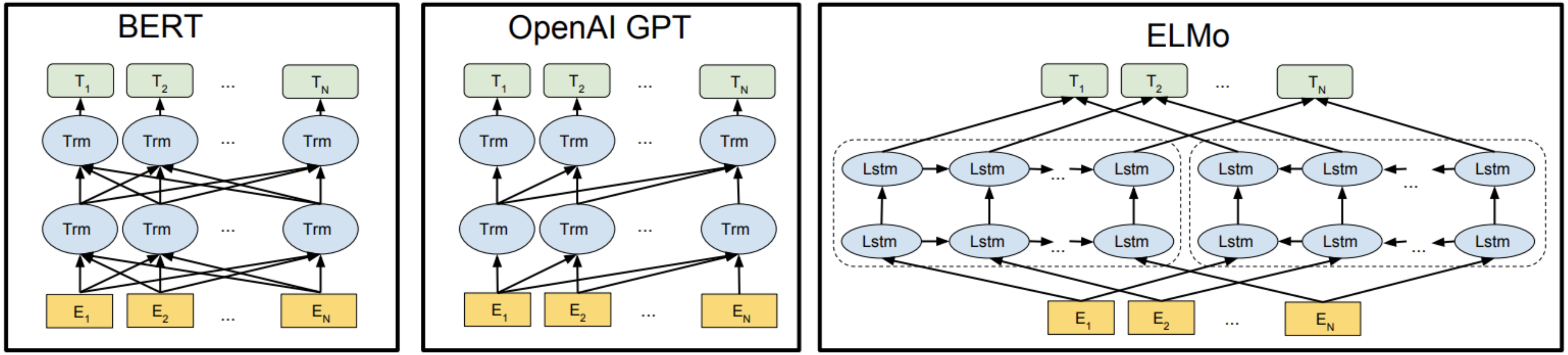
UCLA

Retrofitting Contextualized Word Embeddings with Paraphrases

Weijia Shi[1*], ***Muhao Chen***[1*], Pei Zhou[2], Kai-Wei Chang[1]
[1]University of California, Los Angeles
[2]University of Southern California

Contextualized Word Embeddings

Representations that considers the **difference of lexical semantics** under **different linguistic contexts**



Such representations have become the backbone of many StoA NLU systems for

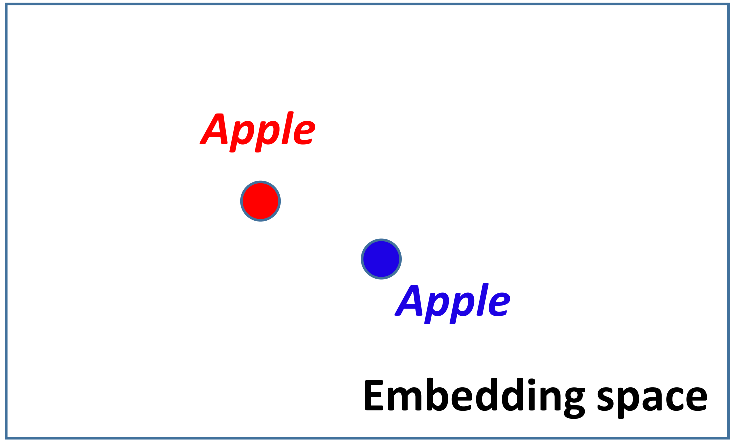
- Sentence classification, textual inference, QA, EDL, NMT, SRL, …

Contextualized Word Embeddings

Aggregating context information in a word vector with a pre-trained **deep neural language model**.

Key benefits:

- More refined semantic representations of lexemes
- Automatically capturing polysemy
    - *Apples* **have been grown for thousands of years in Asia and Europe**.



