A Systematic Search for Compound Semantics in Pretrained BERT Architectures

Filip Miletic & Sabine Schulte im Walde

Institute for Natural Language Processing, University of Stuttgart, Germany

{filip.miletic, schulte}@ims.uni-stuttgart.de

Introduction

Motivation

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

guinea pig *snail mail* *health insurance*

→ *more compositional*

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

Research questions

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

Experimental setup

Data and model

- 280 English compounds with human compositionality ratings (Cordeiro et al., 2019)
- Corpus data: ENCOW16 (Schäfer and Bildhauer, 2012; Schäfer, 2015)
- BERT-base-uncased, no fine-tuning (Devlin et al., 2019)

Modeling approach

For each compound:

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

Experimental parameters

Preprocessing

# sequences	10, 100, 1k
Seq. length	any, ≥ 20 tokens

Embedding computation

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

Compositionality estimation

Direct	pairwise cosine
Composite	ADD, MULT, COMB

⇒ 41,496 parameter constellations

Results

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

Performance range

ours		SOTA		
best	0.706	word2vec	0.726	(Cordeiro et al., 2019)
worst	-0.649	BERT	0.370	(Garcia et al., 2021)

Effects of modeled tokens

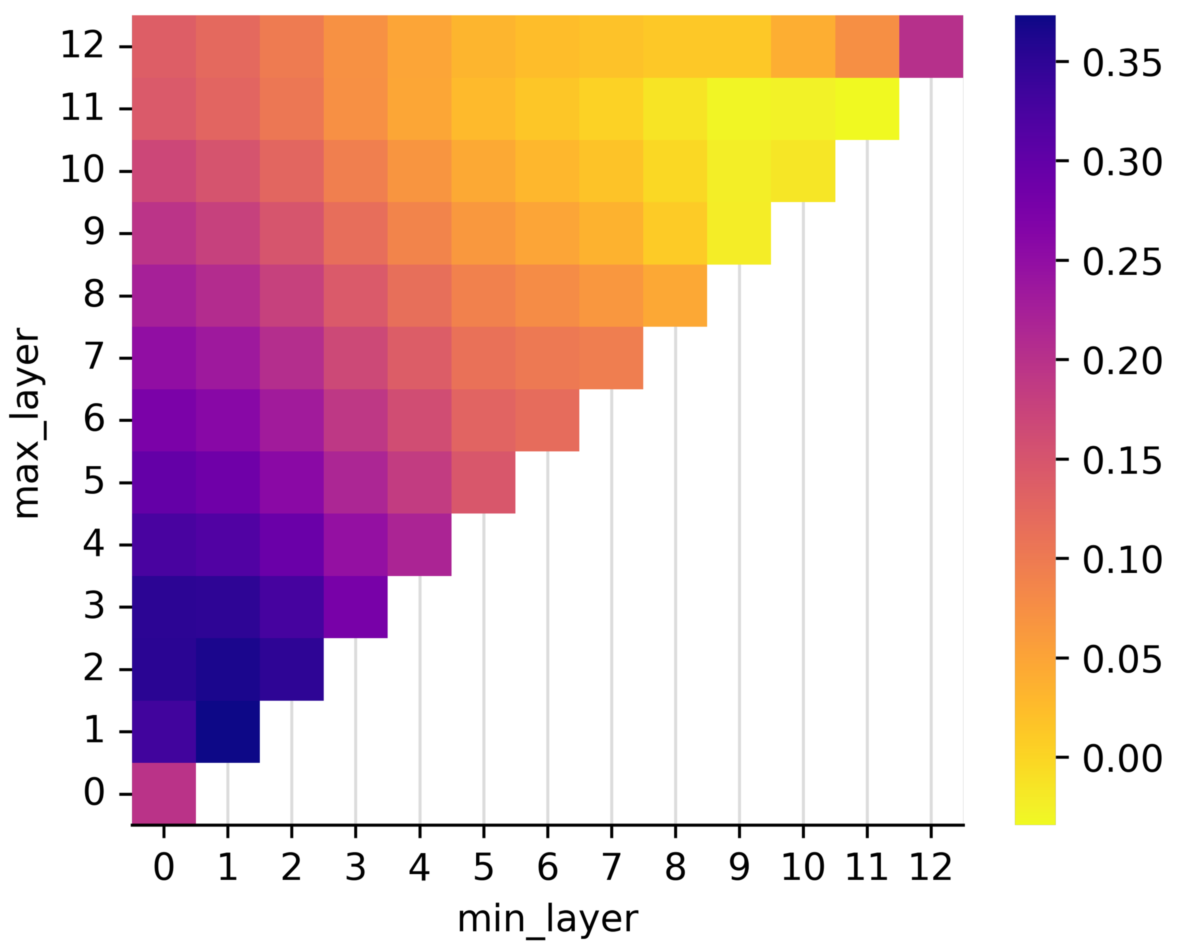
Best performance across prediction targets

