Word Embeddings Revisited: Do LLMs Offer Something New?

Matthew Freestone, Shubhra Kanti Karmaker Santu
BDI Lab, Auburn University, Alabama, USA
{maf0083, sks0086}@auburn.edu

1 Introduction

Learning meaningful word embeddings is key to training a powerful language model. The rise of Large Language Models (LLMs) has provided us with many new word embedding models recently. Although LLMs have shown remarkable advancement in various NLP tasks (Bubeck et al., 2023; Dai et al., 2022; Du et al., 2022; Smith et al., 2022), it is still unclear whether the performance improvement is merely because of scale or whether the underlying embeddings they produce are significantly different from classical encoding models like Sentence-BERT or Universal Sentence Encoders. In this paper, we systematically investigate this issue by comparing classical embedding techniques against LLM-based embeddings in terms of their latent vector semantics.

2 Experiments and Results

2.1 Embeddings

A sampling of models from the last five years will be used to create a multitude of embeddings for experimentation. We classify these models into two broad categories: Large Language Models (LLMs) – models with over 1B parameters, and "Classical" models – those with under 1B parameters. These embedding models, with their respective embedding length, are as follows:

1. GPT-ADA-002 (LLM) (Brown et al., 2020) - 1536
2. LLAMA2-7B (LLM) (Touvron et al., 2023) - 4096
3. PaLM-Gecko-001 (LLM) (Chowdhery et al., 2022) - 768
4. USE (Classical) (Cer et al., 2018) - 512
5. LASER (Classical) (Artetxe and Schwenk, 2019) - 1024
6. SBERT (Classical) (Reimers and Gurevych, 2019) - 384

2.2 Similar Word Closeness

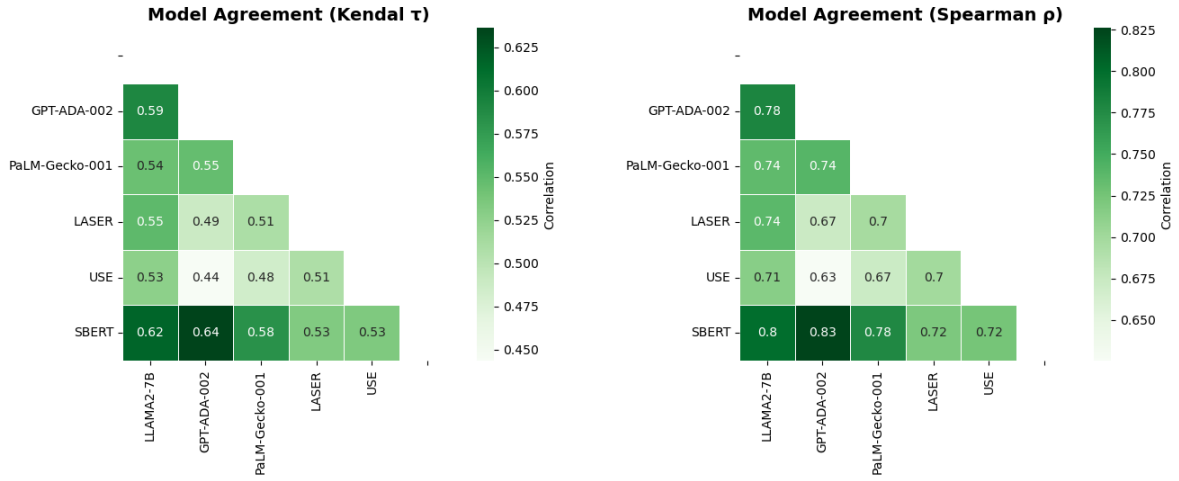
Q: Do LLM embeddings capture related-word similarity more strongly than classical models?
To answer this question, a corpus of approximately $80,000$ well-distributed words was created from WordNet (Fellbaum, 1998). The vector representations for all of these words were created from each embedding. The distribution of cosine similarities of all pairs of words (≈ 6.4 billion pairs) was found as a baseline for future tests. The BATS dataset provides pairs of related words across many categories, which describe how those words are related. The distribution of cosine similarities between these pairs of words (by broad category) next to the distribution of non-related pairs of words for all embeddings is shown in Figure 2.
From this visualization, we can see that ADA, PALM, and SBERT showed the strongest separation of related words from all words across both broad categories of analogies. In all models, morphologically-related words were more similar to each other than semantically-related ones. In some models, especially USE, the semantically-related word pairs were distributed in similarity nearly the same as all words pairs.
Some LLMs, particularly ADA and PALM, capture semantic relations between words more strongly than classical models.