- **With that market capacity,** ***Apple*** **is worth over 1% of the world's GDP.**

The Paraphrased Context Problem

The pre-trained language models are not aware of the semantic relatedness of contexts

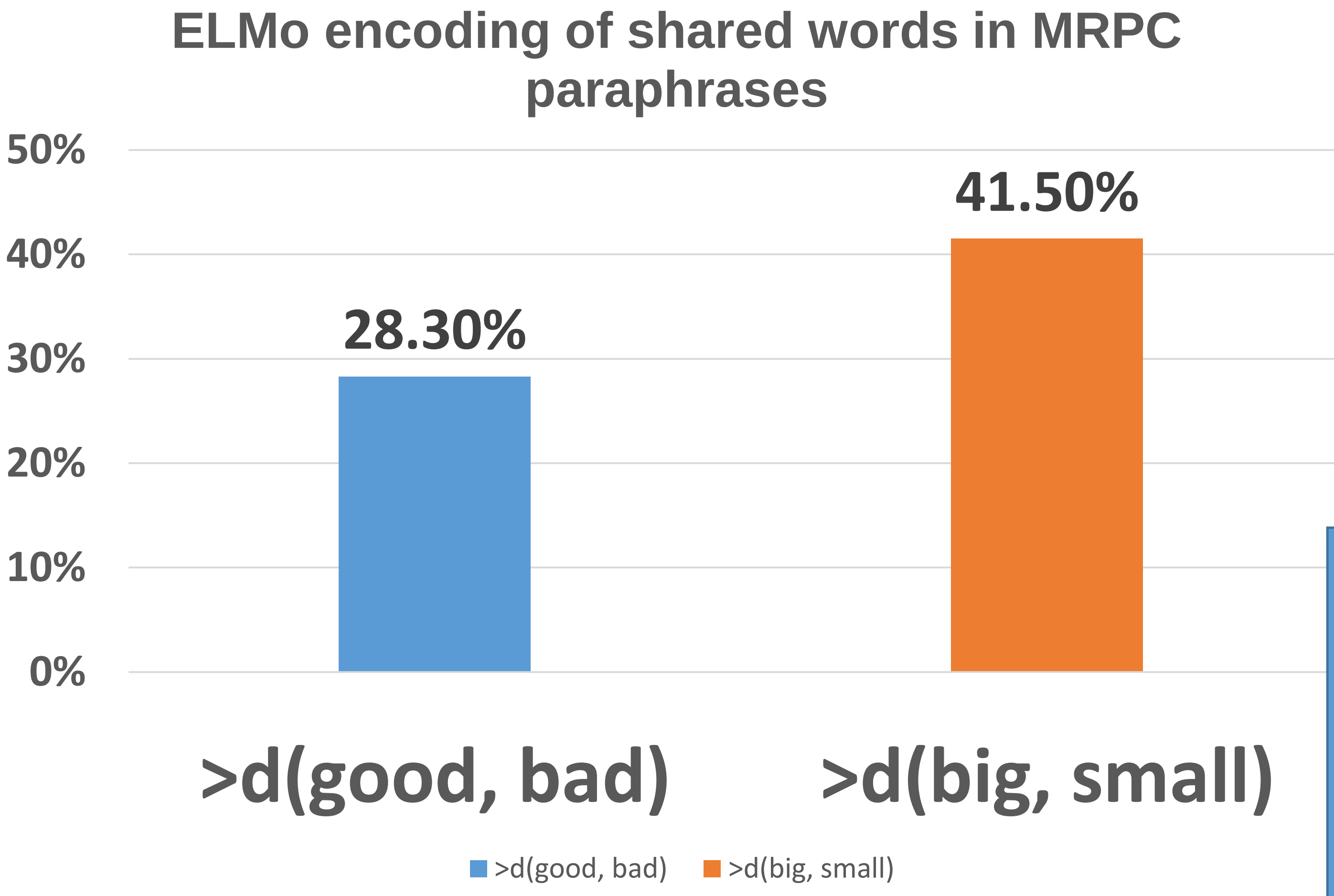
The same word can be represented more differently than opposite words in unrelated contexts

	Contexts	L2 distance by ELMo
Paraphrases	How can I make **bigger** my arms? How do I make my arms **bigger?**	6.42
Paraphrases	Some people believe earth is **flat**, why? Why do people still believe in **flat** earth?	7.59
	It is a very **small** window. I have a **large** suitcase.	5.44

The Paraphrased Context Problem

UCLA

Consider ELMo distances of the same words (**excluding stop words**) in paraphrased sentence pairs from MRPC:



Contextualization:

- **Can be oversensitive to paraphrasing,**
- **and further impair sentence representations.**

Outline

- Background
- Paraphrase-aware retrofitting
- Evaluation
- Future Work

Paraphrase-aware Retrofitting (PAR)

