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

Chris Jenkins and **Filip Miletić** and **Sabine Schulte im Walde**
Institute for Natural Language Processing, University of Stuttgart
{christopher.jenkins,filip.miletic,schulte}@ims.uni-stuttgart.de

Abstract

We investigate the effect of sub-word tokenization on representations of German noun compounds: single orthographic words which are composed of two or more constituents but often tokenized into units that are not morphologically motivated or meaningful. Using variants of BERT models and tokenization strategies on domain-specific restricted diachronic data, we introduce a suite of evaluations relying on the masked language modelling task and compositionality prediction. We obtain the most consistent improvements by pre-splitting compounds into constituents.

1 Introduction

Contextual word embedding models such as BERT (Devlin et al., 2019) have opened up exciting avenues in computational lexical semantics. But since their vocabulary size is limited, these models rely on sub-word tokenization, potentially splitting words into smaller units that are not morphologically or semantically motivated. This linguistically counter-intuitive mechanism does not seem to be detrimental, at least judging by the models' success on downstream tasks.

We challenge this assumption by investigating a type of linguistic structure where it is especially important for sub-word representations to be meaningful. Specifically, we vary and analyze BERT representations of German noun compounds.[1] These are conventionally represented as a single orthographic word (e.g. *Zitronensaft* 'lemon juice') with two clearly identifiable constituents (*Zitrone* 'lemon' and *Saft* 'juice'). But BERT would tokenize our example as [*Zit*, *##ronen*, *##sa*, *##ft*], ignoring both constituents and instead including semantically unrelated sub-word fragments which occur in a highly diverse set of also irrelevant contexts. Moreover, these suboptimal representations are especially problematic in under-resourced settings, where a compound may only occur a few times.

This study aims to identify the most robust BERT representations of German compounds by experimenting with variants derived from the base and re-trained models using three tokenization strategies: the default tokenizer, a re-trained tokenizer, and pre-splitting compounds into their constituents. We evaluate on (i) the model's likelihood of predicting a compound in the masked language modelling task, additionally using the semantic relatedness of these predictions to the target compound as defined by GermaNet paths; and (ii) the correlation of model-internal measures of compound–constituent similarity with human compositionality ratings. In order to maximally enforce the need for robust representations, we run the experiments[2] on a diachronic corpus as a notoriously restricted type of data. The most consistent improvements are obtained when compounds are pre-split into constituents; as a simple but efficient preprocessing step, this has direct implications for contextual representations of other types of complex words.

2 Prior Work

Impact of tokenizers on contextual embeddings Rust et al. (2021) and Agarwal et al. (2023) evaluated multilingual contextual language models on downstream tasks such as question answering and event detection. They noted that a lack of representation of some languages in the tokenizer vocabulary for models like multilingual-BERT hurts the performance, but that this effect can be mitigated with a monolingual tokenizer (Rust et al., 2021) or learning a function to aggregate sub-word tokens and to compensate for the relatively higher fragmentation of tokens in under-represented languages (Agarwal et al., 2023). These works were

[1] We focus our investigation on the subset of noun-noun compounds for simplicity.

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

Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 16131–16136
December 6-10, 2023

an important step toward addressing the negative effect that sub-word-tokenization can have on semantic representation, going beyond the approach by Devlin et al. (2019) to use the first sub-word token as the representation of a longer sequence.

Semantic representation in contextual embeddings Shwartz and Dagan (2019) found contextualized embeddings to be more successful than static embeddings as a basis for classification tasks involving detecting semantic differences from the typical meaning of several kinds of multi-word expressions. Ethayarajh (2019) explored the geometry of contextual vector spaces, showing that randomly selected words' vector representations are more similar (via cosine similarity) than would be expected if the vectors were distributed uniformly. They also found that vector representations formed from later layers were more context-specific, and those from earlier layers functioned better when used to create static representations. These works will inform our baseline expectations for forming contextual semantic representations.

3 Resources

Here we describe our corpus of historical German texts used to train and adapt semantic representations, human compositionality ratings of German compounds, and the lexical taxonomy GermaNet.

Deutsches Textarchiv – DTA The DTA (Berlin-Brandenburgischen Akademie der Wissenschaften, 2022) is a diachronic, curated selection of German texts of various genres. We use a portion of the corpus (1814–1900) to train and fine-tune our models. We rely on the orthographically modernized, normalized, lemmatized versions of the texts made available in the DTA. All of ≈4M sentences (≈89M tokens) were shuffled, and 10% were held out as evaluation data.

