Probing for Hyperbole in Pre-Trained Language Models

Nina Skovgaard Schneidermann[1], **Daniel Hershcovich**[2] and
Bolette Sandford Pedersen[1]
[1]Center for Language Technology,
[2]Department of Computer Science
University of Copenhagen
ninasc@hum.ku.dk, dh@di.ku.dk, bspedersen@hum.ku.dk

Abstract

Hyperbole is a common figure of speech, which is under-explored in NLP research. In this study, we conduct edge and minimal description length (MDL) probing experiments for three pre-trained language models (PLMs) in an attempt to explore the extent to which hyperbolic information is encoded in these models. We use both word-in-context and sentence-level representations as model inputs as a basis for comparison. We also annotate 63 hyperbole sentences from the HYPO dataset according to an operational taxonomy to conduct an error analysis to explore the encoding of different hyperbole categories. Our results show that hyperbole is to a limited extent encoded in PLMs, and mostly in the final layers. They also indicate that hyperbolic information may be better encoded by the sentence-level representations, which, due to the pragmatic nature of hyperbole, may therefore provide a more accurate and informative representation in PLMs. Finally, the inter-annotator agreement for our annotations, a Cohen's Kappa of 0.339, suggest that the taxonomy categories may not be intuitive and need revision or simplification.

1 Introduction

Hyperbole is a common figure of speech that involves the use of exaggerated language for emphasis or effect (Claridge, 2010). Humans exaggerate in a variety of registers and contexts, spanning from the colouring of informal, everyday speech to a literary trope or a rhetorical means of persuasion. Hyperboles intentionally augment or diminish a feature of some referent of discourse, presenting this feature on some more or less abstract scale of magnitude. The task of hyperbole identification poses a challenge to natural language processing in that it is highly pragmatic and utilizes context and background knowledge to distinguish between literal and exaggerated usage of a given lexical unit. As an illustration of the pragmatic nature of hyperbole, we can inspect the following two example sentences, wherein (1A) is hyperbolic and (1B) is literal:

(1A) I've seen this movie *at least eighty thousand times*.

(1B) These products are tested *at least eighty thousand times*.

In (1A), it is reasonable to assume that the speaker is exaggerating the number of times they have seen this particular movie to emphasize their enjoyment or familiarity with it because this would otherwise be a significant and unrealistic time investment. However, when it comes to a particular product, it has likely gone through rigorous testing and quality control measures, which means that the statement in (1B) can reasonably be interpreted literally.

Hyperbole identification has recently attracted the interest of NLP researchers who have collected datasets manually or semi-automatically and shown that computational modelling of hyperbole is indeed plausible (Troiano et al., 2018). However, it remains an under-explored area of research in figurative language processing (FLP), primarily because its subjective and contextual nature complicates computational modelling of the phenomenon and makes it challenging to apply a standard for collecting high-quality annotated data (Biddle et al., 2021).

This paper seeks to contribute to the growing research on hyperbole identification in two ways: Firstly, we perform probing tasks to investigate whether pre-trained language models (PLMs) encode hyperbolic information in its representation without fine-tuning on task-specific data.[1] In recent years, probing tasks

[1]By "hyperbolic", we consistently refer to the figure of speech, not the mathematical space.

Proceedings of the 61st Annual Meeting of the
Association for Computational Linguistics - Student Research Workshop, pages 200–211

have emerged as a popular approach in NLP for interpreting and analyzing model representations, and it has previously been shown that PLMs do encode both simile and metaphorical knowledge (Chen et al., 2022). However, to our knowledge, hyperbole probing remains so far unexplored. Therefore, we replicate edge and minimal description length (MDL) probing experiments for metaphor described by Aghazadeh et al. (2022) on a small hyperbole dataset constructed by Troiano et al. (2018). We expect that encoding hyperbole may present a larger challenge to PLMs than metaphor because hyperbole knowledge is primarily pragmatic rather than semantic (McCarthy and Carter, 2004).

Secondly, we build an operational taxonomy based on a meta-analysis of the linguistic treatment of hyperbole, and annotate an existing dataset according to said taxonomy (McCarthy and Carter, 2004; Mora, 2009; Claridge, 2010; Burgers et al., 2016; Troiano et al., 2018). We then use these annotations to analyze errors in model predictions to further shed light on the types of hyperboles that may pose a particular challenge to PLMs, as well as when constructing training corpora for the phenomenon. Our work will hopefully provide insight into the challenges of PLMs in identifying hyperbole, as well as contribute to developing an operational annotation standard for computational modelling of hyperbole.[2]

The remainder of this paper is structured as follows: Section 2 contains an overview of related work in hyperbole research, as well as probing experiments on other figures of speech. Section 3 provides a background on the linguistic research that is the framework for our operational taxonomy and annotation. Section 4 is a short explanation of probing tasks for PLMs, which we relate to the aim of our experiments. Section 5 outlines our experimental setup and describes the modifications made to the HYPO dataset. Section 6 provides our results and preliminary error analysis, and section 7 is a discussion of said results, as well as ideas for future research. Section 8 contains a summary and conclusions.

