644, 2018.

[2] A. C. Kozlowski, M. Taddy, and J. A. Evans, "The geometry of culture: Analyzing the meanings of class through word embeddings," *American Sociological Review*, vol. 84, no. 5, pp. 905–949, 2019.

[3] X. Dai, M. Bikdash, and B. Meyer, "From social media to public health surveillance: Word embedding based clustering method for twitter classification," in *SoutheastCon 2017*, pp. 1–7, 2017.

[4] K. Kurita, N. Vyas, A. Pareek, A. W. Black, and Y. Tsvetkov, "Measuring bias in contextualized word representations," in *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, (Florence, Italy), pp. 166–172, Association for Computational Linguistics, Aug. 2019.

[5] M. Ribeiro, S. Singh, and C. Guestrin, ""why should I trust you?": Explaining the predictions of any classifier," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, (San Diego, California), pp. 97–101, Association for Computational Linguistics, June 2016.

[6] J. An, H. Kwak, and Y.-Y. Ahn, "SemAxis: A lightweight framework to characterize domain-specific word semantics beyond sentiment," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (Melbourne, Australia), pp. 2450–2461, July 2018.

[7] B. Mathew, S. Sikdar, F. Lemmerich, and M. Strohmaier, "The polar framework: Polar opposites enable interpretability of pre-trained word embeddings," in *Proceedings of The Web Conference 2020*, pp. 1548–1558, 2020.

[8] L. K. Şenel, F. Şahinuç, V. Yücesoy, H. Schütze, T. Çukur, and A. Koç, "Learning interpretable word embeddings via bidirectional alignment of dimensions with semantic concepts," *Information Processing & Management*, vol. 59, no. 3, p. 102925, 2022.

[9] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *Proceedings of the 1st International Conference on Learning Representations*, vol. 2013, 01 2013.

[10] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1532–1543, 2014.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 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. 4171–4186, June 2019.

[12] C. E. Osgood, G. J. Suci, and P. H. Tannenbaum, *The Measurement of Meaning*. No. 47, University of Illinois press, 1957.

[13] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[14] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, (Brussels, Belgium), pp. 353–355, Association for Computational Linguistics, Nov. 2018.

[15] P. Rajpurkar, R. Jia, and P. Liang, "Know what you don't know: Unanswerable questions for SQuAD," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, (Melbourne, Australia), pp. 784–789, July 2018.

Figures

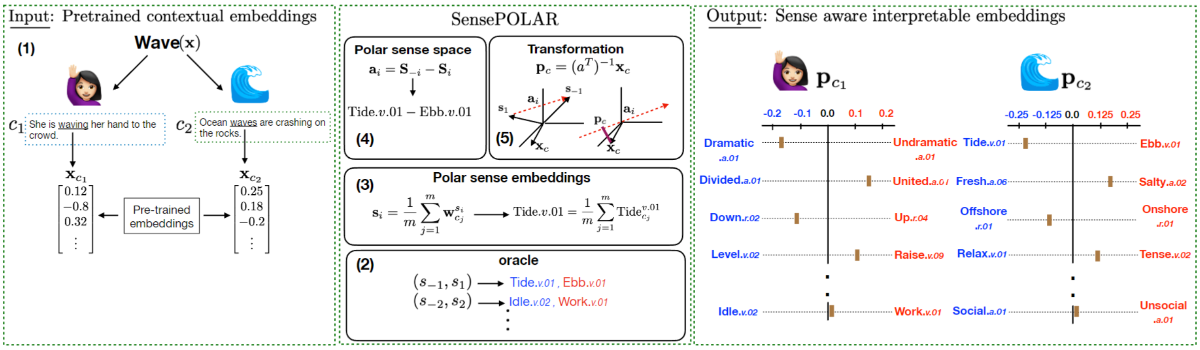
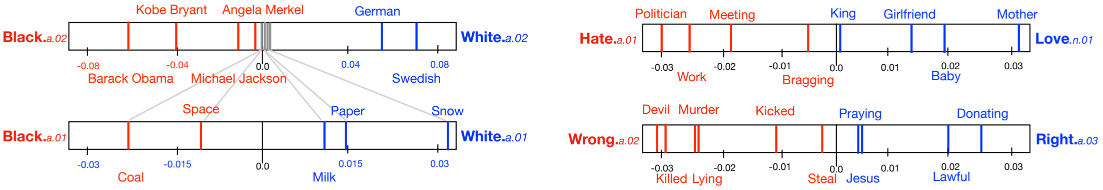