2.3 Analogy Tasks

Q: How do LLMs perform on word analogy tasks, and are certain methods more or less effective on them?
The analogy tasks proposed by (Mikolov et al., 2013) were evaluated on different models using BATS word pairs. Different approaches such as 3CosAdd, PairDistance, 3CosMul, 3CosAvg, LRCos, SimilarToAny, and SimilarToB were tested. We measure top-1 accuracy, excluding analogy input words from the possible answers. The same word corpus was used for all models. Results

Method Model Name	3CosAdd	3CosAvg	3CosMul	LRCos	PairDistance	SimilarToAny	SimilarToB
ADA-002	**0.412**	0.447	**0.424**	0.375	**0.232**	0.058	0.135
LASER	0.227	0.260	0.237	0.284	0.121	0.032	0.076
LLAMA2	0.145	0.200	0.145	0.131	0.053	0.039	0.082
PaLM 2	0.398	**0.458**	0.417	**0.534**	0.193	0.060	0.123
SBERT	0.243	0.261	0.267	0.487	0.086	**0.067**	**0.141**
USE	0.174	0.212	0.187	0.450	0.025	0.043	0.107

Table 1: Results of BATS analogy task for each model by method.



(a) Kendal τ

(b) Spearman ρ

Figure 1: Correlation coefficients for each pair of models, found using a large dataset of pairs of words.

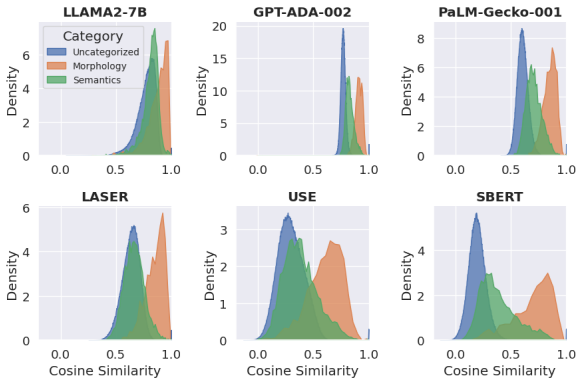


Figure 2: Distribution of random set of words plotted against distribution of lexicographically and semantically related words.

showed ADA and PALM performing well with 3Cos-style methods, whereas LLAMA lagged behind. LRCos notably boosted accuracy for SBERT and USE and achieved the highest overall accuracy with PALM embeddings. A short summary of each model's accuracy is given in Figure 1.

LLMs (ADA and PALM) performed very well with the 3CosAvg method, and all LLMs saw less benefit from the LRCos method than classical models.

2.4 Model Agreement on Similarity

Q: Do LLM Embeddings agree on the relative similarity of words compared to classical models?

In order to create a direct comparison on the similarity of models in general, statistical measures of correlation can be used. First, all the cosine similarity of all pairs of words in the Section 2.2 corpus is found for each model. These similarities can now act as an annotated "score" for each word pair, and because the corpus of words is common, correlation between two different embeddings' scores can be measured. Both Kendall's τ and Pearson's ρ between each pair of models is shown in Figure 1. This metric shows SBERT and ADA-002 to be the most similar models, while USE and ADA are most different.

Two of the LLMs, PaLM and ADA tended to agree with each other, but they also surprisingly meaningfully agree with SBERT.

