across subsequent layers in an almost monotonic fashion in the Compositional Probe experiment. In general, both *xlm-mlm-xnli15-1024* and *xlnet-base-cased* tend to represent fine-grained dissimilarity patterns (as measured with RSA) better in later layers. At the same time, the individual dimensions of vectors produced by these two models become less decodable in later layers. In the case of the BERT-style models, it is clear that the final layer is not the best for representing the thematic relation signal and we recommend exploring the use of middle layers for downstream tasks that require similar relational-semantic information.

5. Limitations and Future Work

One of the main limitations of our analysis was the relatively small size of the datasets we used compared to other datasets that are used for evaluation in natural language processing research. In this area we are limited by the lack of large, annotated noun-noun compound datasets, as the process of labeling noun-noun compounds with thematic relations is a time-consuming process for human annotators. This is particularly the case for the 18-dimensional setting, where every potentially applicable relation must be considered, rather than choosing one main relation. The fact that we can only use the 300 compound dataset to compare within each of the 60 groups means that we are limited to a total of 540 comparisons between compound representations using RSA, as the upper triangle of the ground truth RDM structure in Figure 4 (excluding the main diagonal and the compound pair marked in gray) allows for 9 comparisons within each group. We identify this annotation task as a key recommendation for future work for extending our analyses. While the 60 compound dataset provides a richer annotation of the underlying relation between head and modifier nouns, this dataset again allows for relatively few comparisons (1,770 pairs for RSA and the 2 vs. 2 test). In part due to the limited number of samples available to us, we chose to use data analysis techniques developed by researchers in the area of cognitive science, as this field is often limited by both the number of subjects that can be observed, and the number of stimuli that can be presented to a human subject within a single session.

Another area where our analysis is limited is the range of Transformer-based language models we investigated. While these models were chosen to cover a range of Transformer types, it is impossible to make generalizable judgments about certain classes of model (i.e., multilingual models or distilled models) based on our analysis, as we only feature one type of model in each of those classes. One exception is our analysis of *bert-base-uncased*, *roberta-base*, and *distilroberta-base*, which allows us to make generalizable statements about this class of Transformer. This is particularly true of the *bert-base-uncased* model, as we carry out an analysis on 25 different instantiations of this same class of model. In any case, future work should expand the analysis of how Transformers process noun-noun compounds to cover several models within each one of these areas. A related recommendation towards building a more robust analysis is to train several versions of the other five types of model, allowing for variation within a constrained architecture and choice of hyperparameters (McCoy, Min, and Linzen 2020).

As part of our experimental design, we do not consider the effect of fine-tuning Transformer-based language models. This choice allows us to probe Transformer-based language models for their capacity to automatically capture implicit semantic information about noun-noun compounds at the expense of limiting the generalizability of our findings in fine-tuning settings that are commonplace in the application of Transformer-based language models for solving downstream tasks. This consideration is particularly

relevant in the context of probing representations with simple linear models, as latent high-level semantic information about noun-noun compounds may require non-linear processing facilitated by fine-tuning several layers in order for this information to become available for linear regression models in the final layers. After juxtaposing the high RSA correlations found in later layers of *xlm-mlm-xnli15-1024* and *xlnet-base-cased* in the Relation Vector RSA experiments with the relatively low decodability scores for later layers of these two models seen in the Compositional Probe experiment, we would expect fine-tuning to be particularly useful for these models. One further consideration for expanding the analysis to allow for fine-tuning is that any sensitivity to noise incurred by a small amount of training samples may be amplified during this fine-tuning process, and thus the thematic relation vector dataset may need to be expanded before introducing this experimental extension. Another potential limitation with our experimental design is the choice to limit our RSA and probing tasks to token spans within the noun-noun compounds. As was seen in all experiments, Transformer-based language models tend to compose and distribute semantic information across many token vectors. Accordingly, it could be the case that information about the thematic relation between the head and modifier word could be distributed across tokens in other parts of the sentence, rather than just in tokens in the compound. A related concern to our choice of representation is the question of whether mean-pooling across tokens preserves thematic relation information and whether other approaches should be explored, such as concatenating a max-pooled vector with the mean-pooled token, or constructing a non-linear recurrent neural network probe to preserve all token vector information. In our experiments, we chose to consider a priori one choice of token representation that is most simply and most straightforwardly related to the compound (i.e., mean pooled token representations from within the compound), in order to avoid complications due to "researcher degrees of freedom" (Wicherts et al. 2016). Nevertheless, future work could investigate a wider range of possible model representations, to examine the extent to which these Transformer-based language models distribute information about compound relation semantics across the sentence, and whether such information is recoverable from a whole-sentence representation.

