A Systematic Search for Compound Semantics in Pretrained BERT Architectures

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

Abstract

To date, transformer-based models such as BERT have been less successful in predicting compositionality of noun compounds than static word embeddings. This is likely related to a suboptimal use of the encoded information, reflecting an incomplete grasp of how the models represent the meanings of complex linguistic structures. This paper investigates variants of semantic knowledge derived from pretrained BERT when predicting the degrees of compositionality for 280 English noun compounds associated with human compositionality ratings.

Our performance strongly improves on earlier unsupervised implementations of pretrained BERT and highlights beneficial decisions in data preprocessing, embedding computation, and compositionality estimation. The distinct linguistic roles of heads and modifiers are reflected by differences in BERT-derived representations, with empirical properties such as frequency, productivity, and ambiguity affecting model performance. The most relevant representational information is concentrated in the initial layers of the model architecture.

1 Introduction

The meaning of multiword expressions such as noun compounds is notoriously difficult to model, particularly because of variability in their degree of compositionality, i.e. the relatedness of the meaning of a compound (e.g. *flea market*) to that of the individual constituents (*flea* and *market*). The degree of compositionality has been successfully predicted using static word embeddings, but transformer-based models such as BERT (Devlin et al., 2019) have so far been less successful. This might be related to a suboptimal use of the information encoded by the models, reflecting our incomplete grasp of how they represent the meanings of complex linguistic structures.

In this paper, we aim to improve this understanding, as well as produce actionable methodological recommendations. We predict the degrees of compositionality of 280 English noun compounds associated with human compositionality ratings. We extract their occurrences from a web corpus and model them using pretrained BERT. Like previous work, we assume that the contextualized nature of these representations may capture key aspects of compound semantics. But we do not expect this information to be equally accessible across the model or independent from underlying linguistic properties. We therefore experiment with variants of BERT-derived semantic knowledge (comprising over 40,000 ways of computing the degree of compositionality), and analyze the linguistic roles of compound constituents and their empirical properties (frequency, ambiguity, and productivity). We provide the following contributions:

- We identify a robust setup to extract compositionality information from pretrained BERT. It strongly improves on earlier unsupervised implementations and highlights beneficial decisions in data preprocessing, embedding computation, and compositionality estimation.
- We show that the distinct linguistic roles of heads and modifiers are reflected by differences in BERT-derived representations. Further focusing on compound heads, we find clear effects of their empirical properties on model performance.
- Our results support the view that pretrained BERT encodes at least some aspects of the semantics of multiword expressions, and also show that the most relevant information is found in the model's initial layers.

The remainder of this paper is organized as follows. We first review related studies (§2), and then introduce our data (§3) and experimental setup (§4). We then analyze and discuss the results (§5) and provide a conclusion (§6).

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 1499–1512
May 2-6, 2023 ©2023 Association for Computational Linguistics

2 Related work

The meaning of noun compounds is modeled from a broad range of perspectives. In psycholinguistics, for example, there is a long tradition of research on human processing of compound semantics. Its focus is usually on semantic transparency, which is operationalized using measures including the semantic relatedness of the constituents and the retention of their meaning in the compound (e.g. Bell and Schäfer, 2016; Auch et al., 2020; Günther et al., 2020). Computational studies have examined a similar range of compound properties, aiming to predict the meaning of the whole compound (Mitchell and Lapata, 2008; Dima et al., 2019), the semantic relations between a compound's constituents (Ó Séaghdha, 2007; Dima et al., 2014), or the compound's degree of compositionality. The latter issue is also the focus of our work.

The ability of computational models of compound semantics to predict the degree of compositionality is usually evaluated on a ranking task with gold standard data in form of human compositionality ratings, which exist for languages including English (Reddy et al., 2011), German (von der Heide and Borgwaldt, 2009; Schulte im Walde et al., 2016), French and Portuguese (Cordeiro et al., 2019). Strong results have been obtained using static word embeddings, generally by learning dedicated representations for the whole compound and comparing them against the representations of the individual constituents, often combined using composition functions (Reddy et al., 2011; Schulte im Walde et al., 2013, 2016; Salehi et al., 2014, 2015; Cordeiro et al., 2019; Alipoor and Schulte im Walde, 2020). Most studies predict the compositionality of the whole compound, but differences between heads and modifiers have been underscored. Reported effects on model performance are related to their distinct linguistic roles and empirical properties such as frequency, productivity, and ambiguity (Schulte im Walde et al., 2013, 2016; Alipoor and Schulte im Walde, 2020).

More recently, compositionality prediction has been addressed using pretrained transformer-based models. While BERT-derived representations used in a subsequently trained classifier performed well on binary classification (Shwartz and Dagan, 2019), their unsupervised use in the standard ranking task formulation has been less successful. For instance, an implementation based on comparisons between contextualized and non-contextualized compound representations obtained significantly poorer results compared to static word embeddings, leading to a suggestion that these models do not capture compositionality in a way similar to human annotators (Garcia et al., 2021a). Stronger correlations with human judgment were obtained in a probing study, but using external linguistic knowledge – gold standard synonyms of noun compounds (Garcia et al., 2021b). We are unaware of unsupervised implementations of BERT-derived representations that are competitive with static word embeddings on this task.

This might be related to a suboptimal use of the information encoded by BERT, as the layers in its architecture capture different aspects of linguistic structure (Rogers et al., 2020). For instance, it has been suggested that semantic knowledge in general (Jawahar et al., 2019) and word sense information in particular (Coenen et al., 2019) is encoded in higher layers, but also that type-level information is encoded in lower layers (Vulić et al., 2020). More generally, these and other types of implementation decisions impact BERT performance on other linguistically oriented tasks (e.g. Laicher et al., 2021). To the best of our knowledge, these patterns have not been investigated in detail for compound semantics, with the cited studies generally relying on widely used solutions (e.g. computing a token-level embedding by averaging over the last four layers).

3 Data

This section introduces the data resources we used. For details on licenses, see Appendix A.

3.1 Gold standard of noun compounds

We use the set of 280 English noun compounds introduced by Cordeiro et al. (2019). It includes an initial set of 90 compounds created by Reddy et al. (2011)[1] and a further 190 compounds annotated by Cordeiro and colleagues using the same rating procedure.[2] Human annotators were asked to provide compositionality ratings in terms of literality, on a scale from 0 (not at all literal) to 5 (very literal). They provided scores for the interpretation of the whole compound (e.g. *crash course*), as well as for the use of the modifier (*crash*) and the head (*course*) within it. Sample compounds and their ratings are shown in Table 1.