Input: Pretrained contextual embeddings
(1)
Wave(x)
c_1 She is waving her hand to the crowd.
c_2 Ocean waves are crashing on the rocks.
$\mathbf{x}_{c_1}$
$\mathbf{x}_{c_2}$
0.12
−0.8
0.32
0.25
0.18
−0.2
Pre-trained embeddings
SensePOLAR
Polar sense space
$\mathbf{a}_i = \mathbf{S}_{-i} - \mathbf{S}_i$
Tide.v.01 − Ebb.v.01
(4)
Transformation
$\mathbf{p}_c = (a^T)^{-1}\mathbf{x}_c$
(5)
(3)
Polar sense embeddings
$\mathbf{s}_i = \frac{1}{m}\sum_{j=1}^{m}\mathbf{w}_{c_j}^{s_i} \longrightarrow \text{Tide.v.01} = \frac{1}{m}\sum_{j=1}^{m}\text{Tide}_{c_j}^{v.01}$
(2)
oracle
$(s_{-1}, s_1) \longrightarrow$ Tide.v.01 , Ebb.v.01
$(s_{-2}, s_2) \longrightarrow$ Idle.v.02 , Work.v.01
Output: Sense aware interpretable embeddings
$\mathbf{p}_{c_1}$
$\mathbf{p}_{c_2}$
-0.2 -0.1 0.0 0.1 0.2
-0.25 -0.125 0.0 0.125 0.25
Dramatic.a.01
Undramatic.a.01
Divided.a.01
United.a.01
Down.r.02
Up.r.04
Level.v.02
Raise.v.09
Idle.v.02
Work.v.01
Tide.v.01
Ebb.v.01
Fresh.a.06
Salty.a.02
Offshore.r.01
Onshore.r.01
Relax.v.01
Tense.v.02
Social.a.01
Unsocial.a.01

Figure 1: SensePOLAR overview. Pre-trained contextual word embeddings are transformed into an interpretable space where the word's semantics are rated on scales individually encoded by opposite senses such as "good" ↔ "bad". The scores across the dimensions are representative of the strength of relationship (between word and dimension) which allows us to rank the dimensions and thereby identify the most discriminative dimensions for a word. In this example, the word "wave" is used in two senses: *hand waving* and *ocean wave*. SensePOLAR not only generates dimensions that are representative of individual contextual meanings, the alignment to the respective sense spaces also aligns well with human judgement. SensePOLAR generates neutral scores for dimensions not related to the word in the given context (e.g., "idle" ↔ "work", "social" ↔ "unsocial"). We follow the WordNet [13] convention to represent a particular sense of a word. For example, "Tide.v.01" represents the word "tide" in the sense of *surge (rise or move forward)*.



(a) Benefits of sense dimensions

(b) Connotative meanings across dimensions

Figure 2: Illustration of polar sense dimensions. (a) SensePOLAR allows for interpretability along multiple senses. "black" ↔ "white" in the sense of *ethnicity* (top) can be differentiated from "black" ↔ "white" in the sense of *color* (bottom). Words like "snow"or "coal" - which are not semantically related to ethnicity - score neutral on the upper scale while being clearly distinguishable on the lower scale. (b) The connotative meanings of words can also be investigated through SensePOLAR. For example, "politician" is associated with "hate" while "mother" is associated with "love".