6. Conclusion

In this work, we used two English noun-noun compound datasets in order to probe Transformer-based language models for their knowledge of semantic relations between head and modifier nouns. To this end, we constructed three experiments to measure the representation of semantic relation information at a coarse and fine-grained level. In our layer-wise analysis, we find evidence that head-modifier thematic relation information is encoded in the token vector representations of six different Transformer-based language models. Of the six models we looked at, we find that the four English monolingual models strongly represent this information at both the coarse and fine-grained levels. Our compositional probe experiment shows that representations of these four models significantly benefit from head and modifier nouns being processed in the same context on a relation vector decoding task. Furthermore, we find evidence that these models gain significant levels of decodability from this concurrent compositional mode. These results suggest that the models that best encode relational information dynamically integrate their knowledge of the intrinsic properties of the head and modifier concepts in order to represent the semantic relation between these words, rather than only relying on distributional information of concept-relation frequency.

References

Abnar, Samira, Lisa Beinborn, Rochelle Choenni, and Willem Zuidema. 2019. Blackbox meets blackbox: Representational similarity & stability analysis of neural language models and brains. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 191–203. https://doi.org/10.18653/v1/W19-4820

Alishahi, Afra, Yonatan Belinkov, Grzegorz Chrupała, Dieuwke Hupkes, Yuval Pinter, and Hassan Sajjad. 2020. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*.

Anderson, Andrew James, Douwe Kiela, Jeffrey R. Binder, Leonardo Fernandino, Colin J. Humphries, Lisa L. Conant, Rajeev D. S. Raizada, Scott Grimm, and Edmund C. Lalor. 2021. Deep artificial neural networks reveal a distributed cortical network encoding propositional sentence-level meaning. *Journal of Neuroscience*, 41(18):4100–4119. https://doi.org/10.1523/JNEUROSCI.1152-20.2021, PubMed: 33753548

Baroni, Marco. 2020. Linguistic generalization and compositionality in modern artificial neural networks. *Philosophical Transactions of the Royal Society B*, 375(1791):20190307. https://doi.org/10.1098/rstb.2019.0307, PubMed: 31840578

Coil, Jordan and Vered Shwartz. 2023. From chocolate bunny to chocolate crocodile: Do Language models understand noun compounds? *arXiv preprint arXiv:2305.10568*. https://doi.org/10.18653/v1/2023.findings-acl.169

Csordás, Róbert, Kazuki Irie, and Juergen Schmidhuber. 2021. The devil is in the detail: Simple tricks improve systematic generalization of transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 619–634. https://doi.org/10.18653/v1/2021.emnlp-main.49

Devereux, Barry and Fintan Costello. 2005. Investigating the relations used in conceptual combination. *Artificial Intelligence Review*, 24(3–4):489–515. https://doi.org/10.1007/s10462-005-9007-5

Devereux, Barry and Fintan Costello. 2006. Modelling the interpretation and interpretation ease of noun-noun compounds using a relation space approach to compound meaning. In *28th Annual Conference of the Cognitive Science Society*.

Devereux, Barry J. and Fintan J. Costello. 2012. Learning to interpret novel noun-noun compounds: Evidence from category learning experiments. In *Cognitive Aspects of Computational Language Acquisition*. Springer, pages 199–234. https://doi.org/10.1007/978-3-642-31863-4_8

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *CoRR abs/1810.04805*.

Downing, Pamela. 1977. On the creation and use of English compound nouns. *Language*, pages 810–842. https://doi.org/10.2307/412913

Edgington, Eugene and Patrick Onghena. 2007. *Randomization Tests*. Chapman and Hall/CRC. https://doi.org/10.1201/9781420011814

Ester, Martin, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *The International Conference on Knowledge Discovery and Data Mining*, volume 96, pages 226–231.

Estes, Zachary and Uri Hasson. 2004. The importance of being nonalignable: A critical test of the structural alignment theory of similarity. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 30:1082–1092. https://doi.org/10.1037/0278-7393.30.5.1082, PubMed: 15355137

Ettinger, Allyson. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48. https://doi.org/10.1162/tacl_a_00298

Fares, Murhaf, Stephan Oepen, and Erik Velldal. 2018. Transfer and multi-task learning for noun–noun compound interpretation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1488–1498. https://doi.org/10.18653/v1/D18-1178

Gagné, Christina L. and E. J. Shoben. 1997. Influence of thematic relations on the comprehension of modifier-noun combinations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23:71–78. https://doi.org/10.1037//0278-7393.23.1.71