A Systematic Search for Compound Semantics in Pretrained BERT Architectures

Filip Miletic & Sabine Schulte im Walde

Institute for Natural Language Processing, University of Stuttgart, Germany
(filip.miletic, schulte)@ims.uni-stuttgart.de

Introduction
Motivation
Modeling the meanings of noun compounds is challenging because they vary in terms of their degree of compositionality.

<table>
<thead>
<tr>
<th>guinea pig</th>
<th>snail mail</th>
<th>health insurance</th>
</tr>
</thead>
<tbody>
<tr>
<td>more compositional</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

When predicting the degree of compositionality of noun compounds, static word embeddings still outperform transformer-based models (Cordeiro et al., 2019; Garcia et al., 2021).

Research questions
- Can BERT information be used more efficiently?
- How does BERT represent compound semantics?
- What is the impact of compound properties?

Experimental setup
Data and model
- 280 English compounds with human compositionality ratings (Cordeiro et al., 2019)
- Corpus data: ENCow16 (Schäfer and Bildhauer, 2012; Schäfer, 2015)
- BERT-base-uncased, no fine-tuning (Devlin et al., 2019)

Modelling approach
For each compound:
- Take a sample of corpus occurrences
- Feed each occurrence into BERT and retain all provided embeddings
- Use a subset of the embeddings to estimate the degree of compositionality

Experimental parameters
Preprocessing
- # sequences: 10, 100, 1k
- Seq. length: any, ≥ 20 tokens

Embedding computation
- Tokens: modifier, head, compound, context, CLS
- Layers: 0–12, all contiguous combinations
- Aggregation: token-level, type-level
- Pooling: avg, sum

Compositionality estimation
- Direct
- Composite: ADD, MULT, COMB

⇒ 41,496 parameter constellations

Results
Spearman’s rank correlation coefficient was used to evaluate the predicted vs. gold standard compositionality ratings.

Performance range
<table>
<thead>
<tr>
<th>ours</th>
<th>best</th>
<th>worst</th>
<th>(Cordeiro et al., 2019)</th>
<th>(Garcia et al., 2021)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.706</td>
<td>0.726</td>
<td>-0.649</td>
<td>0.706</td>
<td>0.611</td>
</tr>
</tbody>
</table>

Effects of modeled tokens
Best performance across prediction targets
- Prediction target: modelled token
- Modelled token: head, comp, cont, CLS

Effects of empirical properties of compounds
- We analyzed the frequency, productivity, and ambiguity of compound heads (cf. Schulte im Walde et al., 2016; Alipoor & Schulte im Walde, 2020)
- For each property, we created subsets (56 compounds each) corresponding to the low, mid, and high ranges
- Model performance was evaluated for each subset independently

Conclusions
- We obtained robust compositionality information from pretrained BERT, but only with a highly optimized experimental setup.
- Strong effects of retained representational information: e.g. preference for lower layers, contextual information.
- Better for heads with lower frequency, productivity, ambiguity.
- BERT appears to encode at least some aspects of compound semantics.

References