[1] http://www.dianamccarthy.co.uk/downloads.html

[2] https://pageperso.lis-lab.fr/carlos.ramisch/?page=downloads/compounds

	Compositionality rating		
Compound	Modifier	Head	Phrase
guinea pig	0.47 ± 0.72	0.47 ± 0.72	0.24 ± 0.56
flea market	0.38 ± 0.81	4.71 ± 0.84	1.52 ± 1.13
biological clock	4.71 ± 0.47	1.76 ± 1.35	2.29 ± 1.21
health insurance	4.53 ± 0.88	4.83 ± 0.58	4.40 ± 1.17

Table 1: Sample gold standard compounds with compositionality ratings (mean and standard deviation).

3.2 Corpus

As corpus data for the modeled noun compounds, we rely on the widely used ENCOW corpus, obtained by crawling web data and containing ≈ 9.6 billion words (Schäfer and Bildhauer, 2012; Schäfer, 2015). For each compound, all tokenized sentences in which it appears are extracted. We only use singular forms so as to avoid potential variability related to grammatical number in BERT.

3.3 Empirical compound properties

Parts of our analysis use information on empirical properties of compound constituents (in particular, their heads), and specifically (i) frequency; (ii) productivity, i.e. the number of compound-types in which they appear; and (iii) ambiguity, i.e. their number of senses. We use the information on these properties created by Schulte im Walde et al. (2016) for the Reddy et al. (2011) dataset. They derived frequency and productivity information from the ENCOW corpus, and calculated ambiguity based on WordNet (Fellbaum, 1998). We apply the same procedures to calculate the information for the compounds from the Cordeiro et al. (2019) dataset.

4 Experimental setup

4.1 BERT representations

We use BERT-base-uncased, a 768-dimension, 12-layer version of the model (110 million parameters) pretrained on English data, from the Hugging Face implementation (Wolf et al., 2020). We run the experiments on a CPU computing server (2×12 3GHz cores with 768GB RAM) over ≈ 5 days. In order to facilitate the analysis of modeling properties, we deliberately adopt a straightforward setup without fine-tuning. Each sequence from the corpus is fed into the model, which returns multiple vector representations for each token in the sequence. For all sequences of a compound, we retain the representations for each token in the sequence; these correspond to the input embedding layer and the outputs of the 12 hidden states. We test different ways of combining the obtained information.

Embedding types. BERT produces contextualized representations for each token in the sequence, which we use both individually and by combining multiple token representations (see pooling functions below). We compute the following types of embeddings. (i) modif: representation of the modifier, corresponding to its contextualized embedding. (ii) head: representation of the head, corresponding to its contextualized embedding. (iii) comp: representation of the full compound, corresponding to pooled modif and head embeddings. (iv) cont: representation of the context in which the compound appears, corresponding to the merged embeddings of all tokens in the sequence apart from modif, head, [CLS], and [SEP]. (v) cls: embedding of the [CLS] token, taken to correspond to a representation of the full sequence.

BERT's tokenizer splits out-of-vocabulary tokens into subwords with known representations. When this occurs for modif or head, the subword representations are pooled into a single embedding.

Layer combinations. We test all contiguous spans of layers, across the input embedding and the 12 hidden state outputs, for a total of 91 layer combinations. The smallest combination is a single layer, and the largest is the full range of 13 layers.

Pooling functions. Multiple vectors can be combined in different ways; we test two options, averaging (avg) or summing (sum) over them. This is applied (i) token-wise, if merging multiple token representations; (ii) layer-wise, if merging multiple layers; (iii) sequence-wise, if creating a type-level representation (see below). In order to streamline the experimental setup, the same pooling function is used in all three cases.

Sequence length. We test if using longer sequences can be beneficial based on the assumption that a larger context may be more semantically discriminating. We either retain only sequences with at least 20 space-separated tokens, or do not impose any threshold; this technically corresponds to a minimum of 3 tokens, i.e. the lowest sequence length in the corpus.

Number of sequences. From a similar perspective, we examine the effect of increasing the number of modeled sequences, experimenting with 10, 100, and 1,000 sequences per compound. The sets of sequences are not resampled; rather, the smaller sets are included in the larger ones. This criterion is combined with that of sequence length: for instance, we extract 1,000 sequences of any length as well as 1,000 sequences with at least 20 tokens, although these may partly overlap. For compounds whose corpus frequency is lower than the threshold, all available occurrences are used.

4.2 Compositionality estimates

As stated above, we expect that the contextualized representations we use may carry semantic information reflecting the degree of compositionality of a compound, but it is unclear which specific combination of representations is the most efficient.

Direct estimates. We directly compute pairwise cosine scores between pairs of embeddings, testing all 10 pairs of embedding types (`head`–`modif`, `head`–`comp`, `head`–`cont`, and so forth).

Composite estimates. We further combine the directly measured information pertaining to the head and the modifier of a compound using the same composition functions as Reddy et al. (2011). Specifically, we use `head` and `modif` embeddings in combination with one of the following: `comp`, `cont`, and `cls`. Taking `comp` as an example, we compute the composite estimates as follows:

$$\begin{aligned}
\text{ADD} &= \cos(\texttt{comp}, \texttt{modif}) + \cos(\texttt{comp}, \texttt{head}) \\
\text{MULT} &= \cos(\texttt{comp}, \texttt{modif}) \cdot \cos(\texttt{comp}, \texttt{head}) \\
\text{COMB} &= \text{ADD} + \text{MULT}
\end{aligned}$$

Token-level vs. type-level. The information from individual occurrences can be aggregated into a single numerical score in different ways. (i) In the token-level approach, we compute a compositionality estimate for each individual occurrence, and then average over those values to get a single score. (ii) In the type-level approach, we first compute a type-level representation by applying a pooling function; as an example, we average or sum over all individual `head` embeddings to produce a type-level `head` embedding. This type-level representation is then used to directly compute a single compositionality estimate.

The combinations of the presented experimental parameters correspond to a total of 41,496 ways of computing a numerical estimate of the degree of compositionality. In a trial run, we also experimented with other parameters (e.g. restrictions on the position of the compound within the sequence), but they did not exhibit strong effects and for clarity are not included in the present discussion.

5 Results and discussion

We evaluate each constellation of parameters by calculating Spearman's rank correlation coefficient between the compositionality estimates it produces and human judgments from the test set. All implementations are evaluated on three prediction targets: compositionality scores for the compound as a whole, the head, and the modifier.

5.1 Overview of model performance

We begin by looking at general trends in model performance (Table 2). The highest correlation coefficient we obtain stands at 0.706 for compound-level compositionality. Compared to previous studies on the same dataset, it is in a similar range as the result reported by Cordeiro et al. (2019) using static word embeddings ($\rho = 0.726$).[3] Their best performance was obtained by training a `word2vec` model on a corpus in which compound occurrences had been joined into single tokens. The cosine scores between those representations and compositionally constructed vectors were then used to predict the degree of compositionality. Interestingly, this procedure strongly relies on the context in which the compounds occur, which is also a key component of our best approaches (see below).

