How Contextual are Contextualized Word Representations?

Kawin Ethayarajh

Stanford University

EMNLP 2019

Background

A brief history of word representations:

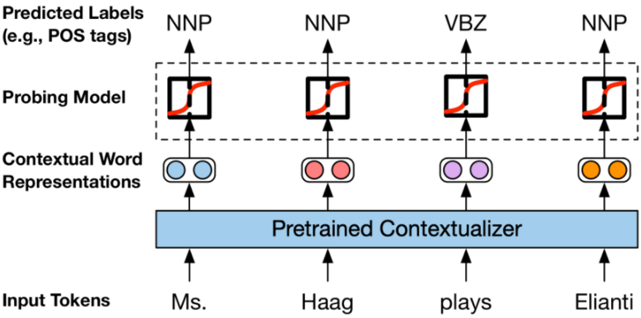
- pre-2018: static (skipgram, GloVe, etc.)
- post-2018: contextualized (ELMo, BERT, etc.)

On virtually every NLP task,

$$\textit{contextualized} \gg \textit{static}$$

Background

Training a linear probe on top of BERT's contextualized representations can achieve near-SOTA on many tasks. (Liu et al., 2019)



(Liu et al., 2019)

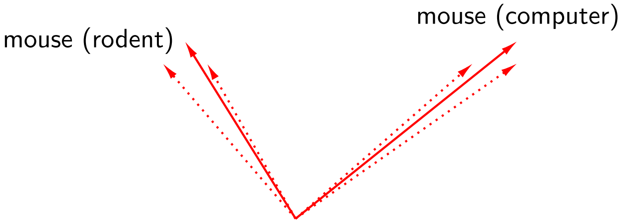
Questions

But *just how contextual* are these contextualized representations?

Questions

But *just how contextual* are these contextualized representations?

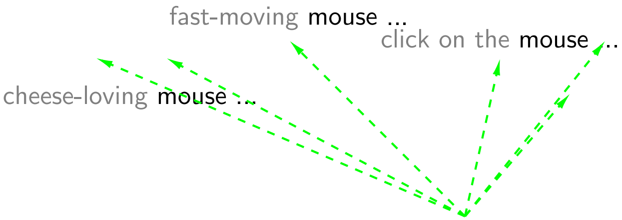
1. Are words essentially given one of a finite set of word-sense vectors?



Questions

But *just how contextual* are these contextualized representations?

1. Are words essentially given one of a finite set of word-sense vectors?
2. Or are there infinitely many context-specific representations?



Questions

More specifically,

1. How do representations of the same word differ across contexts?
2. Do words in the same context have more similar representations?
3. How well can static embeddings replace contextualized ones?

Measures of Contextuality

Consider sentences from SemEval STS data:

- *A panda dog is running on the road.*
- *A dog is trying to get bacon off its back.*

Measures of Contextuality

Consider sentences from SemEval STS data:

- *A panda dog is running on the road.*
- *A dog is trying to get bacon off its back.*

$$\vec{dog} = \vec{dog} \implies \text{no contextualization}$$

Measures of Contextuality

Consider sentences from SemEval STS data:

- *A panda dog is running on the road.*
- *A dog is trying to get bacon off its back.*

$$\vec{dog} = \vec{dog} \implies \text{no contextualization}$$

$$\vec{dog} \neq \vec{dog} \implies \textit{some} \text{ contextualization}$$

Measures of Contextuality

How can we *quantify* contextuality?

Measures of Contextuality

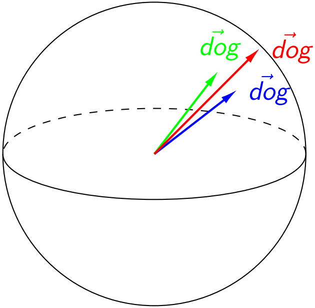
1. self-similarity (SelfSim)

   Average cosine similarity of a word with itself across all contexts, where representations are drawn from the same layer of a given model.

Measures of Contextuality

1. self-similarity (SelfSim)

e.g., high self-similarity for 'dog' across contexts



Measures of Contextuality

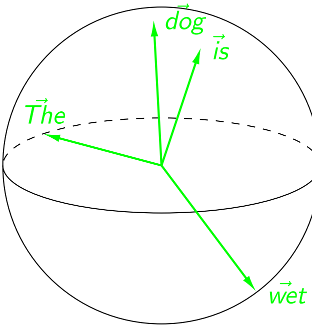
1. self-similarity
2. intra-sentence similarity (IntraSim)

   Average cosine similarity between a word and its context, where the context is represented as the average of its word representations.

Measures of Contextuality

1. self-similarity
2. intra-sentence similarity (IntraSim)

e.g., low intra-sentence similarity for ‘The dog is wet.’