Human compositionality ratings Compositionality ratings measure the degree to which the compound's head or modifier constituents contribute to the meaning of the compound, in our case ranging from 1 (totally unrelated to compound meaning) to 6 (fully accounts for compound meaning). We rely on the ratings for German noun-noun compounds in the GhoSt-NN dataset (Schulte im Walde et al., 2016), but exclude any compound occurring fewer than 20 times in our training data, leaving 185 noun-noun compounds as our set of test items.

bert-base-german-cased[3] This model was used as a basis for fine-tuning on in-domain data from DTA. It was originally trained on German Wikipedia, OpenLegalData,[4] and news articles of unknown provenance. Outside of quotations from older, famous works, there should be minimal overlap between this pretraining data and data from DTA.

GermaNet is a taxonomy of 'synsets', sets of one or more words that are synonymous (Hamp and Feldweg, 1997; Henrich and Hinrichs, 2010). The primary connection between synsets of nouns is established by hypernymy and hyponymy relationships. Paths and distances in this network can be used as a measure of semantic distance (Kim and Baldwin, 2013; Tahmasebi and Risse, 2017).

4 System Configurations

Table 1 provides an overview of our representation variants. We evaluate `bert-base-german-cased` with a range of tokenizer and preprocessing configurations described below. For type-level representations, we also use `word2vec` (Mikolov et al., 2013).

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

Table 1: BERT model configurations. ☑: presence, ☒: absence.

Re-training or fine-tuning BERT configurations `base` and `base-ft-DTA` use the base tokenizer vocabulary, while `voc-rt-DTA` and `split` re-train the tokenizer.[5] All configurations other than the `base` BERT model are trained on DTA data. The `base-ft-DTA` configuration retains the base tokenizer's vocabulary, so its training on DTA data constitutes fine-tuning of the `base` model.

Aggregating layer and token representations Following the suggestion of Schlechtweg et al. (2020), we vary which slices of the BERT embedding layers are used as a representation of a token, taking the sum of either the first, middle, or last four layers (out of 12). Following Montariol et al.

[3] https://www.deepset.ai/german-bert

[4] http://openlegaldata.io/research/2019/02/19/court-decision-dataset.html

[5] Changes to the model's tokenizer vocabulary necessitate learning word embeddings from scratch.

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
Schauspiel (play [theater])	Schauspiel	Schauspiel	schauen + Spiel
Traumbild (vision [imagined])	Traum + ##bild	Traum + ##bild	Traum + Bild

Table 2: Example noun compounds and their tokenizations.

(2021), we average their representations to combine the representations of several tokens, whether they are whole-word or sub-word tokens.

Training settings All BERT models (other than the original `base` configuration) were run for 5 epochs over the training data, with a learning rate of `5e-5`, using a single Nvidia GeForce RTX A6000 GPU, running for approximately 54 hours. Training on DTA data used default settings from Devlin et al. (2019) (masking 15% of tokens, 768-dimensional hidden layer, vocabulary of 30k), with a maximum sequence length of 128 tokens. The word2vec models were trained over 5 epochs, with a minimum term count of 1 and no limit on the vocabulary size, with all other parameters set to default values, using the `gensim` Python package v.4.3.1 (Řehůřek and Sojka, 2010).

Compound pre-processing The `split` configurations for BERT and word2vec perform compound splitting as pre-processing before training. We use the `SimpleCompoundSplitter` (Weller-Di Marco, 2017), which relies on word frequency (at the level of whitespace-delimited tokens) and part-of-speech tags from the training corpus. We conducted a post-hoc evaluation of the splitter using a list[6] of nominal compounds from GermaNet. We adopted the formulations of precision:

$$\frac{correct\ split}{correct\ split\ +\ wrong\ split}$$

and recall:

$$\frac{correct\ split}{correct\ split\ +\ wrong\ split\ +not\ split}$$

as defined by Weller-Di Marco (2017), and obtained a precision of 0.69 and a recall of 0.64. This reduced performance in comparison to the precision of 0.92 and recall of 0.91 in Weller-Di Marco (2017) reflects the smaller size of the data that we used to train the splitter (they used ≈1.5 billion tokens to our ≈89 million), as well as the inclusion of many modern terms in the test set used.

[6]We used version 13.0, which was the closest available version to the list used by Weller-Di Marco (2017).

5 Evaluation

Here we present our two evaluation perspectives: the in-context masked-language-model prediction of target compounds, and the decontextualized exploration of how the compounds and constituents are represented in each model's embedding space.

5.1 Masked Language Model (MLM) Task

If a BERT model has learned an adequate representation of a target compound, we expect it to generate (a subpart of) that compound as a mask-filler in an appropriate context. For each target compound, we extract its occurrences from the DTA evaluation data and construct query sentences by replacing the compound token with a number of `[MASK]` tokens. This number corresponds to the number of tokens in which the compound would be split by that configuration's tokenizer (see examples in Table 2). We then compute two evaluation metrics by (i) directly comparing the predictions against the targets and (ii) characterizing their semantic relatedness based on an external linguistic resource.