[2]Our code for the probing tasks is available at https://github.com/NiSc91/HyperboleProbe

2 Related Work

In this section, we outline previous research related to both hyperbole and probing experiments on other figures of speech.

Hyperbole in NLP. While tropes such as metaphor and sarcasm have received considerable attention within figurative language processing research (Abulaish et al., 2020; Rai and Chakraverty, 2020; Moores and Mago, 2022), the automatic modelling of hyperbole is still at a relatively early stage. Research within this area can be roughly split into two objectives, hyperbole identification (HI) and hyperbole generation (HG).

Within the first, and for our purposes most interesting, category, Troiano et al. (2018) introduce the task of hyperbole detection by showing that classical machine learning pipelines can identify hyperboles with beyond-chance accuracy. For this purpose, they collect HYPO, the only manually constructed corpus of 709 English hyperboles, and include with the hyperbolic sentence s two contrasting corpora: One consisting of the manually constructed literal paraphrases to each of the sentences, and another consisting of a contrastive non-hyperbolic example using the same minimal lexical unit. They then identify a set of hand-crafted features targeting qualitative and quantitative aspects of exaggeration and report the best-performing classifier to be logistic regression using the literal paraphrases as negative examples, which achieves a 76% F1 score. In the same realm, Kong et al. (2020) address hyperbole detection using deep learning techniques on a constructed Chinese corpus and find that an LSTM with hand-crafted and embedding features produced superior results (85.4% accuracy). Biddle et al. (2021) construct a multitask learning classification architecture for hyperbole detection using a multitask BERT-based approach, wherein the model is fine-tuned on the HYPO dataset and takes the literal paraphrases as privileged information using triplet sampling. The authors find

that their model improves the logistic regression baseline described by Troiano et al. (2018) by 10%. The authors also devise a series of test sentences to linguistically probe their model for extreme case formulations (ECFs), quantitative, and qualitative hyperboles, as described by Mora (2009), and find that their model particularly excels at hyperboles containing ECFs, which may be due to the lexical substitution between the hyperbole and the literal paraphrase being minimal.

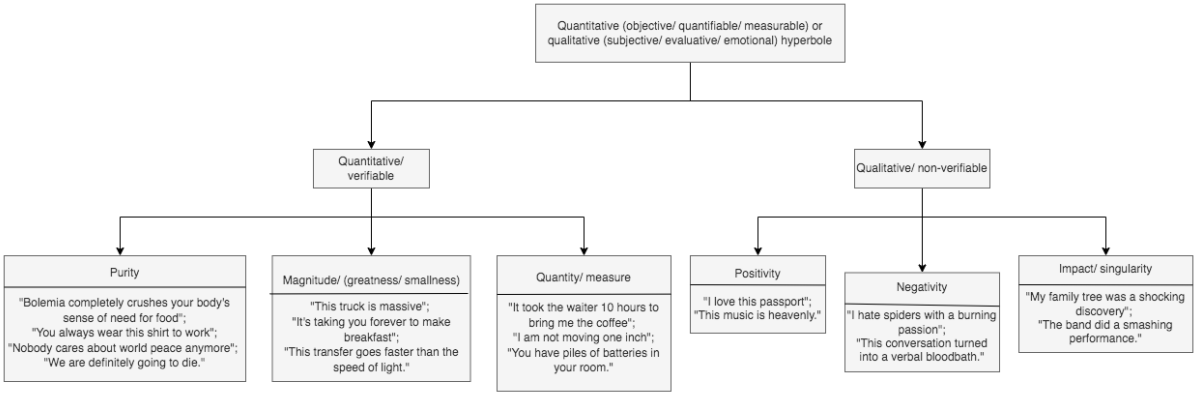
Recent frameworks have also leveraged pre-trained language models to generate hyperbole and expand on existing hyperbole data in a semi-supervised way. Specifically, Tian et al. (2021) construct a sentence-level hyperbole generation model by fine-tuning it on sentences from a Reddit corpus using the syntactic pattern known as the "so ... that" pattern, which is said to be a productive strategy for hyperbole (McCarthy and Carter, 2004). The authors annotate the data with semantic relationships within the sentence and feed the annotations to COMeT models (Bosselut et al., 2019) trained to generate commonsense and counterfactual inference. They then train a classifier to rank hyperbole candidates and use a paraphrase model to generalize to more syntactic patterns. An HG approach by Zhang and Wan (2021) involves constructing a large-scale hyperbole corpus, HypoXL, and proposes an unsupervised approach to hyperbole generation wherein a fine-tuned BART model is used to fill in masked hyperbolic spans.

