**Motivation**

How can we diagnose strengths and weaknesses of transformer-based language models beyond traditional accuracy metrics?

We extract snapshots of acquired knowledge at sequential stages of the training process.

**Linguistic Metrics**

Part-of-Speech Overprediction Rate

$$\text{POSOR}(\text{pos}) = \frac{(LM_{\text{pos}} - GLM_{\text{pos}}) \cdot 100}{GLM_{\text{pos}}}$$

POSOR allows us to identify part-of-speech deficiencies of our language models. This metric is framed within the context of probing task literature.

**Probing Tasks**

- wh-words
- prepositions
- coordinate conjunctions
- negation
- coordination
- EOS
- spatial

**Knowledge Graph Metrics**

**Graph-Edit-Distance**

$$GED(g_1, g_2) = \min_{(e_1, \ldots, e_k) \in \mathcal{E}(g_1, g_2)} \sum_{i=1}^{k} c(e_i)$$

These metrics are inspired from graph literature to quantitatively compare KG extracts.

**Knowledge Graph Results**

<table>
<thead>
<tr>
<th>Target Model (distance from RoBERTa)</th>
<th>Graph-Edit-Distance on the extracted knowledge graph</th>
<th>Euclidean distance on the graph2vec embeddings</th>
</tr>
</thead>
<tbody>
<tr>
<td>RoBERTa 1e</td>
<td>141.25</td>
<td>0.2260</td>
</tr>
<tr>
<td>RoBERTa 3e</td>
<td>135.00</td>
<td>0.1733</td>
</tr>
<tr>
<td>RoBERTa 5e</td>
<td>130.50</td>
<td>0.1607</td>
</tr>
<tr>
<td>RoBERTa 7e</td>
<td>121.50</td>
<td>0.1605</td>
</tr>
<tr>
<td>DistilBERT</td>
<td>28.50</td>
<td>0.0284</td>
</tr>
<tr>
<td>BERT</td>
<td>16.50</td>
<td>0.0202</td>
</tr>
</tbody>
</table>

Across both quantitative graph metrics, we see the distance from each model to pretrained RoBERTa reduce as the number of epochs and the amount of training data increase.

**Research Questions**

1. Quantitatively compare knowledge acquisition across language models
2. Analyze the same model at different stages over time (early training)
3. Compare knowledge graphs linguistically

**Datasets**

- SQuAD
- Google-RE

Datasets are inspired from Facebook’s LAMA paper