The highest performance we obtain substantially improves on the best BERT-derived score reported by Garcia et al. (2021a) based on comparisons between contextualized and non-contextualized representations of a compound ($\rho = 0.37$). In another BERT-based experiment, Garcia et al. (2021b) obtain a higher correlation using similarity measurements between compounds and gold standard synonyms ($\rho = 0.67$); by contrast, we do not rely on external linguistic knowledge.

The implementations we tested are strongest at predicting compound-level compositionality. This is followed by the compositionality of the head (maximum $\rho = 0.645$) and then of the modifier (maximum $\rho = 0.553$). Predicted values for the three types of scores are strongly correlated ($\rho > 0.9$ for all three pairwise comparisons), but the best performing parameter constellations clearly differ.

More generally, the summary in the table underscores the wide range of obtained values; the

[3] This result is reported for a subset of the same dataset containing 180 compounds.

	ρ	layers	pool	len	seqs	estimate		agg	ρ	layers	pool	len	seqs	estimate		agg
COMP	0.706	1-1	sum	3	1k	COMB	cont	token	-0.642	2-3	avg	3	1k	cont	cls	type
	0.706	1-1	avg	3	1k	COMB	cont	token	-0.644	3-3	avg	3	1k	cont	cls	type
	0.706	1-1	sum	20	1k	MULT	cont	token	-0.645	1-5	avg	3	1k	cont	cls	type
	0.706	1-1	avg	20	1k	MULT	cont	token	-0.646	2-4	avg	3	1k	cont	cls	type
	0.706	1-1	sum	3	1k	MULT	cont	token	-0.649	1-4	avg	3	1k	cont	cls	type
HEAD	0.645	1-1	sum	3	1k	head	cont	token	-0.598	0-7	avg	3	1k	cont	cls	type
	0.645	1-1	avg	3	1k	head	cont	token	-0.599	1-4	avg	3	1k	cont	cls	type
	0.638	1-1	sum	3	1k	COMB	cont	token	-0.600	0-6	avg	3	1k	cont	cls	type
	0.638	1-1	avg	3	1k	COMB	cont	token	-0.604	1-5	avg	3	1k	cont	cls	type
	0.638	1-1	sum	3	1k	ADD	cont	token	-0.606	1-6	avg	3	1k	cont	cls	type
MODIF	0.553	1-1	avg	20	1k	modif	cont	token	-0.464	2-4	avg	3	1k	cont	cls	type
	0.553	1-1	sum	20	1k	modif	cont	token	-0.465	1-5	avg	3	1k	cont	cls	type
	0.548	1-1	sum	3	1k	modif	cont	token	-0.471	1-3	avg	3	1k	cont	cls	type
	0.548	1-1	avg	3	1k	modif	cont	token	-0.474	1-4	avg	3	1k	cont	cls	type
	0.546	1-1	avg	20	1k	modif	cont	type	-0.476	1-2	avg	3	1k	cont	cls	type

Table 2: Best (left) and worst (right) evaluated implementations. Abbreviations: *pool* = pooling function; *len* = minimum tokens per sequence; *seqs* = minimum number of modeled sequences; *agg* = aggregation of occurrences (type vs. token-level).

weakest implementations yield negative correlations, with a low of -0.649. This confirms the relevance of the parameters we tested and the importance of understanding the optimal choices in implementing them. Trends suggested by the initial overview include better performance of (i) the first hidden layer in isolation; (ii) token-level rather than type-level modeling; (iii) cont embeddings, when paired with another relevant type of information; (iv) embeddings targeting the compound as a whole, the head, or the modifier for the corresponding compositionality score. Strikingly, all the weakest constellations are closely similar to one another, the only distinguishing characteristic being the span of layers. But some of these parameter choices are also found in the best performing implementations; more generally, the direction and relevance of all trends are not immediately apparent. We therefore now more closely examine individual parameters.

Sequence length. Using sequences with at least 20 tokens leads to lower mean correlations. However, their minimum values are higher and maximum values are comparable (Table 3). The overall lack of a clear effect suggests that beneficial distinctions made by the model are primarily based on the most immediate linguistic context.

	min. 3 tokens	min. 20 tokens
COMP	0.146 (-0.649, 0.706)	0.134 (-0.587, 0.706)
HEAD	0.102 (-0.606, 0.645)	0.087 (-0.561, 0.637)
MODIF	0.099 (-0.476, 0.548)	0.093 (-0.460, 0.553)

Table 3: Spearman's ρ (mean, min, max) for minimum sequence length.

	10 sequences	100 sequences	1,000 sequences
C	0.135 (-0.394, 0.622)	0.142 (-0.607, 0.689)	0.143 (-0.649, 0.706)
H	0.093 (-0.384, 0.565)	0.094 (-0.551, 0.621)	0.096 (-0.606, 0.645)
M	0.089 (-0.367, 0.495)	0.101 (-0.459, 0.544)	0.098 (-0.476, 0.553)

Table 4: Spearman's ρ (mean, min, max) for number of sequences. C, H, M = compound, head, modifier.

Number of modeled sequences. There is a clear trend towards an increase in performance with an increase in the number of modeled sequences (Table 4), likely because this facilitates disambiguation and limits the effect of sampling differences. Looking at the maximum values, the shift from 10 to 100 occurrences leads to a stronger improvement ($\approx$ 0.06 points) than the shift from 100 to 1,000 ($\approx$ 0.02 points); this suggests that it is especially important to avoid very low numbers of examples. While increasing the number of occurrences also leads to the lowest performances overall, this may be due to the detrimental effect of other parameters.

We further assessed the impact of sampling differences in the condition with 10 sequences, which is the most inherently unstable. We ran the compositionality estimation 10 times, each time randomly resampling the modeled occurrences. The mean difference between the minimum and maximum values obtained by a parameter constellation is 0.028. This variability does not alter the overall parameter-level trends (see Appendix B).

Some compounds do not exceed the threshold frequencies for some parameter combinations. This affects 10 compounds for the 100 sequence threshold, and 97 compounds for the 1,000 sequence

	avg	sum
Comp	0.139 (-0.649, 0.706)	0.141 (-0.587, 0.706)
Head	0.094 (-0.606, 0.645)	0.095 (-0.563, 0.645)
Modif	0.095 (-0.476, 0.553)	0.097 (-0.460, 0.553)

Table 5: Spearman's ρ (mean, min, max) for pooling functions.

	token-level	type-level
Comp	0.150 (-0.584, 0.706)	0.130 (-0.649, 0.699)
Head	0.103 (-0.556, 0.645)	0.085 (-0.606, 0.628)
Modif	0.100 (-0.460, 0.553)	0.092 (-0.476, 0.546)

Table 6: Spearman's ρ (mean, min, max) for token- vs. type-level processing.

threshold. As a check, we calculated correlations on a smaller set of compounds, excluding the 10 most affected by frequency issues. The results were near-identical, with marginal improvements for the top predictions (e.g. best ρ increasing from 0.706 to 0.710 for compound scores) and the same general trend. As for the compounds with fewer than 1,000 sequences, the mean number of sequences available for these items was 510, which is still a strong increase compared to the preceding threshold.