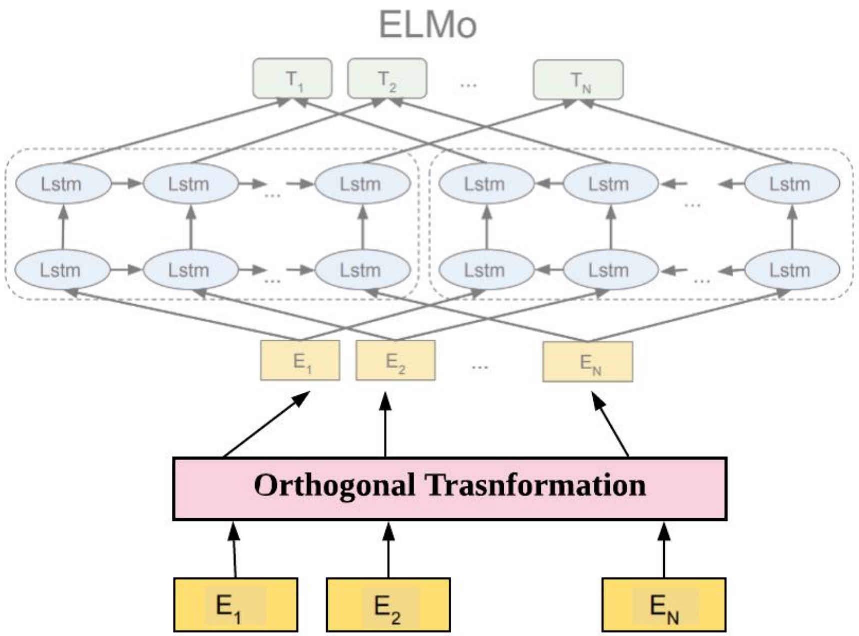
Method

- An orthogonal transformation **M** to retrofit the input space
- Minimizing the variance of word representations on paraphrased contexts
- Without compromising the varying representations on unrelated contexts

Orthogonal constraint:

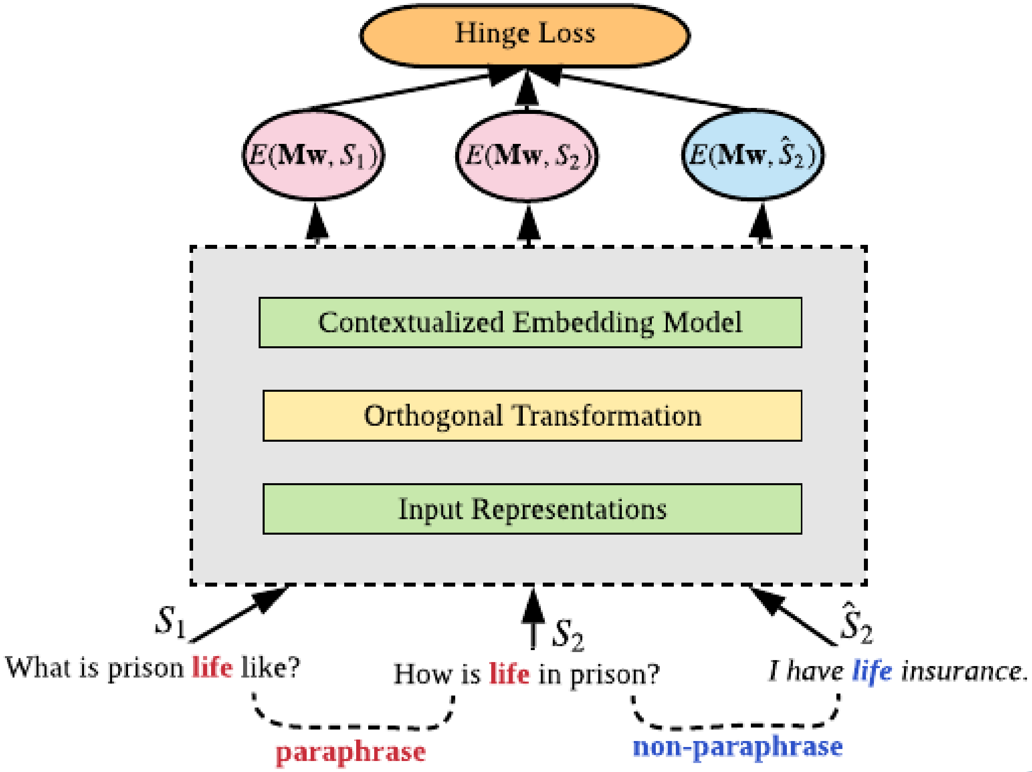
$$L_O = \| \mathbf{I} - \mathbf{M}^\top \mathbf{M} \|_{F},$$

Keeping the relative distance of raw embeddings before contextualization



Paraphrase-aware Retrofitting (PAR)

Learning objective



Input:

Paraphrase 1: What is prison **life** like?

