To Split or Not to Split: Composing Compounds in Contextual Vector Spaces

Motivation

- Contextual embeddings can represent more nuanced semantic and syntactic relationships than bag-of-words models.
- What makes a good (vectorized) representation of a compound?
  - By extension: what structure(s) should the overall vector space have?
  - Guak: models that can represent polysemy, composition
- How does sub-word tokenization in contextual embedding models like BERT affect the representation of compound semantics?
  - Is it detrimental that sub-word splits often do not correspond to morphological boundaries?

Setup

Compound Splitting (Preprocessing)

- The split configurations operate on a copy of the training data that has had the SimpleCompoundSplitter (Weller-Di Marco[2017]) run on it.
- It uses word frequencies and POS tags from the training corpus to inform its splits.

Tokenizers

- For two configurations (voc-rt-DTA and split), we re-allocate the (WordPiece) tokenizer’s vocabulary based on our training dataset.
  - Aligning the granularity of tokens and sub-tokens with in-domain data.
- When the tokenizer is changed, the model can’t benefit from any pre-training that was done with the original base tokenizer.

Training and Fine-Tuning

- BERT models
  - Run for 5 epochs, learning rate 5e-5, on a Nvidia GeForce RTX A6000 for ~54 hours.
  - Default settings from [Devlin et al.(2019)](Devlin, Chang, Lee, and Toutanova): e.g. 30k vocabulary
  - Maximum sequence length of 128 tokens

Vector Space Similarity and Compositionality Ratings

- Decontextualized vector representations for compounds and constituents: prompted with- out sentence context, representations for multiple tokens averaged.
  - Use of either the first, middle or last 4 layers.
- Correlation between

Data Description

Deutsches Textarchiv (DTA) (1814-1900 siles)
[Berlin-Brandenburgischen Akademie der Wissenschaften(2022)]
- Curated selection of German texts
- ≈4M sentences, ≈8M tokens; 10% held out as evaluation data

Compositionality Ratings of Noun Compounds
[Schulte im Walde et al.(2016)](Schulte im Walde, Hády, Bott, and Khlivařiová)
- Compositionality with respect to compound modifier or head (scale of 1-6) rated by experts and by crowd workers.
- 185 compounds occur ≥ 20 times in DTA training data, remainder excluded

German BERT
- Trained on Wikipedia, OpenLegalData, news (unknown time period)

In-Context Masked-Language-Model Task

- Fill [MASK] tokens for target compound nouns in eval data.
- Partial: match any token in top 10 predictions; Full: match all tokens (among top 10 predictions for each split).
- GermaNet path similarity: Top model predictions queried in GermaNet, and a path between that item and a Synset representing the target compound is searched for.
  - Only words that could be found in GermaNet have scores reported in Path Sim. The Precision measure shows the proportion of predicted words that did not return a result from GermaNet.

Table 3: MLM task evaluations over the four preprocessing / tokenizer configurations.

- split model outperforms or competes with the fine-tuned model (base-ft-DTA), without benefiting from pre-training.

Vector Space Similarity and Compositionality Ratings

- No significant correlations found for compound + modifier pairs.
- Correlation between

Table 4: Coarse similarity between BERT compound and constituent vectors versus compositionality ratings.

- The base models perform the best, with a split model showing a stronger correlation with human annotations than the fine-tuned configurations.

References

[Schulte im Walde et al.(2016)](Schulte im Walde, Hády, Bott, and Khlivařiová) run on it.
[Schulte im Walde et al.(2016)](Schulte im Walde et al., 2016).

<table>
<thead>
<tr>
<th>Component</th>
<th>Base Tokenizer</th>
<th>Pre-Trained Tokenizer</th>
<th>Split Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>murmeln + Tier</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Traum + Bild</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zitrone + Saft</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Geschmack + Sache</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Example noun compounds and their tokenizations.

Our code is available at https://gitlab.com/cjenk/representations-composition