While these efforts point towards the possibility of successfully training computational models for the task of identifying hyperbole, the research so far also has significant gaps: Firstly, hyperbole in NLP lacks a unifying definition or linguistically motivated formal theory to describe the phenomenon. This is reflected in a lack of a consistent annotation scheme and procedure for hyperbole identification in the available data, which makes hyperbole studies relatively far behind investigations of metaphor, where most annotated data use either the Metaphor Identification Procedure and its extensions (MIP/MIPVU; Group, 2007; Steen et al., 2019), or Conceptual Metaphor Theory (CMT; Lakoff and Johnson, 1980) as a procedure for annotation. This consistency of theoretical framework and annotation procedure makes it easier to perform experiments generalizing across languages and datasets. Secondly, limited attempts have been made to probe pre-trained language models on how well they encode hyperbole without any fine-tuning. This makes it unclear whether models simply reconstruct the hyperboles found in the fine-tuning objective, and how well the model is able to learn hyperbolic information in a zero-shot or few-shot setting.

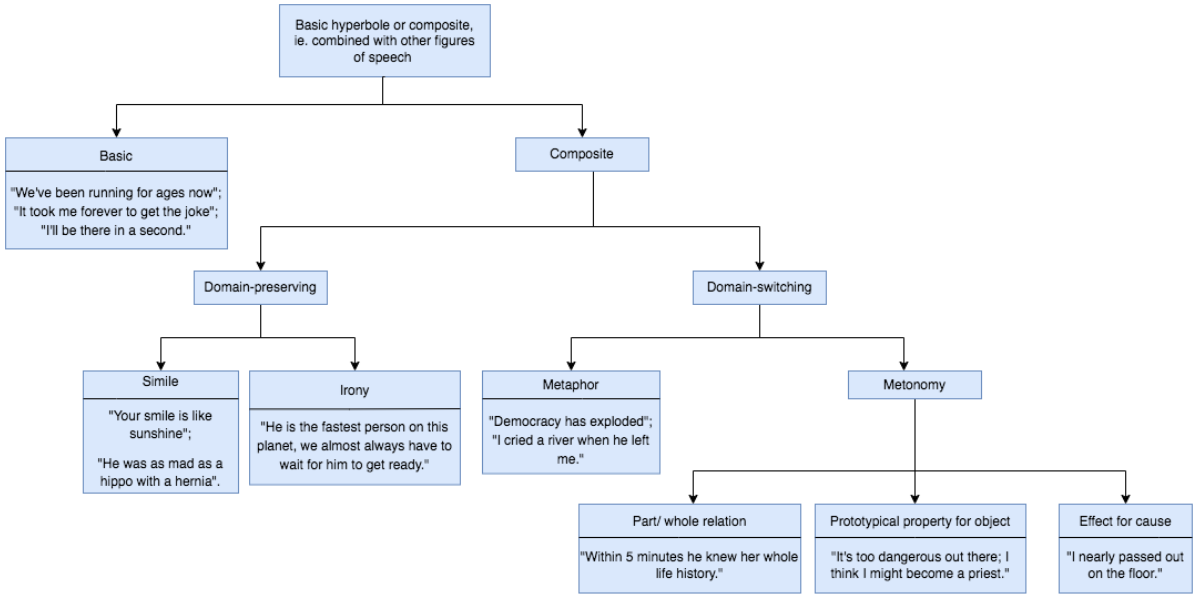
Our experiment is, to our knowledge, the first one to not utilize a fine-tuned model on hyperbolic sentences and to instead use probing methods to test for the encoding of hyperbolic information in PLMs.

Probing PLMs for Figurative Language Information. Probing techniques provide ways to understand and interpret the internal representations learned by deep neural networks (Belinkov, 2022). They typically involve extracting particular features or representations from a model's intermediate layers to gain insights into its structure or decision-making process. Several recent experiments have been designed to probe PLMs for information on figurative language. Namely, Chen et al. (2022) tackle similarity interpretation (SI) and generation (SG) tasks by probing simile knowledge from PLMs by testing it on similarity triple completion, i.e. sentences that take the form *[NP1] is as [ADJ] as [NP2]*. Their approach is to manually construct masked sentences with this syntactic pattern and predict the candidate words in the masked position. To that end, they adopt an auxiliary training process with the MLM loss to enhance the prediction diversity of candidate words. While this kind of probing works well to generate particular syntactic constructions, it would be ineffective for hyperbole due to its relatively limited dependence on syntax.

Instead, we choose to adapt several experiments conducted for metaphor probing by Aghazadeh et al. (2022) for hyperbole. The



(a) Subtree and examples for the Dimension category.



(b) Subtree and examples for the Type category.