Paraphrase 2: How is **life** in prison?

Negative sample: *I have life insurance.*

Loss Function:

Orthogonal constraint

$$L = \sum_{(S_1,S_2)\in P} \sum_{w \in S_1 \cap S_2} \left[d_{s_1,s_2}(\mathbf{Mw}) + \gamma - d_{\widehat{S_1},\widehat{S_2}}(\mathbf{Mw})\right]_+ + \boxed{\lambda L_O}$$

$$d_{S_1,S_2}(\mathbf{w}) = \left\| E(\mathbf{w}, S_1) - E(\mathbf{w}, S_2) \right\|_2 .$$

Intuition: the shared words in paraphrases should be embedded closer than those in non-paraphrases.

UCLA

Experiment Settings

Paraphrase pair datasets

- The positive training cases of MRPC (2,753 pairs)
- Sampled Quora (20,000 pairs) and PAN (5,000 pairs)

Tasks

- **Sentence classification**: MPQA, MR, CR, SST-2
- **Textual inference**: MRPC, SICK-E
- **Sentence relatedness scoring**: SICK-R, STS-15, STS-16, STS-Benchmark
- **Adversarial SQuAD**

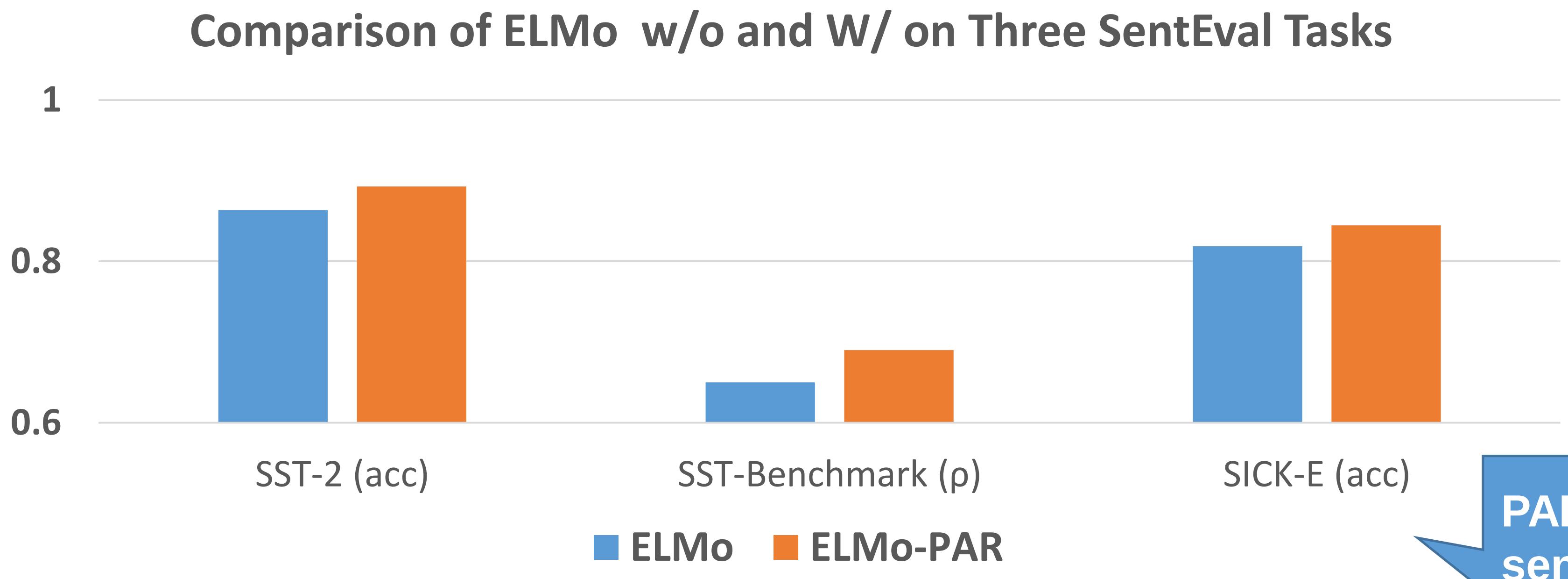
* The first three categories of tasks follow the settings in **SentEval** [Conneau et al, 2018].

