Universität Stuttgart

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

Chris Jenkins
Filip Miletić
Sabine Schulte im Walde

IMS

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?
  - Goals: 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?

Compound	**Base Tokenizer**	**Re-Trained Tokenizer**	**Split Tokens**
Geschmackssache (matter of taste) 😗👌	Geschmack + ##ss + ##ache	Geschmack + ##ssache	Geschmack + Sache
Zitronensaft (lemon juice) 🍋🧃	Zit + ##ronen + ##sa + ##ft	Zitrone + ##ns + ##aft	Zitrone + Saft
Murmeltier (marmot) 🦫	Mur + ##mel + ##ti + ##er	Murm + ##eltier	murmeln + Tier
Schauspiel (play [theater]) 🎭	Schauspiel	Schauspiel	schauen + Spiel
Klavierspiel (piano music) 🎹	Klavier + ##spiel	Klavier + ##spiel	Klavier + Spiel
Traumbild (vision [imagined]) 😴🖼️	Traum + ##bild	Traum + ##bild	Traum + Bild

Table 1: *Example noun compounds and their tokenizations.*

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 (Word-Piece) 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

Configuration	Pre-Train	DTA	Re-Train	Split
base	☑	☒	☒	☒
base-ft-DTA	☑	☑	☒	☒
voc-rt-DTA	☒	☑	☑	☒
split	☒	☑	☑	☑

Table 2: *BERT model configurations. ☑: presence, ☒: absence.*

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

Data Description

Deutsches Textarchiv (DTA) (1814-1900 slice)
[Berlin-Brandenburgischen Akademie der Wissenschaften(2022)]

- Curated selection of German texts
- ≈ 4M sentences, ≈89M tokens; 10% held out as evaluation data

Compositionality Ratings of Noun Compounds
[Schulte im Walde et al.(2016)Schulte im Walde, Hätty, Bott, and Khvtisavrishvili]

- Compositionality with respect to compound modifier or head (scale of 1-6) rated by experts and by crowd workers
- 185 compounds occur $\geq$ 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 slot)
- 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.

Configuration	Prediction		GermaNet	
	Partial	Full	Path Sim	Prec.
base	0.06	0.02	**0.37**	0.11
base-ft-DTA	0.23	**0.15**	**0.37**	0.24
voc-rt-DTA	0.10	0.07	0.35	0.40
split	**0.26**	0.11	0.36	**0.52**

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

- Decontextualized vector representations for compounds and constituents: prompted without sentence context, representations for multiple tokens averaged.
  - Use of either the first, middle or last 4 layers.
- Correlation between

$cos\_sim(\overrightarrow{\text{🍋🧃}}, \overrightarrow{\text{🧃}})$

Compositionality ratings:
1: 🧃 unrelated to 🍋🧃
6: 🧃 totally related to 🍋🧃
Annotator average: 5.75 (very related)

BERT Configuration	Layer	Constituent	ρ
base	last	head	0.368
split	first	head	0.313
base	first	head	0.288
base-ft-DTA	last	head	0.287
split	mid	head	0.282
base	mid	head	0.261
split	last	head	0.232

Table 4: *Cosine similarity between BERT compound and constituent vectors ∼ compositionality ratings.*

- No significant correlations found for compound + modifier pairs.
- The **base** models perform the best, with a **split** model showing a stronger correlation with human annotations than the fine-tuned configurations.

References

[Berlin-Brandenburgischen Akademie der Wissenschaften(2022)] Berlin-Brandenburgischen Akademie der Wissenschaften. Deutsches Textarchiv. Grundlage für ein Referenzkorpus der neuhochdeutschen Sprache. https://www.deutschestextarchiv.de/, 2022.

[Devlin et al.(2019)Devlin, Chang, Lee, and Toutanova] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. URL `https://aclanthology.org/N19-1423`.

[Hamp and Feldweg(1997)] Birgit Hamp and Helmut Feldweg. GermaNet - a lexical-semantic net for German. In *Proceedings of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, 1997.

[Henrich and Hinrichs(2010)] Verena Henrich and Erhard Hinrichs. GernEdiT - the GermaNet editing tool. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010. European Language Resources Association (ELRA). ISBN 2-9517408-6-7.

[Schulte im Walde et al.(2016)Schulte im Walde, Hätty, Bott, and Khvtisavrishvili] Sabine Schulte im Walde, Anna Hätty, Stefan Bott, and Nana Khvtisavrishvili. GhoSt-NN: A representative gold standard of German noun-noun compounds. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2285–2292, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL `https://aclanthology.org/L16-1362`.

[Weller-Di Marco(2017)] Marion Weller-Di Marco. Simple compound splitting for German. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 161–166, Valencia, Spain, April 2017. Association for Computational Linguistics. URL `https://aclanthology.org/W17-1722`.