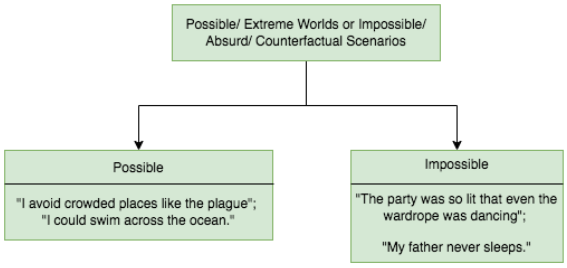
Figure 1: The first two categories in the proposed taxonomy for hyperbole with examples for each.

authors conduct probing in two ways: First, they train a linear probing classifier on 3 different PLMs to evaluate the accuracies and extractabilities with which they encode metaphorical knowledge. Secondly, they use MDL probing to analyze the depth of the encoding of metaphorical information in multi-layer representations. The authors further extend their experiment by generalizing across four datasets and four languages. The results suggest that contextual representations in PLMs do encode metaphorical knowledge, mostly in their middle layers, and that it is possible to transfer this information across languages and datasets provided the annotation is consistent across training and testing sets.

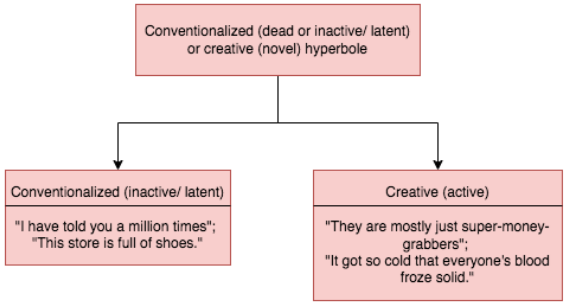
While we can replicate the basic probing experiments, we cannot test the model's generalizability given the scarce hyperbole data. However, we do expect that it is possible via these techniques to learn something about the internal representations of hyperbole.

3 A Taxonomy for Hyperbole

In simple terms, hyperbole involves exaggerating a feature's property X beyond what is justified by the literal state of affairs (Claridge, 2010; Troiano et al., 2018). Stated in a more discourse-centred way, hyperbole occurs when an expression is more extreme than justified given the ontological referent, i.e. the entity in the world referenced by the text (Burgers et al.,



(a) Subtree and examples for the Possibility category.



(b) Subtree and examples for the Conventionality category.

Figure 2: The last two categories in the taxonomy.

2016). While much of the work on hyperbole has previously been subsumed under studies of metaphor, humour, and verbal irony, recent corpus linguistic analyses have shed light on more fine-grained characteristics. Namely, the consensus in the treatment of hyperbole in literature is that the phenomenon is, among others, characterized by the presence of extreme case formulations (ECF), the ability of hyperbole to create either extreme possible worlds or downright counterfactual and absurd scenarios, and its augmentation of some property along a qualitative or quantitative scale (McCarthy and Carter, 2004; Mora, 2009; Claridge, 2010).

In the following, we outline some of the key characteristics and visualize them in an operational taxonomy (see Figures 1 and 2).

Dimension. There is widespread agreement that hyperbole occurs on a scale of magnitude along two main dimensions: a quantitative scale and a qualitative scale (Mora, 2009; Claridge, 2010; Troiano et al., 2018). The distinction between these scales refers to whether a hyperbole primarily concerns objective and measurable aspects or subjective and evaluative emotional states of affairs. According to Mora (2009), who conducted a corpus analysis of natural conversation on a 52000 word subset of the British National Corpus (BNC), quantitative hyperboles comprise 61% of the analyzed hyperboles and include the semantic fields of completeness, universality, measure, and magnitude. Qualitative (evaluative) hyperboles concern positive or negative sentiments, as well as impact or singularity; e.g. 'shocking', 'smashing' etc. However, an important point to make here is that there is a significant overlap between these dimensions, as hyperboles will generally have an evaluative function: For instance, the expression that somebody has "piles of batteries in their room" could be said to be a negative evaluation of the state of the room, but we choose to annotate such expressions as primarily quantitative, as the exaggerated property is one of measure. Another potentially relevant distinction is that quantitative hyperboles have a verifiable element, whereas purely qualitative hyperboles often serve to convey an internal subjective mental or emotional state (Claridge, 2010): For instance, in the statement, *It was the worst meal I have ever had*, the speaker could either be conveying their honest opinion of the meal, or they could be using exaggeration as a figure of speech to emphasize their disappointment with the meal.

Type. We use the term "type" to refer to whether the hyperbole is basic or composite, i.e., whether it stands alone or is combined with another figure of speech. According to Claridge (2010), hyperboles are basic if they preserve the semantic domain of the corresponding literal paraphrase, and composite if it involves a domain transfer where elements of a source domain is mapped onto a target domain. The latter is primarily the case with metaphor and, to a lesser extent, metonymy <citeclaridge2010hyperbole. In our annotations, we analyze simile as domain-preserving, even though we recognize that simile can be analyzed as an explicit metaphor (Burgers et al., 2018).