Measures of Contextuality

1. self-similarity
2. intra-sentence similarity
3. maximum explainable variance (MEV)

   The variance explained by the first principal component of a word's representations across different contexts.

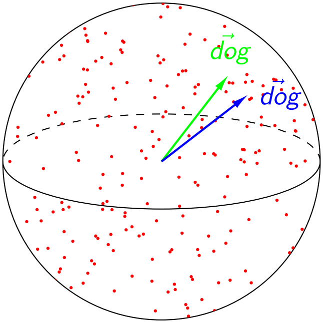
Measures of Contextuality

Generally speaking, we would expect:

1. lower self-similarity
2. higher intra-sentence similarity $\implies$ **MORE** context-specific
3. lower maximum explainable variance

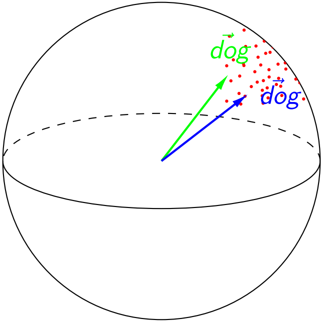
Adjusting for Anisotropy

$SelfSim_\ell(w) = 0.95$ is relatively high if all embeddings are isotropic ...



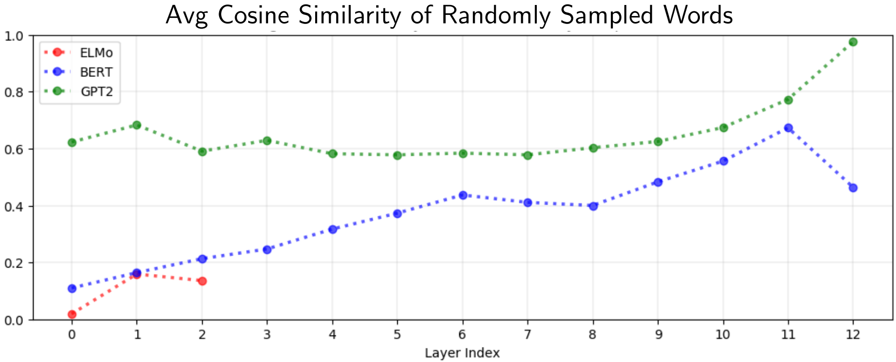
Adjusting for Anisotropy

$SelfSim_{\ell}(w) = 0.95$ is relatively high if all embeddings are isotropic but relatively low if they are anisotropic:



Adjusting for Anisotropy

Do we need to adjust for anisotropy? **Yes!** We find that high anisotropy is inherent to (or at least a by-product of) contextualization.



Adjusting for Anisotropy

We subtract these layer-specific baselines – which are zero for perfectly isotropic vectors – to get the *anisotropy-adjusted measures*:

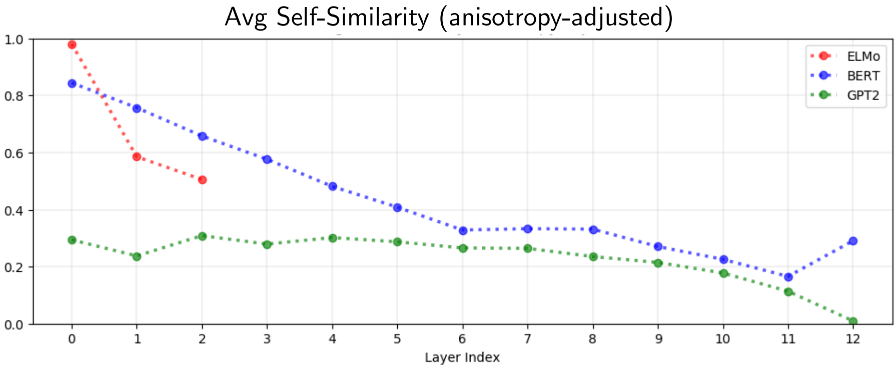
- average similarity of randomly sampled words (for SelfSim, IntraSim)
- variance explained by first PC of randomly sampled words (for MEV)

Questions

Back to our questions:

1. How do representations of the same word differ across contexts?
2. Do words in the same context have more similar representations?
3. How well can static embeddings replace contextualized ones?

On average, contextualized representations are more context-specific in higher layers. The decrease in self-similarity is almost monotonic.



Self-Similarity

Stopwords (e.g., 'the', 'of') have among the lowest self-similarity (i.e., the most context-specific representations).