Pooling functions. Averaging and summing perform similarly when creating a single embedding from multiple vectors (Table 5). They obtain near-identical mean and identical maximum values; averaging leads to lower minimum values.

One of the ways we used pooling functions was to merge representations for out-of-vocabulary tokens that are split up by BERT's tokenizer. This affected 14 compounds; some instances reflect derivational patterns (e.g. `mail`, `##ing` in *mailing list*), but others are more obscure (e.g. `gr`, `##av`, `##y` in *gravy train*). Since it is unclear what some of these representations capture, we checked their impact by calculating correlations on a reduced set of compounds, excluding those with OOV tokens. The results were marginally higher (e.g. ρ increasing from 0.706 to 0.710 for compound scores). This issue therefore does not appear to have a strong detrimental effect, at least when it is limited to a small subset (5%) of target items.

Another frequently used pooling function is concatenation. Applying it across multiple tokens (e.g. for out-of-vocabulary items) or multiple sequences (to create a type-level representation) would result in comparisons between vectors with different numbers of dimensions; we therefore did not include it in the full experimental setup. As a check, however, we ran concatenation across layers, combined with averaging over tokens and sequences. It led to slightly higher mean correlations (0.147 for compound scores), but it did not improve on the maximum results so we did not experiment further.

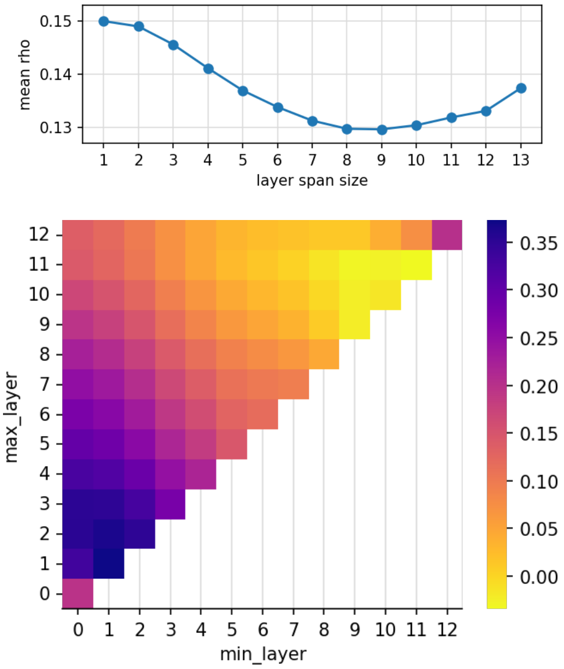

Figure 1: Top: effect of the number of modeled layers on compound-level compositionality prediction. Bottom: mean correlations for compound compositionality prediction across layer combinations. The `min_layer` and `max_layer` values are start and end points of a contiguous span of layers.

Token vs. type-level. There is a clear preference across the board for token-level processing (Table 6). Most improvements over type-level are $\approx$ 0.01–0.02 points; they reach $\approx$ 0.06 points when comparing the worst-performing configurations. A potential explanation is that estimating compositionality on individual occurrences – rather than a merged type-level representation – may be less sensitive to ambiguous or otherwise noisy data.

Layers. The results are affected by the number of modeled layers (Figure 1). The best performance is obtained with a single layer (mean $\rho = 0.150$ for compound scores) and decreases as the span increases up to 9 layers (0.130). Larger spans fare somewhat better, likely because they are bound to capture some relevant representational information; the best is the full range of 13 layers (0.137).

In terms of specific layers, the best result overall is obtained by the first hidden layer in isolation (mean $\rho = 0.373$), followed by other combinations and individual layers in the low-to-mid range. This is plotted in Figure 1 for compound compositionality scores; head and modifier scores follow the same trends (see Appendix C). The traditionally used combination of the last four hidden layers (9–12) is not among the best ones. It is in fact the single weakest in terms of maximum values: 0.248 for compound scores, close to 0.5 below the best implementation. However, it also has a comparatively high minimum value (-0.262), suggesting it is more robust to the effect of other parameters.

Compositionality estimates. We summarize the impact of compositionality estimates by looking at performance across the five types of embeddings which constitute the basis of subsequent score calculations (Table 7); for a summary of results on individual estimates, see Appendix D.

	modif	head	comp	cont	cls
COMP	0.135	0.274	0.245	0.172	-0.128
	-0.383	-0.133	-0.324	-0.649	-0.649
	0.615	0.630	0.666	0.706	0.611
HEAD	0.071	0.242	0.194	0.130	-0.161
	-0.384	-0.130	-0.327	-0.606	-0.606
	0.464	0.645	0.598	0.645	0.558
MODIF	0.106	0.167	0.164	0.133	-0.094
	-0.274	-0.130	-0.229	-0.476	-0.476
	0.553	0.415	0.517	0.553	0.477

Table 7: Spearman's ρ (mean, min, max) for embedding types, across all direct and composite estimates if used.

Looking at the mean values, head performs stronger than modif as well as comp across the prediction targets. This is coherent with the dominant role of the head in the morphological constituency of compounds. However, the maximum values out of the three are obtained by comp for the compound-level score, head for the head score, and modif for the modifier score. This indicates that representations targeting the whole compound or a constituent of interest are successful in capturing information specific to the respective element.

As for the two other embedding types, the mean values for cont follow head and comp across prediction targets. But its maximum values are the single best (compound-level prediction) or joint best (head and modifier predictions). By contrast, cls is clearly in the lower range of performance. Its maximum values remain around 0.1 points behind the best implementations. This shows that using the linguistic context surrounding the compound – modeled by cont – is beneficial, and clearly more so than using a representation of the full sequence.[4]

This might be explained by the redundancy of using a representation of a compound or its constituent, and comparing it with a representation of the full sequence which encodes the same element. It could also be the case that, as suggested before, token-level embeddings are strongly influenced by their immediate linguistic context; a balanced representation of a broader range of information might be complementary. Whatever the case, the results show that the similarity between a compound and its linguistic context is reflective of the degree of compositionality. This might explain why previous implementations using neutral contexts were less successful in capturing these trends.

5.2 Ablation study

In order to further validate the parameter-level trends observed thus far, we conduct an ablation study. We start from the parameter constellation that obtained the best results on predicting the compositionality of the compound as a whole (see Table 2; the top two configurations obtained identical results). For each parameter, we then test all other potential values, one at a time, while keeping the remaining parameters unchanged. The results are presented in Figure 2.