A Appendix

See Table 2 for a more granular description of the performance of each model on specific categories of BATS.

Model	Analogy Method	1. Inflectional Morphology	2. Derivational Morphology	3. Encyclopedic Semantics	4. Lexicographic Semantics
LLAMA2-7B	3CosAdd	0.230	0.271	0.055	0.023
LLAMA2-7B	3CosAvg	0.326	0.362	0.086	0.026
LLAMA2-7B	3CosMul	0.230	0.276	0.053	0.022
LLAMA2-7B	LRCos	0.150	0.148	0.176	0.050
LLAMA2-7B	PairDistance	0.066	0.130	0.013	0.001
LLAMA2-7B	SimilarToAny	0.065	0.043	0.037	0.011
LLAMA2-7B	SimilarToB	0.130	0.118	0.054	0.026
PaLM-Gecko-001	3CosAdd	0.743	0.609	0.118	0.122
PaLM-Gecko-001	3CosAvg	0.794	0.668	0.232	0.136
PaLM-Gecko-001	3CosMul	0.768	0.648	0.128	0.124
PaLM-Gecko-001	LRCos	0.780	0.714	**0.404**	0.238
PaLM-Gecko-001	PairDistance	0.466	0.249	0.048	0.008
PaLM-Gecko-001	SimilarToAny	0.165	0.027	0.011	0.035
PaLM-Gecko-001	SimilarToB	0.270	0.082	0.030	0.108
GPT-ADA-002	3CosAdd	0.761	0.677	0.115	0.097
GPT-ADA-002	3CosAvg	0.802	**0.734**	0.148	0.102
GPT-ADA-002	3CosMul	0.776	0.697	0.122	0.100
GPT-ADA-002	LRCos	0.606	0.482	0.280	0.132
GPT-ADA-002	PairDistance	0.546	0.323	0.052	0.006
GPT-ADA-002	SimilarToAny	0.155	0.044	0.005	0.029
GPT-ADA-002	SimilarToB	0.276	0.134	0.038	0.090
LASER	3CosAdd	0.431	0.434	0.022	0.022
LASER	3CosAvg	0.484	0.506	0.030	0.020
LASER	3CosMul	0.448	0.454	0.023	0.023
LASER	LRCos	0.510	0.482	0.116	0.028
LASER	PairDistance	0.230	0.245	0.009	0.003
LASER	SimilarToAny	0.087	0.027	0.004	0.007
LASER	SimilarToB	0.198	0.072	0.012	0.020
USE	3CosAdd	0.397	0.156	0.039	0.103
USE	3CosAvg	0.442	0.190	0.084	0.132
USE	3CosMul	0.436	0.165	0.049	0.100
USE	LRCos	0.722	0.412	0.396	**0.270**
USE	PairDistance	0.076	0.012	0.008	0.005
USE	SimilarToAny	0.101	0.032	0.006	0.035
USE	SimilarToB	0.204	0.098	0.026	0.098
SBERT	3CosAdd	0.461	0.393	0.046	0.073
SBERT	3CosAvg	0.474	0.418	0.058	0.092
SBERT	3CosMul	0.506	0.424	0.062	0.074
SBERT	LRCos	**0.808**	0.642	0.270	0.228
SBERT	PairDistance	0.135	0.184	0.021	0.003
SBERT	SimilarToAny	0.178	0.065	0.003	0.019
SBERT	SimilarToB	0.302	0.154	0.020	0.088

Table 2: BATS performance across categories with methods.

References

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Daniel Cer, Yinfei Yang, Sheng yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Zhifang Sui, and Furu Wei. 2022. Why can gpt learn in-context? language models secretly perform gradient descent as meta optimizers. *arXiv preprint arXiv:2212.10559*.

Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. 2022. Glam: Efficient scaling of language models with mixture-of-experts. In *International Conference on Machine Learning*, pages 5547–5569. PMLR.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, et al. 2022. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.