Degree of possibility. This distinction is one of degree and refers to the extent to which hy-

perboles generate impossible, absurd, or counterfactual scenarios. This is purely pragmatic and influences the degree to which a statement may be perceived as hyperbolic (McCarthy and Carter, 2004; Troiano et al., 2018).

Level of conventionality. This last dichotomy refers to the fact that hyperboles can use either more conventional or more novel and creative language to express exaggeration. This also impacts the extent to which a statement is perceived as a hyperbole: For instance, to say that one has not seen a person *for ages* is so frequent that it could be considered a latent or dead hyperbole, in the sense that it might not be viewed as intentional exaggeration for a specific purpose (McCarthy and Carter, 2004). However, in our annotation, we do label such frequent sentences as hyperbolic, although a conventionalized one.

4 Probing PLMs for Hyperbole

Probing language models aims to answer questions related to the model's internal representation, such as the location and depth of the encoding of a linguistic property in the multi-layer representation, or which input features contributed to a particular behaviour of the PLM (Belinkov, 2022). Standard probing methods involve training a linear classifier on top of a PLM to predict a linguistic property of interest, where a high probing performance on the task is associated with the model encoding said property. It is common practice to freeze the parameters of the PLM, which serves to prevent the gradients of the probing classifier from back-propagating into the model and thereby altering its pre-trained representation (Tenney et al., 2019). Following Aghazadeh et al. (2022), our experiments are not aimed at improving the accuracy of hyperbole identification tasks; we simply want to check the extent to which hyperbole knowledge may be encoded in the base representations. To that end, we employ edge probing, in which the classifier receives span-level representations from the PLM as inputs after they have been projected to a fixed-dimensional layer, 250 in this case. Thus, we define the span input to the PLM as the minimal lexical unit conveying hyperbolic information as given by the HYPO dataset (Troiano et al., 2018).

One common criticism of edge probing is that it may not be explanatory in the sense that it does not provide insight into whether a model is learning a linguistic property or simply memorizing the task (Belinkov, 2022). An information-theoretic perspective on addressing this limitation is to combine the probing quality of the classifier with some metric of the effort needed to extract the linguistic knowledge. This approach is known as MDL probing (Voita and Titov, 2020), wherein effort intuitively refers to the number of steps required by the PLM to encode a compressed representation of the input sequence. Following Aghazadeh et al. (2022), we use the online coding implementation of MDL, which measures a representation's ability to learn from various portions of the data. We report the compression, which is given by $N \cdot \log_2(K)$. In the context of language modelling, N refers to the size of the dataset, and K is the set of unique sequences being compressed. A random classifier will have a compression of 1, and increased data compression is associated with a better encoding of the given property.

5 Experiments

Here we describe our data and setup.

Dataset and annotation. We utilize HYPO, a manually constructed English hyperbole dataset (Troiano et al., 2018) of 709 hyperboles with corresponding literal paraphrases, as well as a *minimal units corpus* that provides the contrastive negative (literal) examples for each hyperbole (see examples (1A) and (1B) in §1).

For the purpose of our experiment, we first discard the corpus of literal paraphrases as we are interested in contrasting the hyperbolic usage of a particular word or phrase with a literal usage of the same word or phrase. It would otherwise not be possible to construct spans. To obtain span labels for each hyperbole and its negative contrast sentence, we programmatically extract the positions of each minimal

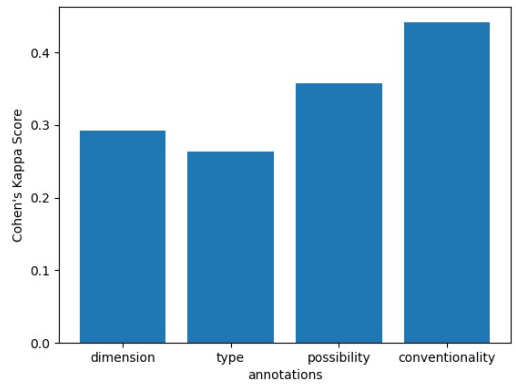


Figure 3: Inter-annotator agreement for the four aspects.

lexical unit and manually adapt the labels as needed; namely, we exclude examples with multiple spans and those without minimal unit contrasts.[3] Our final dataset contains 1396 span-labelled hyperbolic and literal sentences, which we split into training (70%), test (20%), and development (10%) sets.

We meticulously annotate the 63 hyperbolic sentences in the development sample using the operative taxonomy outlined in §3.[4] In order to obtain inter-annotator agreement, we enliste the help of additionally 5 annotators, assigning 12-13 sentences to each. As a result, each sentence is annotated twice. We observe a mean Cohen's Kappa of 0.339 (see Figure 3), suggesting only fair agreement, with particular difficulties on the dimension and type spectra on the taxonomy.