Text Classification/Inference/Relatedness Tasks

UCLA

PAR leads to performance improvement of ELMo by

- **2.59-4.21%** in accuracy on sentence classification tasks
- **2.60-3.30%** in accuracy on textual inference tasks
- **3-5%** in Pearson correlation in text similarity tasks



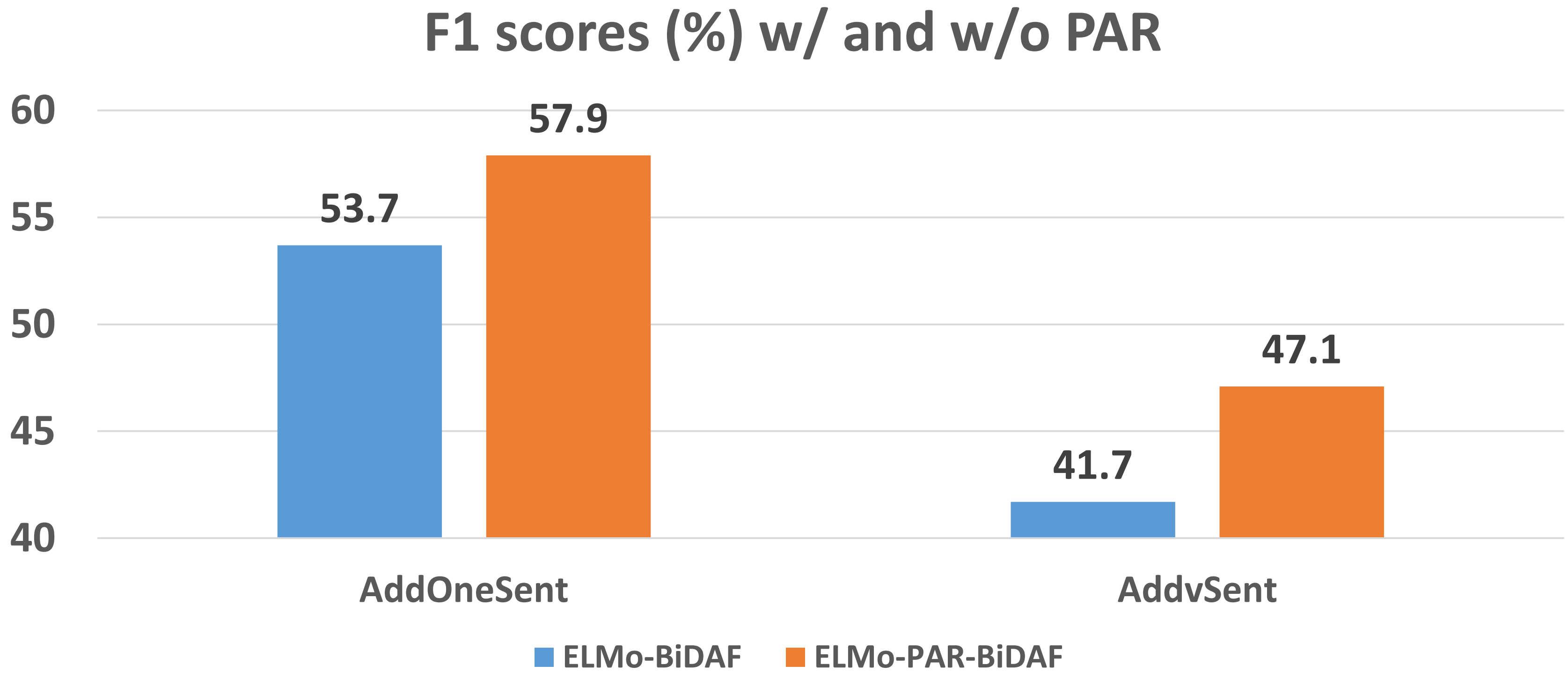
PAR improves ELMo on sentence representation tasks.