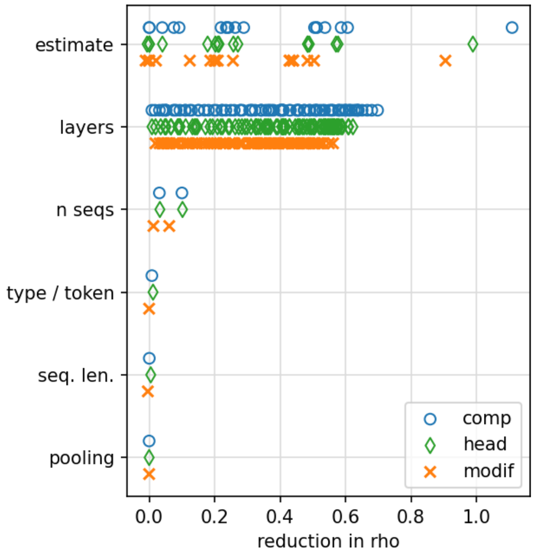

Figure 2: Effect of alternative parameter values compared to the top parameter constellation.

[4] A more direct comparison would involve a merged representation of all the tokens in the sequence rather than the cls embedding. We nevertheless think that we capture the same range of information.

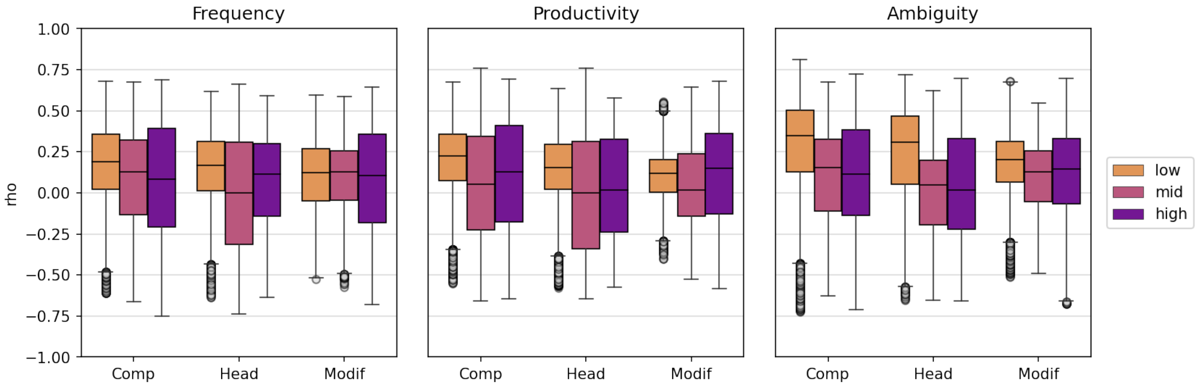

Figure 3: Effect of empirical properties of the head on model performance, observed across the evaluated implementations. Values on the x-axis indicate prediction targets (compound, head, and modifier scores).

Several parameters exhibit no or limited differences with respect to the top parameter constellation: pooling functions, minimum sequence length, and type vs. token-level modeling. Clearer effects on performance are observed for the number of modeled sequences; the drop is strongest for the smallest number of sequences. The choice of layers shows a very strong effect, with a clear tendency towards a drop in performance with both (i) a shift from initial to later hidden states, and (ii) an increase in layer span size. The strongest effect is exhibited by the estimate used to predict the degree of compositionality. All weakest estimates (reduction in $\rho > 0.4$) involve the `cls` embedding. All strongest estimates (reduction in $\rho < 0.1$) involve the `cont` embedding.

These results overall confirm the trends previously discussed for individual parameters. They also provide further evidence that model performance is most strongly affected by the choice of representational information to be used, i.e. the layers and the modeled tokens in the sequence. Suboptimal values for either of these parameters can lead to compositionality predictions that are fully decorrelated from human judgment. Performance does not depend to the same extent on the way in which data is preprocessed and representational information is combined. However, the effect of some of these parameters is not negligible, especially cumulatively, and as such should not be disregarded.

5.3 Empirical properties of compounds

We now turn to the potential impact of empirical properties of the compounds on model performance. We focus on key characteristics of the compounds' heads, given their importance indicated by the trends for compositionality estimates.

As previously stated, we examine the impact of three empirical properties: frequency, productivity, and ambiguity (for sources of this information, see Section 3.3). For each property, we rank the 280 compounds based on the values they exhibit and split them into five sets containing 56 compounds each. We retain the first, third, and fifth set, which we take to clearly reflect the low, mid, and high ranges for each empirical property. The remaining two sets are excluded in order to avoid overlapping or closely similar values in adjoining sets.

A summary of the splits across the three features is presented in Table 8. For each feature, we compute correlations with human judgments separately for each of the three splits of compounds, across all evaluated constellations of parameters. The distribution of the obtained values is plotted in Figure 3 and further discussed below.

Feature	Mean	Std.	Example
Frequency	42	± 30	*silver **spoon***
(thousands)	452	± 108	*labor **union***
	3,614	± 2,438	*crash **course***
Productivity	7	± 5	*night **owl***
	75	± 19	*time **difference***
	448	± 208	*birth **rate***
Ambiguity	2	± 1	*research **project***
	5	± 1	*flea **market***
	13	± 4	*application **form***

Table 8: Mean and standard deviation for the low, mid, and high range splits across empirical features. A sample compound from each split is provided.

Frequency. In predicting the compound and head compositionality scores, the best performance is obtained for low-frequency heads (mean $\rho = 0.18$

and 0.16, respectively). For the compound score, performance across the mid and high range is similar; for the head score, it is clearly lower in the mid range. As for the modifier score, performance is overall stable (mean $\rho = 0.10$ across the three splits); this is not especially surprising given the focus on the frequency of the head. This contrasts the previously reported improvements in performance with a higher number of modeled sequences. However, head frequency is correlated with both productivity and ambiguity ($\rho = 0.87$ and 0.50, respectively, across the 280 compounds). Poorer performance in higher frequency ranges is consistent with an indirect effect of these two properties.

Productivity. Overall, low-productivity heads clearly obtain the best results for compound and head compositionality scores (mean $\rho = 0.21$ and 0.16, respectively). Low- and high-productivity heads obtain similar results for the modifier scores (mean $\rho = 0.11$ and 0.12, respectively). Across the prediction targets, mid-productivity heads have the poorest performance; this drop is the strongest for head scores (mean $\rho = 0.01$). The overall better performance at modeling low-productivity heads is likely explained by their very nature of being used with fewer distinct modifiers, which might facilitate the learning of those compound meanings. This is opposed to higher productivity ranges, which potentially imply more dispersion.