Experimental setup. We conduct edge- and MDL probing experiments for three models, BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), and Electra (Clark et al., 2020). Following Aghazadeh et al. (2022), all the models are initiated from the base versions of the Huggingface Transformer library (Wolf et al., 2020), with 12 layers, 768 hidden size, and 110m parameters. In line with the procedure described in detail by by Tenney et al. (2019), we use the contextual vector representation for each span as inputs to the model, followed by a projection-layer and self-attention pooling to collapse the span vectors down to a fix-length 256-dimensional representation. The edge probing classifier, which in this case is a single linear layer, is then trained on top of the PLM. We do not change the original hyperparameters; we keep the batch size of 32 and the learning rate of $5e-5$, and train over 5 epochs for each experiment. During model training, the development set is used to monitor the model's performance and as a stopping criterion at each epoch. The MDL probe is based on the same structure as the edge probing experiment (Aghazadeh et al., 2022). One minor change we make to accommodate the small size of our data is to delete the smallest fraction trained on by the MDL probe, as it would otherwise amount to a single example. We run our experiments in two configurations: One in which we use the manually labelled hyperbole spans as inputs to the PLM, which follows the classic edge probing procedure. We call this the word-in-context (WiC) representation to emphasize that the model only has access to the rest of the sentence through the context embeddings (Tenney et al., 2019). In the other configuration, which is used as basis for comparison, we feed the entire sentence span to the model - the so-called sentence-level configuration.

[3] See examples in Appendix A.

[4] Similar fine-grained annotations were conducted by citet-troiano2018computational, although they weren't included in the HYPO dataset, and inter-annotator agreement were not measured due to expected degree of difficulty.

Experiment	Word-in-Context Accuracy	Word-in-Context μ-F1	Sentence Level Accuracy	Sentence Level μ-F1
BERT	0.69	0.6895	0.72	0.7184
RoBERTa	0.72	0.7220	0.78	0.7762
ELECTRA	0.73	0.7256	0.78	0.7761

Table 1: Edge probing classification results.

6 Results

All our results are reported on the test set.

Edge probing results. The edge probing classification results are in Table 1 and the classification scores for the hyperboles and the literal sentences are in Table 2. We only report last layer scores, as we just evaluate the base representations.

Experiment	Class	Precision	Recall	F1
Word-in-Context				
BERT	literal	0.70	0.66	0.68
	nonliteral	0.68	0.72	0.70
RoBERTa	literal	0.73	0.71	0.72
	nonliteral	0.71	0.73	0.72
Electra	literal	0.74	0.71	0.72
	nonliteral	0.72	0.74	0.73
Sentence Level				
BERT	literal	0.78	0.61	0.69
	nonliteral	0.68	0.82	0.74
RoBERTa	literal	0.80	0.74	0.77
	nonliteral	0.75	0.82	0.78
ELECTRA	literal	0.84	0.69	0.76
	nonliteral	0.73	0.87	0.79

Table 2: Performance metrics for each of the models.

Annotation	WiC	Sentence	Total
QUAL	0.784	0.865	37
QUANT	0.692	0.731	26
PDOM	0.676	0.765	34
SDOM	0.828	0.862	29
NPOSS	0.769	0.821	39
POSS	0.708	0.792	24
CONV	0.806	0.806	36
NCONV	0.667	0.815	27

Table 3: Recall for word-in-context and sentence-level annotations for each category.

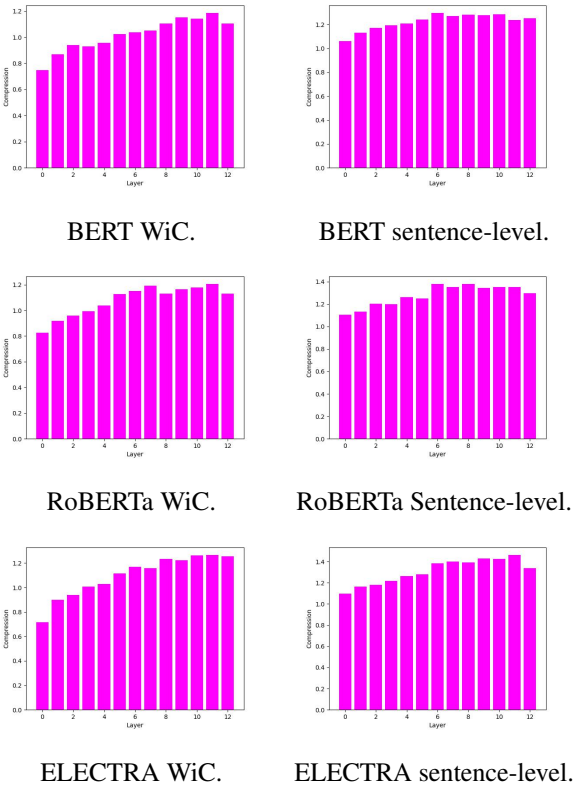


Figure 4: Compression for each of the models.