UCLA

Adversarial SQuAD

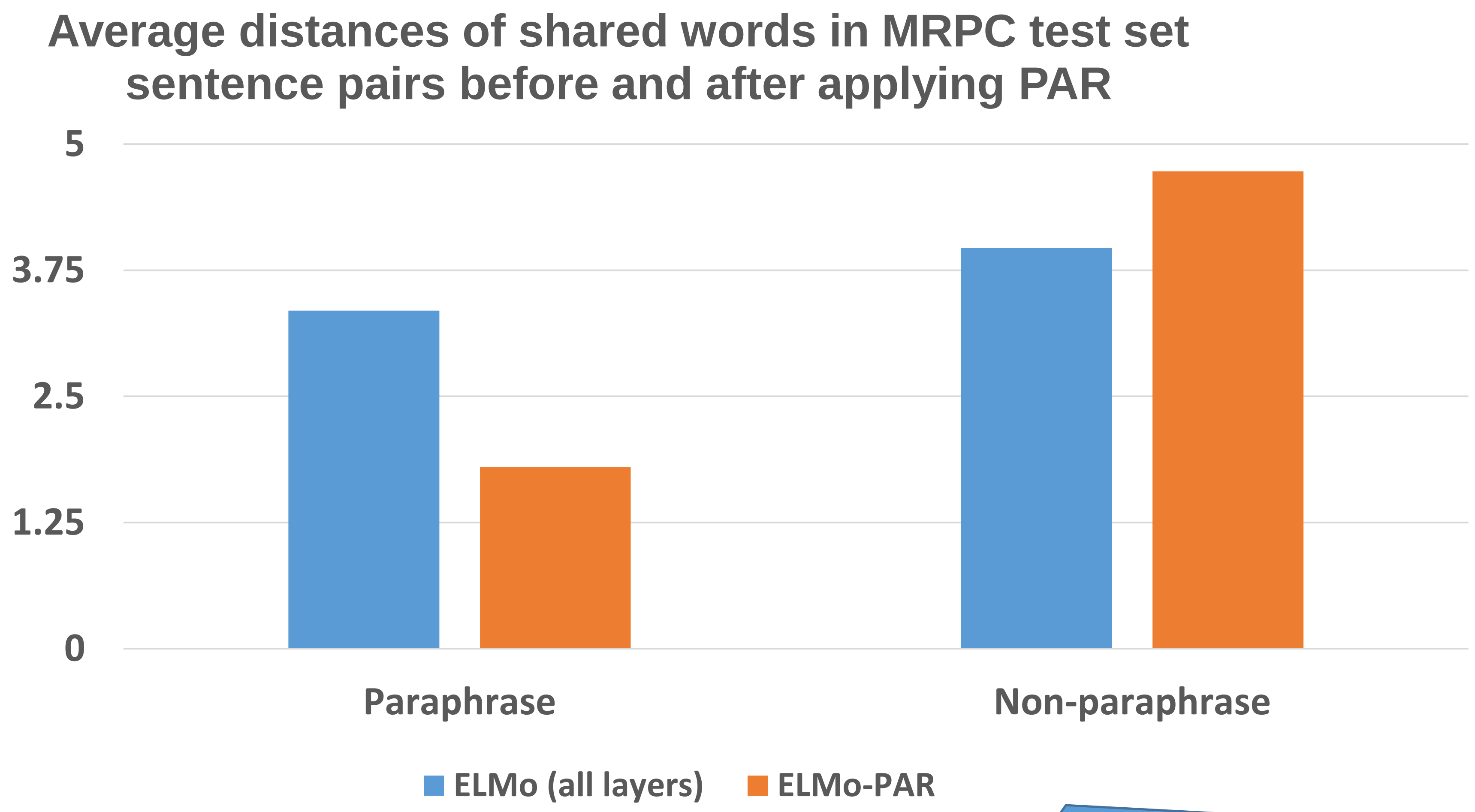
Bi-Directional Attention Flow (BiDAF) [Seo et al. 2017] on two challenge settings

- AddOneSent: add one human-paraphrased sentence
- AddvSent: add one adversarial example sentence that is semantically similar to the question



PAR improves the robustness of a downstream QA model against adversarial examples.

Word Representations



PAR minimizes the differences of a word's representations in paraphrased contexts and preserves the differences in non-paraphrased contexts.

Future Work

Applying PAR on other contextualized embedding models

To modify contextualized word embeddings linguistic knowledge

- Context simplicity aware embeddings
- Incorporating lexical definitions in the word contextualization process

Thank You