Knowledge Base - Based Language Model Pre-training
Xuan Zhang¹, Kevin Duh¹, Hao Cheng², Hoifung Poon², Xiaodong Liu²
Oct 2, 2020

¹ Johns Hopkins University
² Microsoft Research
Figure 1. Overall pre-training and fine-tuning procedures for BERT.

* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Devlin et al.
Domain-Specific Pre-training - Biomedicine

Figure 2. Domain-specific pretraining from scratch for biomedicine.*

*Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing, Gu et al.
Masked LM

Sentence: In the present study, we provide first evidence that agrin is absent from basal lamina of tumor vessels if the TJ molecules occluding, claudin-5 and claudin-1 were lacking in the endothelial cells.

BERT: In the [MASK] study, we provide [MASK] evidence that agrin is [MASK] from basal lamina of tumor [MASK] if the TJ molecules [MASK], claudin-5 and claudin-1 were [MASK] in the endothelial cells.
In the present study, we provide first evidence that agrin is absent from basal lamina of tumor vessels if the TJ molecules occluding, claudin-5 and claudin-1 were lacking in the endothelial cells.

BERT: In the [MASK] study, we provide [MASK] evidence that agrin is [MASK] from basal lamina of tumor [MASK] if the TJ molecules [MASK], claudin-5 and claudin-1 were [MASK] in the endothelial cells.

Entity-Level Masking:
In the [MASK] study, we provide first evidence that [MASK] is absent from basal lamina of tumor [MASK] if the TJ molecules [MASK], [MASK] and [MASK] were lacking in the endothelial cells.
Sentence: In the present study, we provide first evidence that agrin is absent from basal lamina of tumor vessels if the TJ molecules occluding, claudin-5 and claudin-1 were lacking in the endothelial cells.

BERT: In the [MASK] study, we provide [MASK] evidence that agrin is [MASK] from basal lamina of tumor [MASK] if the TJ molecules [MASK], claudin-5 and claudin-1 were [MASK] in the endothelial cells.

Bigram Masking (consecutive words that frequently co-occur):

In the [MASK] [MASK], we provide first evidence that agrin is absent from basal lamina of [MASK] [MASK] if the TJ molecules occluding, claudin-5 and claudin-1 were lacking in the [MASK] [MASK].
Distant Pair Masking (this work)

Sentence: In the present study, we provide first evidence that agrin is absent from basal lamina of tumor vessels if the TJ molecules occluding, claudin-5 and claudin-1 were lacking in the endothelial cells.

BERT: In the [MASK] study, we provide [MASK] evidence that agrin is [MASK] from basal lamina of tumor [MASK] if the TJ molecules [MASK], claudin-5 and claudin-1 were [MASK] in the endothelial cells.

Pair Masking (bigram/distant pairs that frequently co-occur, high pmi score): In the [MASK] [MASK], we provide first evidence that [MASK] is absent from basal lamina of tumor vessels if the TJ molecules [MASK], [MASK] and [MASK] were lacking in the endothelial cells.
Dataset

- **Pre-training:** Biomedical abstracts
  
  #sentences: 171million  
  average length: 22  
  22% sentences contain entities
  
  54 million entity appearances

- **Fine-tuning:**
  
  Named entity recognition: 12k samples

  Relation extraction: 18k samples
Result 1. Train from scratch
Result 2. Continued-Training

NER BC2GM

RE Chemprot
Result 3. Pair Masking

NER BC2GM

RE Chemprot
## Results

<table>
<thead>
<tr>
<th>Method</th>
<th>NER</th>
<th>RE</th>
</tr>
</thead>
<tbody>
<tr>
<td>scratch_xdl</td>
<td>87.5</td>
<td>77.30</td>
</tr>
<tr>
<td>scratch_entity0.33</td>
<td>87.56 (+0.06)</td>
<td>77.75 (+0.45)</td>
</tr>
<tr>
<td>ct_entity_random</td>
<td>87.64 (+0.14)</td>
<td>77.34 (+0.04)</td>
</tr>
<tr>
<td>ct_entity0.33</td>
<td>87.81 (+0.31)</td>
<td>77.33 (+0.03)</td>
</tr>
<tr>
<td>ct_entity1</td>
<td><strong>87.91 (+0.41)</strong></td>
<td>77.29 (-0.01)</td>
</tr>
<tr>
<td>ct_entity0.33_entity0.5</td>
<td>87.70 (+0.20)</td>
<td>77.33 (+0.03)</td>
</tr>
<tr>
<td>ct_entity0.33_entity1</td>
<td>87.80 (+0.30)</td>
<td>77.51 (+0.21)</td>
</tr>
<tr>
<td>ct_bigram0.33</td>
<td>87.56 (+0.06)</td>
<td>76.94 (-0.26)</td>
</tr>
<tr>
<td>ct_bigram0.5</td>
<td>87.54 (+0.04)</td>
<td>76.97 (-0.23)</td>
</tr>
<tr>
<td>ct_bigram1</td>
<td>87.31 (-0.19)</td>
<td>77.67 (+0.37)</td>
</tr>
<tr>
<td>ct_pmi0.5</td>
<td>87.62 (+0.12)</td>
<td>76.79 (-0.51)</td>
</tr>
</tbody>
</table>