MDL probing results. We report the compression for each of the experiments in Figure 4. The best layer is consistently near the top layer, but not the top layer itself.

Error analysis. Our error analysis is conducted for the model with the best recall, RoBERTa, and is only conducted for the hyperbolic examples, i.e. the 63 annotated hyperboles in the development set. We choose the best layer based on the compression displayed in Figure 4; i.e. layer 11 for the WiC representation and layer 8 for the sentence-level representation.

Table 3 report the recalls, i.e. the percentages of correctly predicted hyperboles, for each of the annotated categories, for both of our experiments, along with the distributions of each of the annotations on the 63 samples.

7 Discussion

We observe notably lower scores than for the metaphor probing experiments across the board: Based on the compression reported for the MDL probes, only reaching up to 1.4 in the best configuration, we can conclude that hyperbolic information does appear to a minor extent to be encoded in PLM representations. This is in line with our expected hypothesis that encoding hyperbole may pose a bigger challenge given its primarily pragmatic nature, and also fits with the fact that PLMs have been reported to struggle with pragmatic inference and commonsense knowledge (Rogers et al., 2020). Perhaps more interestingly, we can inspect the compression for each of the 12 layers reported in Figure 4 to understand where hyperbole is best encoded by the representation, which appears to mostly be in the final layers. This is different from metaphor and may lend further credence to the idea that pragmatics is typically encoded deeper into the PLM. However, since we are employing a very small dataset, the extent to which we can draw definite conclusions is limited. In the future, we would like to extend our experiments to more data and languages to measure generalizability.

Upon analyzing the MDL compressions of the two model representations, we make an intriguing observation that the sentence-level representation consistently outperforms the WiC representation, with compressions reaching up to 1.4 for the top layer. This discovery raises thought-provoking questions about the amount of hyperbole information inferred by the contextual embeddings, as hyperbole often surpasses the token or phrase level. For example, consider the sentence, "The temperature was so low, I saw polar bears wearing jackets." In this case, the entire complement sentence creates the hyperbole. This leads to discussions about defining the lexical unit of hyperboles for corpus collection and annotation purposes (Burgers et al., 2016). As for the model representations themselves, while PLMs theoretically encode context in their representation, it is worth exploring how much information is contained within and between subwords in the WiC representation. Employing interpretability metrics could provide further insights into this matter.

Considering the low inter-annotator agreement and that recall seems to generally increase with the frequency of the subcategory in the sample, it is challenging to draw insights from the model error analysis (see Table 3). However, we may tentatively conclude that the models have an easier time with conventional hyperboles, which is the opposite finding to that of Troiano et al. (2018) for traditional machine learning pipelines. Similarly surprisingly is it that the PLMs have better recall for domain-switching hyperboles than domain-preserving ones, which may also be confounded by a strength variable. Furthermore, when manually expecting the false positives, we observe that some sentences predicted to be hyperbolic do indeed contain words and phrases with a potential hyperbolic interpretation, e.g. *paradise* in the sentence "He thought a place awaited him in paradise",, suggesting that analyzing hyperbole in a larger context might provide further insights.

Finally, the low inter-annotator agreement, particularly on the dimension and type dichotomies, suggests that the hyperbole categories are not intuitively well-understood or discriminated. During discussions with annotators upon completion of the task, we had several instances where overlap of the dimension subcategories was so large that annotators could argue for either one, and it also wasn't clear to annotators when a semantic domain-switch was present. The latter suggests that more linguistic training may be necessary to identify combined figures of speech in context, for instance, through application of the hyperbole identification procedure (HIP) (Burgers et al., 2016). As a consequence, we would like to change our approach to hyperbole annotation in future corpus construction and investigate to which extent these categories are indeed computationally relevant. Our negative findings lend credence to the claim by Biddle et al. (2021) that annotation schemes may present a bottleneck for further development of of the task. We would also like to explore approaches for model evaluation of hyperbole types using conceptual knowledge bases and linguistic resources; namely leveraging framenets to explore their utility for metaphorical hyperboles, as well as investigating templates using particular syntactic patterns for evaluating quantitative hyperboles.

8 Conclusions

This study has attempted to probe three pretrained language models (PLMs) for hyperbolic knowledge to better inspect how this information is encoded in their representations. We find, predictably, that knowledge of hyperbole is only to a limited extend encoded by PLMs, and, somewhat more surprisingly, that sentence-level representations appear to be supperior to word-in-context (WiC) representations, which may further highlight that most hyperbolic information does in fact exist beyond the token or phrase level. In the future, we would like to contribute with more hyperbole data with an operational annotation procedure, extend to cross-lingual experiments, as well as investigate the role of linguistic resources for hyperbole identification.