prediction target	modeled token: modif	head	comp	cont	CLS
Comp	0.615	0.630	0.666	**0.706**	0.611
Head	0.464	**0.645**	0.598	**0.645**	0.558
Modif	**0.553**	0.415	0.517	**0.553**	0.477

Prediction targets: Comp, Head, Modif = degree of compositionality for the whole compound, the head, or the modifier, respectively.

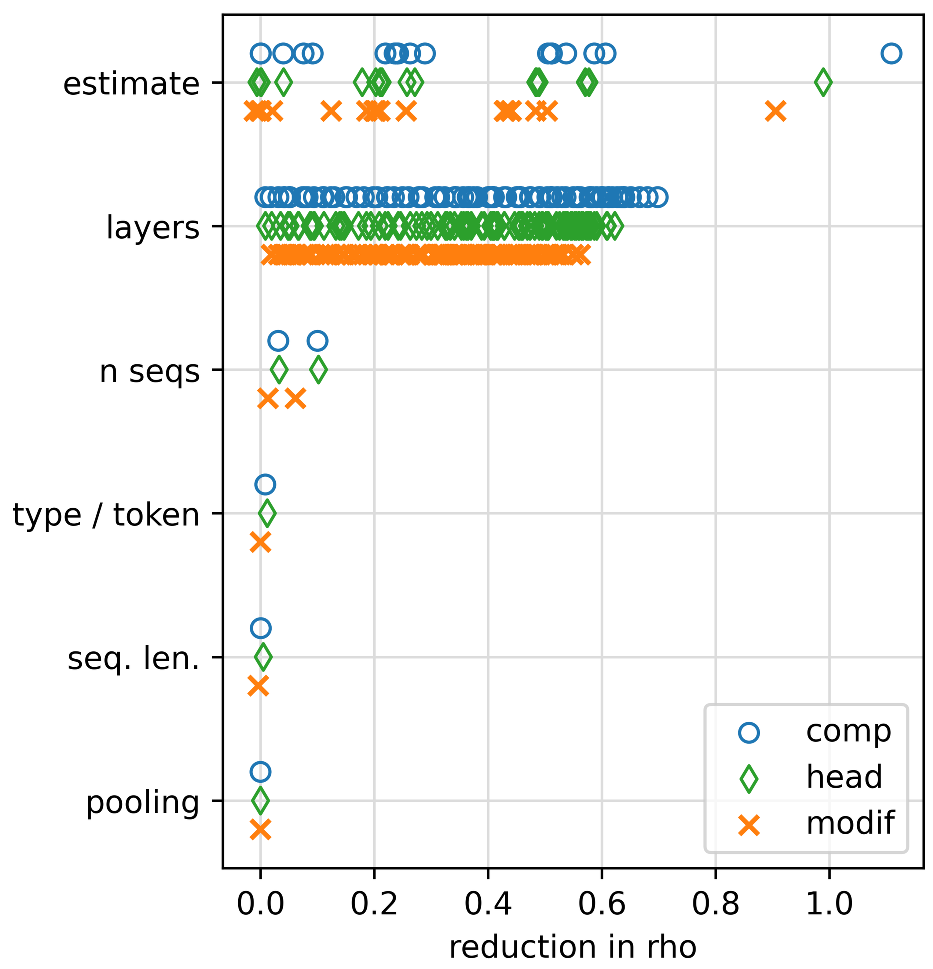
Effects of layer spans

Mean performance on compound-level compositionality



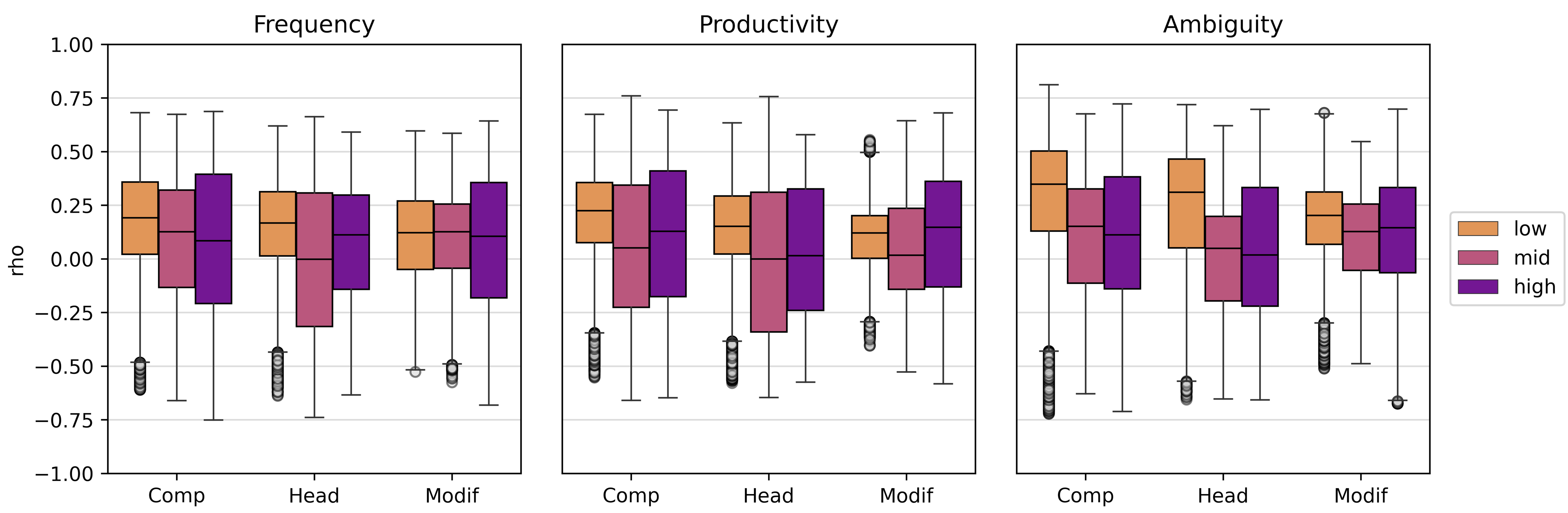
Ablation study

Effects of alternative parameters compared to best



Effects of empirical properties of compounds

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



Conclusions

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

References

Alipoor, P., & Schulte im Walde, S. (2020). Variants of vector space reductions for predicting the compositionality of English noun compounds. In *Proc. LREC*.
Cordeiro, S., Villavicencio, A., Idiart, M., & Ramisch, C. (2019). Unsupervised compositionality prediction of nominal compounds. *Computational Linguistics*, 45(1), 1–57.
Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional Transformers for language understanding. In *Proc. NAACL-HLT*.
Garcia, M., Kramer Vieira, T., Scarton, C., Idiart, M., & Villavicencio, A. (2021). Assessing the representations of idiomaticity in vector models with a noun compound dataset labeled at type and token levels. In *Proc. ACL-IJCNLP*.
Schäfer, R. (2015). Processing and querying large web corpora with the COW14 architecture. In *Proc. CMLC*.
Schäfer, R., & Bildhauer, F. (2012). Building large corpora from the web using a new efficient tool chain. In *Proc. LREC*.
Schulte im Walde, S., Hätty, A., & Bott, S. (2016). The role of modifier and head properties in predicting the compositionality of English and German noun-noun compounds: A vector-space perspective. In *Proc. *SEM*.

University of Stuttgart
Institute for Natural Language Processing

The 17th Conference of the European Chapter of the Association for Computational Linguistics

Dubrovnik, Croatia, May 2023