Directly predicting compounds We take the top 10 predictions[7] for each mask token (predicted simultaneously), and match each with the n^{th} sub-word token of the target compound (e.g. for *Zitronensaft*, the base tokenizer yields $n = 4$). A partial match is formed if any of the n mask predictions contains the n^{th} sub-token of the target compound (accuracy by dividing by the total number of mask tokens predicted), and a full match is formed if all n mask predictions contain matching sub-tokens of the target compound (accuracy by dividing by the total number of compounds in the evaluation data).

GermaNet scoring of MLM predictions All possible combinations of predicted tokens from the MLM task evaluation (limited to the top five predictions for each mask token due to the combinatorial complexity of the operation) are queried in GermaNet, and are semantically compared using the path similarity measure (Wu and Palmer,

[7]This threshold is arbitrary.

1994), yielding a score from 0.0 (no path possible) to 1.0 (exactly matching synsets) against the synset representing the compound as a whole. To form queries from our model outputs, we combine subword-tokenized continuation tokens with the preceding token(s), remove all ## symbols and query each word separately. When a word is included in multiple synsets, the maximum similarity score is chosen; scores from multiple-word phrases are summed. Lemmatization using `spaCy`[8] is applied if a candidate word is not initially found in GermaNet. The path similarity scores are averaged over *only* the number of words successfully queried in GermaNet. A separate precision measure is provided to quantify the proportion of queried words that did not return any result in GermaNet.

Results Table 3 shows that the `split` and the `base-ft-DTA` configurations performed best on the MLM prediction task, with comparable scores in the partial and full match conditions. There was little variance across configuration scores in the GermaNet path similarity measure. There was, however, a wider spread in precision scores, favoring the two configurations with re-trained tokenizer vocabularies (`voc-rt-DTA` and `split`).

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.

Discussion The poor performance of the `base` BERT configuration across tasks confirms the need for training on in-domain data in a low-resource scenario. But exposure to in-domain data while still using the default tokenization strategy (`voc-rt-DTA`) was not sufficient for the MLM prediction, and neither was exposure to in-domain data without re-training the vocabulary (`base-ft-DTA`) for matches with the GermaNet vocabulary. The most successful `split` configuration seems to make effective use of the limited training data via the granularity of tokens enabled by the compound splitting. We attribute this top performance to the reduction in the number and to the meaningfulness of pieces that target compounds are broken into, since the compound-splitter almost always outputs tokens that are not further split by the re-trained tokenizer. The limitation to hyper- and hyponym links between synsets, as well as the selection of maximum similarity scores when more than one synset is available for a given word could both have contributed to a flattening of scores.

[8] `https://spacy.io/models/de` v.3.5.0 and using the `de_core_news_sm` model

5.2 Vector Similarity and Human Ratings

A robust representation of a compound should be positioned in a vector space near to semantically related terms. We assess this goal based on the proximity of compounds to their constituents, which should be higher for more compositional compounds. Since compositionality prediction is usually addressed as a type-level task (i.e. collapsing the potentially multiple senses of a word into a single representation), we compare our BERT representations against word2vec representations. We compute cosine scores for all compound-head and compound-modifier pairs, and evaluate them against the compositionality ratings mentioned above, using Spearman's rank-order correlation coefficient ρ. We report results with $p \leq 0.05$.

Results Table 4 shows correlations between compositionality ratings and cosine scores comparing compound and head vector representations, which were most strongly rank-correlated for the `base` BERT model using the last segment of hidden layers. Of the configurations that were trained on in-domain data, the `split` configuration (using the first segment of hidden layers) shows the strongest correlation. No other configuration had significant correlations, including all of the compound-modifier comparisons.

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 $\sim$ compositionality ratings.

The word2vec `unsplit` configurations (Table 5) obtained overall stronger correlations between vector similarities and compositionality ratings than any of the BERT configurations, for both compound-head and compound-modifier pairings. Regarding the `split` configurations, the word2vec `split` configuration was weaker than the strongest

word2vec Configuration	Constituent	ρ
unsplit	head	0.525
unsplit	modifier	0.371
split	head	0.207

Table 5: Cosine similarity between word2vec compound and constituent vectors $\sim$ compositionality ratings.

`split` configuration from the BERT models.

Discussion As the reference embeddings that are correlated with compositionality ratings in this evaluation are decontextualized, it was expected to observe stronger correlations from the word2vec models, since they have far fewer parameters to estimate. The stronger correlation observed from the `base` BERT models may be attributable to the larger size of the pre-training data ($24\times$ the DTA training data). That we observe the strongest `base` correlation using the last four layers aligns with what we would expect, cf. Ethayarajh (2019), that representations derived from later layers are more context-specific. We conjecture that the opposite trend seen in the `split` configuration, where the performance declines from the first to mid to last layers, can be attributed to the lower rate of subword fragmentation there, obviating the benefit of the greater context-specificity of the last layers.