References

Muhammad Abulaish, Ashraf Kamal, and Mohammed J Zaki. 2020. A survey of figurative language and its computational detection in online social networks. *ACM Transactions on the Web (TWEB)*, 14(1):1–52.

Ehsan Aghazadeh, Mohsen Fayyaz, and Yadollah Yaghoobzadeh. 2022. Metaphors in Pre-Trained Language Models: Probing and Generalization Across Datasets and Languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2037–2050, Dublin, Ireland. Association for Computational Linguistics.

Yonatan Belinkov. 2022. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics*, 48(1):207–219.

Rhys Biddle, Maciek Rybinski, Qian Li, Cecile Paris, and Guandong Xu. 2021. Harnessing privileged information for hyperbole detection. In *Proceedings of the the 19th Annual Workshop of the Australasian Language Technology Association*, pages 58–67.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Christian Burgers, Britta C Brugman, Kiki Y Renardel de Lavalette, and Gerard J Steen. 2016. HIP: A method for linguistic hyperbole identification in discourse. *Metaphor and Symbol*, 31(3):163–178.

Christian Burgers, Kiki Y Renardel de Lavalette, and Gerard J Steen. 2018. Metaphor, hyperbole, and irony: Uses in isolation and in combination in written discourse. *Journal of Pragmatics*, 127:71–83.

Weijie Chen, Yongzhu Chang, Rongsheng Zhang, Jiashu Pu, Guandan Chen, Le Zhang, Yadong Xi, Yijiang Chen, and Chang Su. 2022. Probing Simile Knowledge from Pre-trained Language Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5875–5887, Dublin, Ireland. Association for Computational Linguistics.

Claudia Claridge. 2010. *Hyperbole in English: A Corpus-Based Study of Exaggeration*. Cambridge University Press.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-Training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Pragglejaz Group. 2007. MIP: A Method for Identifying Metaphorically Used Words in Discourse. *Metaphor and Symbol*, 22(1):1–39.

Li Kong, Chuanyi Li, Jidong Ge, Bin Luo, and Vincent Ng. 2020. Identifying Exaggerated Language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7024–7034, Online. Association for Computational Linguistics.

George Lakoff and Mark Johnson. 1980. Conceptual metaphor in everyday language. *The journal of Philosophy*, 77(8):453–486.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach.

Michael McCarthy and Ronald Carter. 2004. "There's millions of them": Hyperbole in everyday conversation. *Journal of Pragmatics*, 36(2):149–184.

Bleau Moores and Vijay Mago. 2022. A survey on automated sarcasm detection on twitter. *arXiv preprint arXiv:2202.02516*.

Laura Cano Mora. 2009. All or nothing: A semantic analysis of hyperbole. *Revista de Lingüística y Lenguas Aplicadas*, 4(1):25–35.

Sunny Rai and Shampa Chakraverty. 2020. A survey on computational metaphor processing. *ACM Computing Surveys (CSUR)*, 53(2):1–37.

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.

Gerard Steen, Aletta G Dorst, J Berenike Herrmann, Anna Kaal, Tina Krennmayr, and Tryntje Pasma. 2019. Mipvu: A manual for identifying metaphor-related words. *Metaphor identification in multiple languages: MIPVU around the world*, pages 24–40.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Yufei Tian, Arvind krishna Sridhar, and Nanyun Peng. 2021. HypoGen: Hyperbole Generation with Commonsense and Counterfactual Knowledge.

Enrica Troiano, Carlo Strapparava, Gözde Özbal, and Serra Sinem Tekiroğlu. 2018. A computational exploration of exaggeration.

Elena Voita and Ivan Titov. 2020. Information-Theoretic Probing with Minimum Description Length.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi 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 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.

Yunxiang Zhang and Xiaojun Wan. 2021. MOVER: Mask, over-generate and rank for hyperbole generation. *arXiv preprint arXiv:2109.07726*.

A Fine-grained Annotation Examples

Table 4 shows example data, along with the spans and annotations (taken from the development set of the data). The annotations are constructed along dimension (QUANT/QUAL), type (PDOM/SDOM), possibility (POSS/NPOSS), and conventionality (CONV/NCONV).

Hyperbole	**Literal**	**Dim.**	**Type**	**Poss.**	**Conv.**
Marriage is the *grave* of love.	I have gone to visit the grave of a friend.	QUAL	SDOM	NPOSS	CONV
So much snow that it is like walking in the *firmament*.	Some stars in the firmament have a name.	QUANT	PDOM	NPOSS	NCONV
The ancient castle was so big that it took *a week* to walk from one end to the other.	It took a week to walk from one end of the region to the other.	QUANT	PDOM	POSS	CONV
His feet are *colder than the arctic*.	The Antarctic is colder than the Arctic.	QUANT	PDOM	NPOSS	NCONV

Table 4: Sample data with annotations. Token spans are marked by italics around the word or phrase.