Word Embeddings Revisited: Do LLMs Offer Something New?

AUBURN UNIVERSITY
Samuel Ginn College of Engineering

Matthew Freestone, Shubhra Kanti Karmaker Santu
{maf0083, sks0086}@auburn.edu

KEY FINDINGS

- LLMs are not always better than classical models in capturing semantic similarity (e.g. SBERT vs LLaMa)
- ADA and PALM perform significantly better than classical models on word analogy tasks. SBERT (a classic model) is often ranked as third.
- Two of the LLMs, PaLM and ADA, tended to agree with each other, but they also surprisingly meaningfully agreed with SBERT.
- SBERT can be an efficient alternative to LLMs when resources are constrained.

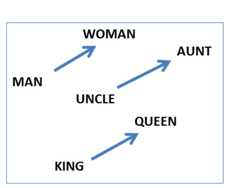
TASKS AND DATASETS

Word Analogy Task

- Measures the ability of an embedding to encode information about the relation of words.
- For words a,b,c,d, analogy a:b::c:d, embedding function f(x):

$$f(b) - f(a) + f(c) \approx f(d)$$

- Above method is *3CosAdd;* other methods have been proposed and tried here



Linguistic Regularities in Continuous Space Word Representations (Mikolov et al., NAACL 2013)

Mikolov (Google) Analogy Set

- 9 morphological, 5 semantic categories
- 20-70 word-pairs per category
- Unbalanced; most semantic questions are country:capital

Linguistic Regularities in Continuous Space Word Representations (Mikolov et al., NAACL 2013)

Bigger Analogy Test Set (BATS)

- 20 morphological categories, 20 semantical categories
- 50 word pairs per category
- Allows multiple correct answers

Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. (Gladkova et al, NAACL 2016)

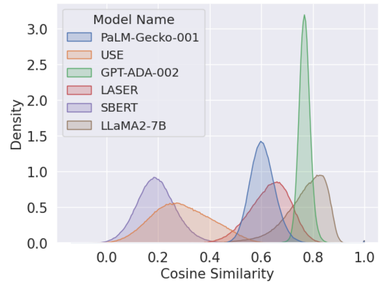
MODELS

(1) LLaMA2-7B (dim=4096), Meta AI
(2) ADA-002 (dim=1536), OpenAI
(3) PaLM2-Gecko-001 (dim=768), Google
(4) LASER (dim=1024), Meta AI
(5) Universal Sentence Encoder (dim=512)
(6) Sentence-BERT (dim=384)

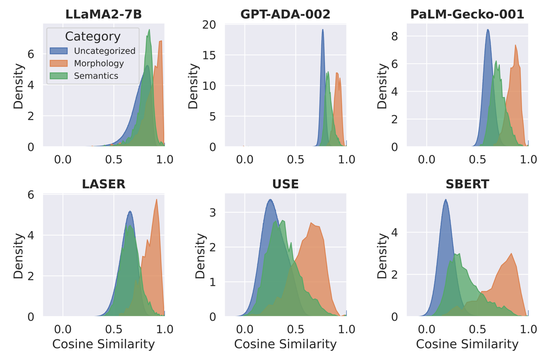
Google AI OpenAI Meta

ANALYSIS 1: WORD-PAIR SIMILARITY

- ≈ 80,000 distinct words sampled from WordNet.
- Cosine simiarity of all pairs (≈ 6.4 billion) computed for all models.
- Distribution of the cosine similarities for each model is shown right.
- ADA and LLaMA2 yield higher expected similarity for random pairs of words than other models.



- The Bigger Analogy Test Set (BATS) provides related word-pairs.
  - *Morphologically Related Pairs*
  - *Semantically Related Pairs*
- *Uncategorized Pairs:* Random word pairs from the corpus.
- The distrobution of cossims between pairs in each category are shown for all six embeddings.



- ADA and PaLM can effectively distiguish semantically related pairs from unrelated pairs, but so can SBERT.

ANALYSIS 2: WORD ANALOGY TASK

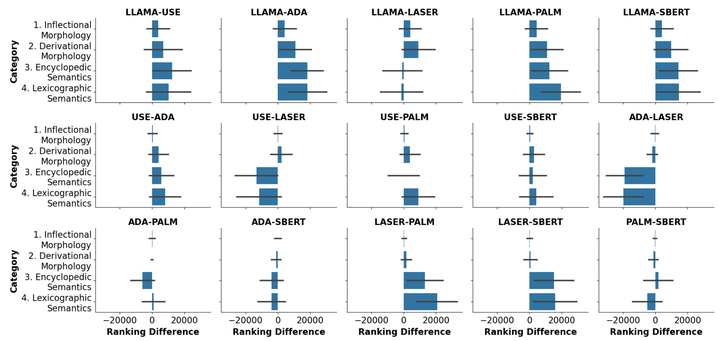
- The word analogy task was evaluated for each model using the BATS wordpairs.
- Methods tested: *3CosAdd*, *Pair Distance*, *3CosMul*, *3CosAvg*, *LRCos*
- Uniform Corpus for each model; Top-1 accuracy measured.
- ADA and PaLM performed very well with 3Cos style methods.
- LLaMA performed worst among LLMs.
- SBERT performed quite well, often ranked as the third best.

Method	3CosAdd	3CosAvg	3CosMul	LRCos	PairD
SBERT	0.243	0.261	0.267	**0.487**	0.086
USE	0.174	0.212	0.187	0.450	0.025
LASER	0.227	0.260	0.237	0.284	0.121
ADA-002	**0.412**	**0.447**	**0.424**	0.375	**0.232**
LLaMA2	0.145	0.200	0.145	0.131	0.053
PaLM 2	**0.398**	**0.458**	**0.417**	**0.534**	**0.193**

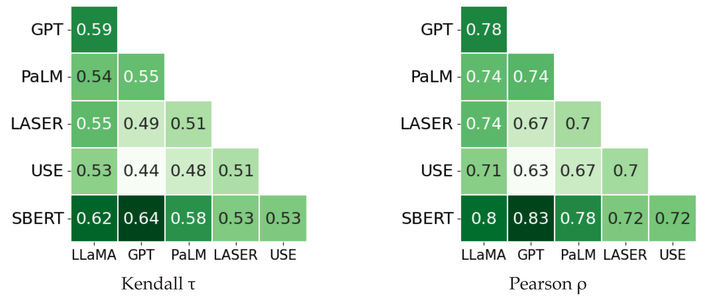
Table 1: Performance on BATS Analogy. Blue denotes the best accuracy; bold black denotes the second best.

DO LLMS OFFER SOMETHING NEW?

- We investigated the inter-model agreement on the similarity of related word pairs.
- The difference in rank between related words is calculated for each pair of models.
  - If two models agree, this value should have a mean of 0 and a small variance.



- Direct statistical measures of correlation can be used for a more robust evaluation.
- Kendall's tau and Pearson's rho were computed in terms of all word-pair similarities between each pair of models.



- Two of the LLMs, PaLM and ADA tended to agree with each other and with SBERT.

LIMITATIONS

- Only six models were analyzed; additioan work needs to be done to draw general conclusions about differences between LLMs and Classical Model embeddings.
- Existing works has illuminated issues in the word analogy task for evaluating word embedding quality.
  - We avoid maing claims implying one embedding to be 'better' than another.
- We rely on cosine similarity to compare vectors, and recent work has questioned the widespread use of the method.
  - Cosine Similarity is still the most popular metric in NLP literature.