Ambiguity. Across the prediction targets, low-ambiguity heads clearly have the strongest performance (mean $\rho = 0.30$ for compound scores; 0.25 for head scores; 0.18 for modifier scores). The difference with respect to the two other ambiguity ranges is the strongest for head scores; it amounts to ≈ 0.20 points. Performance is similar for mid- and high-range ambiguity. For compound and head scores, the mid range performs slightly better than the high range; for modifier scores, it is the reverse. Similarly to low productivity, the better results for low ambiguity suggest that limited semantic dispersion across the occurrences of a given word makes it easier for BERT to learn its meaning. This is in turn beneficial for derived representations of more complex linguistic structures such as compounds.

Overall, these results have shown that compound properties affect predictions of the degree of compositionality, with better performance in the lower ranges of all three properties. This directly echoes the results for type-level word embeddings reported by Schulte im Walde et al. (2016), who similarly suggested that prediction performance is related to broader effects of compound properties on the quality of the underlying meaning representations. Moreover, the prediction of compound and head scores generally follows the same pattern across the three empirical features. Although modifier scores do not align as closely, they too point to an effect of head properties. These results further underscore the importance of the representations of compound heads for the modeling approach we adopted.

6 Conclusion

We have presented an experiment on predicting the degree of compositionality of 280 English noun compounds using a wide range of variants of semantic knowledge derived from pretrained BERT. We have identified a competitive best implementation achieving $\rho = 0.706$ with human judgement and highlighted clear takeaways.

In terms of preprocessing, stronger results were obtained when modeling a larger number of occurrences per compound, whereas controlling for sequence length did not have a clear effect. On embedding computation, different pooling functions led to comparable performance, but there was a clear advantage for layers in the low-to-mid range, which strongly improved on layer combinations used in earlier studies. As for compositionality estimates, it was clearly beneficial to compute them on the token rather than type level, as well as to use (i) representations targeting the constituent of interest; (ii) representations of the surrounding context; (iii) comparisons across complementary – rather than redundant – representational information.

Looking at empirical properties of compounds, low-frequency, low-productivity, and low-ambiguity heads obtain better compositionality predictions. This trend confirms that more limited semantic dispersion makes it easier to model compound meaning. The fact that it holds across compound, head, and modifier compositionality scores highlights the importance of the head in the linguistic structure – and computational modeling – of compounds. Taken together, our results point to important practical decisions when running similar implementations and contribute to our understanding of the way in which BERT represents lexical meaning, supporting the view that the pretrained model encodes at least some aspects of compound semantics.

Limitations

Our experiments were limited to noun compounds in a single pretrained model for English, with potential implications for the generalizability of our results. They may be partly related to the architecture of this specific model. From a linguistic standpoint, compound properties vary widely across languages. For instance, where English has productive patterns combining two nouns, often in an open (space-separated) compound, German has closed compounds; Romance languages widely rely on N-Prep-N patterns; the structure in many Slavic languages involves patterns of nominal declension; and so forth. The model might not capture the information relevant for compositionality prediction in the same way across these cases. Additionally, our results are strongly related to the central role of compound heads; they may therefore be different in multiword expressions with a different linguistic structure, such as particle verbs and idioms. Finally, we used token-level representations to predict type-level compositionality judgments. This is relevant in terms of assessing the general ability to infer type-level information, but (i) performance may be improved by controlling for the senses in the modeled occurrences vs. those in the stimuli used to collect human judgements; (ii) further work is needed to fully understand the factors contributing to individual token-level representations of a compound.

Acknowledgements

The research 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

Pegah Alipoor and Sabine Schulte im Walde. 2020. Variants of vector space reductions for predicting the compositionality of English noun compounds. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4379–4387, Marseille, France. European Language Resources Association.

Leah Auch, Christina L. Gagné, and Thomas L. Spalding. 2020. Conceptualizing semantic transparency: A systematic analysis of semantic transparency measures in English compound words. *Methods in Psychology*, 3:100030.

Melanie J. Bell and Martin Schäfer. 2016. Modelling semantic transparency. *Morphology*, 26(2):157–199.

Andy Coenen, Emily Reif, Ann Yuan, Been Kim, Adam Pearce, Fernanda Viégas, and Martin Wattenberg. 2019. Visualizing and measuring the geometry of BERT. *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*.

Silvio Cordeiro, Aline Villavicencio, Marco Idiart, and Carlos Ramisch. 2019. Unsupervised compositionality prediction of nominal compounds. *Computational Linguistics*, 45(1):1–57.

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.

Corina Dima, Daniël de Kok, Neele Witte, and Erhard Hinrichs. 2019. No word is an island—A transformation weighting model for semantic composition. *Transactions of the Association for Computational Linguistics*, 7:437–451.

Corina Dima, Verena Henrich, Erhard Hinrichs, and Christina Hoppermann. 2014. How to tell a schneemann from a milchmann: An annotation scheme for compound-internal relations. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1194–1201, Reykjavik, Iceland. European Language Resources Association (ELRA).

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021a. Assessing the representations of idiomaticity in vector models with a noun compound dataset labeled at type and token levels. 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 2730–2741, Online. Association for Computational Linguistics.

Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021b. Probing for idiomaticity in vector space models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3551–3564, Online. Association for Computational Linguistics.

Fritz Günther, Marco Marelli, and Jens Bölte. 2020. Semantic transparency effects in German compounds: A large dataset and multiple-task investigation. *Behavior Research Methods*, 52(3):1208–1224.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Severin Laicher, Sinan Kurtyigit, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2021. Explaining and improving BERT performance on lexical semantic change detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 192–202, Online. Association for Computational Linguistics.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *Proceedings of ACL-08: HLT*, pages 236–244, Columbus, Ohio. Association for Computational Linguistics.

Diarmuid Ó Séaghdha. 2007. Annotating and learning compound noun semantics. In *Proceedings of the ACL 2007 Student Research Workshop*, pages 73–78, Prague, Czech Republic. Association for Computational Linguistics.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 210–218, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2014. Using distributional similarity of multi-way translations to predict multiword expression compositionality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 472–481, Gothenburg, Sweden. Association for Computational Linguistics.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2015. A word embedding approach to predicting the compositionality of multiword expressions. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 977–983, Denver, Colorado. Association for Computational Linguistics.

Roland Schäfer. 2015. Processing and querying large web corpora with the COW14 architecture. In *Proceedings of Challenges in the Management of Large Corpora 3 (CMLC-3)*, Lancaster.

Roland Schäfer and Felix Bildhauer. 2012. Building large corpora from the web using a new efficient tool chain. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 486–493, Istanbul, Turkey. European Language Resources Association.

Sabine Schulte im Walde, Anna Hätty, and Stefan Bott. 2016. The role of modifier and head properties in predicting the compositionality of English and German noun-noun compounds: A vector-space perspective. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 148–158, Berlin, Germany. Association 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).

