Retrofitting Contextualized Word Embeddings with Paraphrases

Weijia Shi\textsuperscript{1*}, *Muhao Chen*\textsuperscript{1*}, Pei Zhou\textsuperscript{2}, Kai-Wei Chang\textsuperscript{1}

\textsuperscript{1}University of California, Los Angeles
\textsuperscript{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

<table>
<thead>
<tr>
<th>Contexts</th>
<th>L2 distance by ELMo</th>
</tr>
</thead>
<tbody>
<tr>
<td>How can I make <strong>bigger</strong> my arms?</td>
<td>6.42</td>
</tr>
<tr>
<td>How do I make my arms <strong>bigger</strong>?</td>
<td></td>
</tr>
<tr>
<td>Some people believe earth is <strong>flat</strong>, why?</td>
<td>7.59</td>
</tr>
<tr>
<td>Why do people still believe in <strong>flat</strong> earth?</td>
<td></td>
</tr>
<tr>
<td>It is a very <strong>small</strong> window.</td>
<td>5.44</td>
</tr>
<tr>
<td>I have a <strong>large</strong> suitcase.</td>
<td></td>
</tr>
</tbody>
</table>
The Paraphrased Context Problem

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:

$$L_O = \left\| I - M^T M \right\|_F,$$

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:

\[
L = \sum_{(S_1, S_2) \in P} \sum_{w \in S_1 \cap S_2} \left[ d_{S_1, S_2}(Mw) + \gamma - d_{\hat{S}_1, \hat{S}_2}(Mw) \right] + \lambda L_0
\]

\[
d_{S_1, S_2}(w) = \| E(w, S_1) - E(w, S_2) \|_2.
\]

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

Orthogonal constraint
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].
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

Comparison of ELMo w/o and W/ on Three SentEval Tasks

PAR improves ELMo on sentence representation tasks.
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

<table>
<thead>
<tr>
<th>Setting</th>
<th>ELMo-BiDAF</th>
<th>ELMo-PAR-BiDAF</th>
</tr>
</thead>
<tbody>
<tr>
<td>AddOneSent</td>
<td>53.7</td>
<td>57.9</td>
</tr>
<tr>
<td>AddvSent</td>
<td>41.7</td>
<td>47.1</td>
</tr>
</tbody>
</table>

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

Average distances of shared words in MRPC test set sentence pairs before and after applying PAR

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