- variety of contexts, rather than inherent polysemy, drives variation
- suggests words are not essentially being assigned a word-sense vector

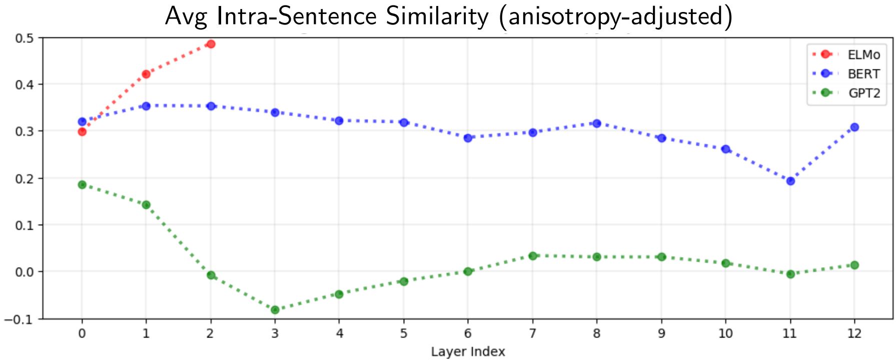
Questions

Back to our questions:

1. How do representations of the same word differ across contexts?
2. Do words in the same context have more similar representations?
3. How well can static embeddings replace contextualized ones?

Intra-Sentence Similarity

Context-specificity manifests differently in ELMo, BERT, and GPT-2, both across models and across different layers of the same model.



Intra-Sentence Similarity

Implications:

- BERT's contextualization is more nuanced than ELMo's; two words sharing the same context do not necessarily have a similar meaning.
- Unlike anisotropy, a high intra-sentence similarity is not inherent to contextualization.

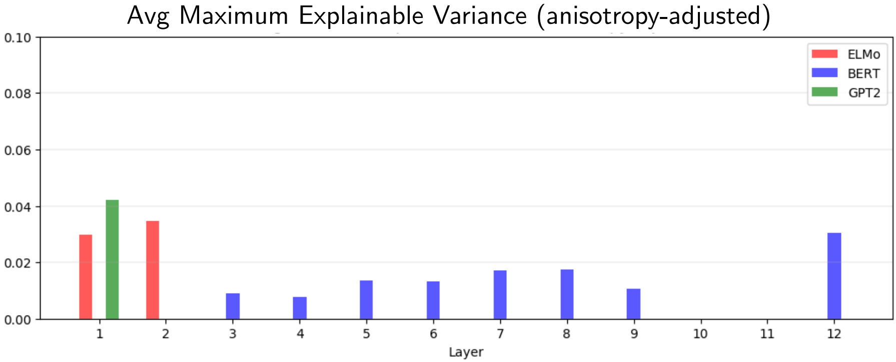
Questions

Back to our questions:

1. How do representations of the same word differ across contexts?
2. Do words in the same context have more similar representations?
3. How well can static embeddings replace contextualized ones?

Maximum Explainable Variance

On average, less than 5% of the variance in a word's contextualized representations can be explained by a static embedding.



Maximum Explainable Variance

The 5% threshold represents the best-case scenario:

- no guarantee that word2vec, for example, would maximize MEV
- low MEV is contrary to the idea of model assigning word-sense vectors

A New Type of Static Embedding

What if we created a static embedding for each word by taking the first principal component of its contextualized representations?

A New Type of Static Embedding

Principal components of contextualized representations in lower layers of BERT outperform GloVe and FastText on static embedding benchmarks.

	SimLex999	MEN	WS353	RW	Google	MSR	SemEval2012
GloVe	0.194	0.216	0.339	0.127	0.189	0.312	0.097
FastText	0.239	**0.239**	**0.432**	0.176	0.203	0.289	0.104
ELMo, Layer 1	0.276	0.167	0.317	0.148	0.170	0.326	0.114
ELMo, Layer 2	0.215	0.151	0.272	0.133	0.130	0.268	0.132
BERT, Layer 1	0.315	0.200	0.394	**0.208**	**0.236**	**0.389**	**0.166**
BERT, Layer 2	**0.320**	0.166	0.383	0.188	0.230	0.385	0.149
BERT, Layer 11	0.221	0.076	0.319	0.135	0.175	0.290	0.149
BERT, Layer 12	0.233	0.082	0.325	0.144	0.184	0.307	0.144

Future Work

Why did we use cosine similarity to measure embedding similarity?

- precedence
- transparency
- straightforward comparison across different layers and models

Future Work

1. What if we measured similarity using other metrics?
2. What if we tried forcing contextualized representations to be more isotropic? (e.g., Mu et al. (2018))

Conclusion

Takeaways:

- For ELMo, BERT, and GPT-2, upper layers produce more context-specific and anisotropic representations.
- However, context-specificity manifests very differently across models, particularly w.r.t. intra-sentence similarity.
- On average, less than 5% of the variance in a word's contextualized representations can be explained by a static embedding.

References

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual representations. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

Jiaqi Mu, Suma Bhat, and Pramod Viswanath. 2018. All-but-the-top: Simple and effective postprocessing for word representations. In Proceedings of the 7th International Conference on Learning Representations (ICLR).