From the configurations trained on DTA, the stronger correlations observed from the `split` configurations may be due to situations where the compound was correctly split and is also highly compositional with respect to its head, since the vector of the compound only differs from the vector of the head constituent by being averaged with the modifier's vector. For example, for *Traumbild* 'vision' ('dream' + 'image'), the cosine similarity calculation in the `split` configuration is $cos(avg(\overrightarrow{Traum}, \overrightarrow{Bild}), \overrightarrow{Bild})$ (rather than $\overrightarrow{\#\#bild}$, as in the other configurations). The average compositionality rating with respect to the head constituent for our test compounds is 4.21 (slightly compositional), which may have been advantageous for the `split` configuration, per the example above.

6 Conclusion

We endeavored to investigate the effects of subword tokenization on the semantic representation of German noun compounds with contextualized embedding models, using the contrasting perspectives afforded by an in-context and a decontextualized evaluation. Splitting compounds prior to training BERT embeddings resulted in better in-context performance, thus meeting or surpassing the fine-tuned base model, while making more effective use of limited training data. Furthermore, the `split` model's higher performance compared with the `voc-rt-DTA` model suggests an advantage for this pre-processing approach in data-limited settings, as both of these models were only trained on the DTA data. In the decontextualized evaluation, word2vec models surpassed all BERT models, where the `base` model had the best performance, and the `split` model was competitive with the fine-tuned model. It remains to be seen whether the advantages of splitting compounds would remain if this technique was applied over the full pre-training of German BERT, but overall, we recommend paying closer attention to tokenizer granularity in limited data contexts.

Limitations

Our analysis is limited to a single language: German. Languages like English exhibit similar patterns of noun-noun compound formation, albeit more often written in open (space-separated) or hyphenated forms, while languages that are less typologically similar may use other constructions (e.g. the noun-preposition-noun pattern seen in Romance languages) to productively combine nouns. These other orthographic or grammatical patterns would likely affect the relative importance of our conclusions.

The comparisons between models with and without pre-split compounds would have benefited from an additional configuration, i.e., applying compound splitting to the full German BERT training set, and training that model from scratch. But the full set of training data is not immediately available, and the computational requirements would be excessive.

Lastly, the two perspectives offered by our evaluations (in-context and decontextualized) should only be interpreted from a complementary perspective; when considered alone, they each have inherent limitations. In-context performance may lack generalizability to out-of-domain data, and decontextualized representations are to some degree at odds with the initial model training (where target compounds always occur in some context).

Acknowledgements

Our work was supported by the DFG Research Grant SCHU 2580/5-1 (*Computational Models of the Emergence and Diachronic Change of Multi-Word Expression Meanings*).

References

Shantanu Agarwal, Steven Fincke, Chris Jenkins, Scott Miller, and Elizabeth Boschee. 2023. Impact of subword pooling strategy on cross-lingual event detection. https://arxiv.org/abs/2302.11365.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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. Association for Computational Linguistics.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Birgit Hamp and Helmut Feldweg. 1997. 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*, pages 9–15, Madrid.

Verena Henrich and Erhard Hinrichs. 2010. GernEdiT – The GermaNet editing tool. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, pages 2228–2235, Valletta, Malta. European Language Resources Association (ELRA).

Su Nam Kim and Timothy Baldwin. 2013. A lexical semantic approach to interpreting and bracketing english noun compounds. *Natural Language Engineering*, 19(3):385–407.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. https://arxiv.org/abs/1301.3781.

Syrielle Montariol, Matej Martinc, and Lidia Pivovarova. 2021. Scalable and interpretable semantic change detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4642–4652, Online. Association for Computational Linguistics.

Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? On the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135, Online. Association for Computational Linguistics.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.

Sabine Schulte im Walde, Anna Hätty, Stefan Bott, and Nana Khvtisavrishvili. 2016. 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. European Language Resources Association (ELRA).

Vered Shwartz and Ido Dagan. 2019. Still a pain in the neck: Evaluating text representations on lexical composition. *Transactions of the Association for Computational Linguistics*, 7:403–419.

Nina Tahmasebi and Thomas Risse. 2017. Finding individual word sense changes and their delay in appearance. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 741–749, Varna, Bulgaria. INCOMA Ltd.

Marion Weller-Di Marco. 2017. Simple compound splitting for German. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 161–166, Valencia, Spain. Association for Computational Linguistics.

Zhibiao Wu and Martha Palmer. 1994. Verb semantics and lexical selection. In *32nd Annual Meeting of the Association for Computational Linguistics*, pages 133–138, Las Cruces, New Mexico, USA. Association for Computational Linguistics.