Sabine Schulte im Walde, Stefan Müller, and Stefan Roller. 2013. Exploring vector space models to predict the compositionality of German noun-noun compounds. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 255–265, Atlanta, Georgia, USA. Association for Computational Linguistics.

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.

Claudia von der Heide and Susanne Borgwaldt. 2009. Assoziationen zu Unter-, Basis- und Oberbegriffen. Eine explorative Studie. In *Proceedings of the 9th Norddeutsches Linguistisches Kolloquium*, pages 51–74.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Licenses for data resources

The gold standard datasets of noun compounds and extensions with empirical properties are publicly distributed by their authors without a specific license. The ENCOW corpus and WordNet were acquired prior to this study. Their licenses do not have restrictions regarding research use.

B Effect of sampling differences

We found that resampling corpus occurrences (for $n = 10$ occurrences per compound) led to minor differences in model performance. However, they did not impact parameter-level trends overall, with the effect of parameter choices remaining the same in the majority of cases. To illustrate this, we are reporting the mean correlations with human judgment based on different samples of occurrences: three individual samples (those whose overall mean correlations are the lowest, the closest to the mean, and the highest), as well as the mean and standard deviation for the 10 samples. Similarly to the main analysis, the mean values are presented for different parameter choices; for layers, this is limited to a sample of layer spans. The results are split across the three prediction targets: for the compound as a whole in Table 9; for the head in Table 10; for the modifier in Table 11.

	Sample shuffles			Mean	Std
Seq. len.					
3	0.136	0.148	0.157	0.146	0.006
20	0.117	0.135	0.151	0.138	0.011
Agg.					
token	0.134	0.146	0.161	0.149	0.008
type	0.120	0.137	0.147	0.134	0.008
Pooling					
avg	0.126	0.141	0.153	0.141	0.008
sum	0.127	0.142	0.155	0.143	0.008
Estimate					
head	0.265	0.276	0.280	0.274	0.006
modif	0.132	0.134	0.145	0.136	0.006
comp	0.240	0.243	0.250	0.245	0.005
cont	0.156	0.172	0.194	0.180	0.013
cls	-0.153	-0.119	-0.105	-0.126	0.016
Layers					
0-0	0.196	0.198	0.195	0.200	0.006
1-1	0.118	0.139	0.153	0.136	0.010
11-11	-0.062	-0.036	-0.009	-0.040	0.016
12-12	0.179	0.209	0.226	0.200	0.014
1-4	0.310	0.321	0.317	0.318	0.006
8-12	-0.009	0.011	0.044	0.011	0.015

Table 9: Mean correlations for compositionality prediction (of the compound as a whole) across parameter choices.

	Sample shuffles			Mean	Std
Seq. len.					
3	0.095	0.098	0.120	0.104	0.009
20	0.083	0.086	0.106	0.095	0.010
Agg.					
token	0.094	0.097	0.123	0.106	0.010
type	0.083	0.087	0.103	0.093	0.008
Pooling					
avg	0.089	0.091	0.113	0.099	0.009
sum	0.089	0.092	0.113	0.100	0.009
Estimate					
head	0.238	0.240	0.249	0.241	0.006
modif	0.075	0.064	0.086	0.075	0.007
comp	0.196	0.187	0.202	0.195	0.005
cont	0.113	0.126	0.163	0.143	0.019
cls	-0.165	-0.148	-0.132	-0.152	0.013
Layers					
0-0	0.145	0.151	0.150	0.151	0.007
1-1	0.090	0.095	0.120	0.101	0.010
11-11	-0.055	-0.052	-0.013	-0.047	0.015
12-12	0.185	0.184	0.228	0.192	0.017
1-4	0.252	0.261	0.263	0.262	0.009
8-12	-0.015	-0.018	0.028	-0.006	0.015

Table 10: Mean correlations for compositionality prediction (of the head) across parameter choices.

	Sample shuffles			Mean	Std
Seq. len.					
3	0.097	0.103	0.091	0.095	0.006
20	0.079	0.095	0.098	0.091	0.009
Agg.					
token	0.091	0.099	0.098	0.096	0.005
type	0.085	0.099	0.092	0.090	0.007
Pooling					
avg	0.087	0.098	0.094	0.092	0.006
sum	0.088	0.099	0.096	0.094	0.006
Estimate					
head	0.160	0.170	0.167	0.166	0.004
modif	0.101	0.106	0.101	0.102	0.007
comp	0.158	0.164	0.160	0.162	0.006
cont	0.134	0.137	0.130	0.133	0.009
cls	-0.117	-0.090	-0.088	-0.101	0.013
Layers					
0-0	0.166	0.160	0.157	0.164	0.005
1-1	0.075	0.095	0.087	0.084	0.008
11-11	-0.070	-0.038	-0.045	-0.055	0.012
12-12	0.105	0.146	0.131	0.128	0.014
1-4	0.227	0.227	0.216	0.223	0.007
8-12	-0.029	0.000	-0.003	-0.016	0.011

Table 11: Mean correlations for compositionality prediction (of the modifier) across parameter choices.

C Effect of layer combinations

Model performance across layer combinations is plotted in Figure 4.

D Effect of compositionality estimates

Model performance across individual compositionality estimates is plotted in Figure 5.

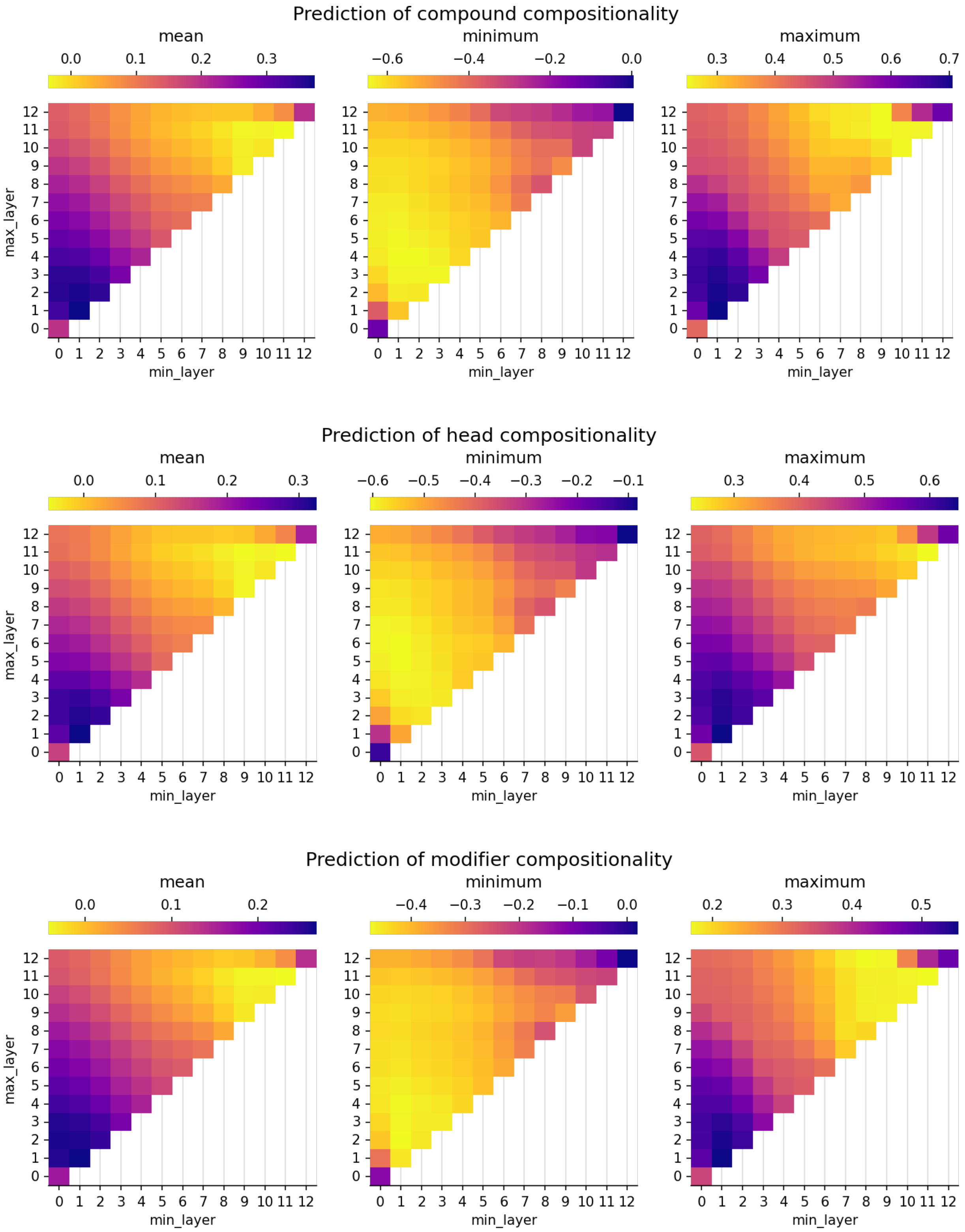

Figure 4: Correlations for compositionality prediction across layer combinations. The `min_layer` and `max_layer` values are start and end points of a contiguous span of layers. For each prediction target (compound, head, and modifier compositionality score), the left panel corresponds to the mean correlation per layer combination; the middle and the right panels correspond to the minimum and maximum values, respectively.

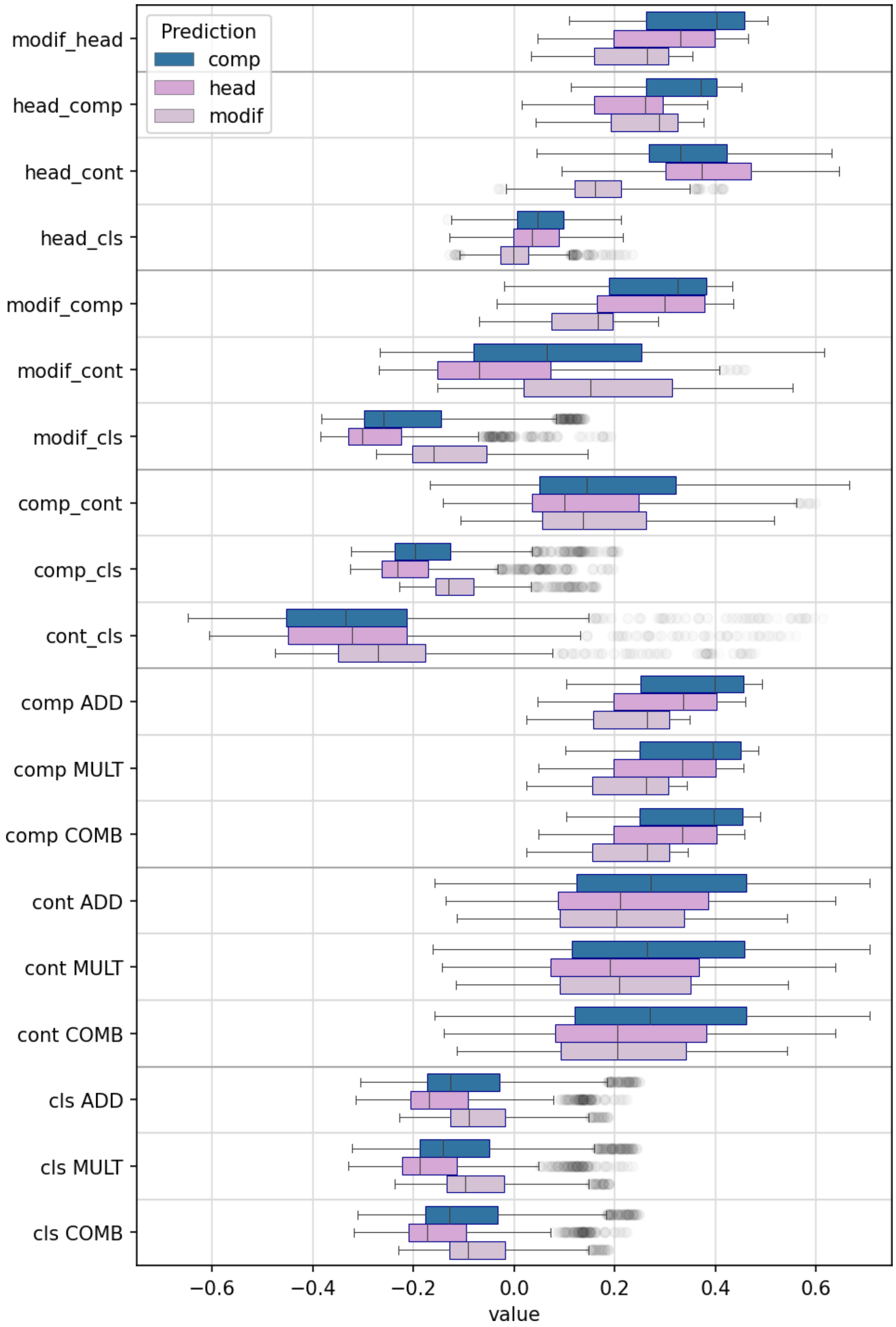

Figure 5: Distribution of Spearman's correlation coefficient for compositionality prediction across compositionality estimates. Embedding types: `modif` = contextualized representation of the modifier; `head` = contextualized representation of the head; `comp` = pooled representation of `modif` and `head`; `cont` = pooled representation of the surrounding context (full sequence without the compound); `cls` = representation of the `[CLS]` token. The use of ADD, MULT, and COMB involves the composite estimates described in Section 4.