A howling success or a working sea? Testing what BERT knows about metaphors

Paolo Pedinotti and **Eliana Di Palma** and **Ludovica Cerini** and **Alessandro Lenci**
University of Pisa
paolo.pedinotti@phd.unipi.it, e.dipalma@studenti.unipi.it,
ludovica.cerini@phd.unipi.it, alessandro.lenci@unipi.it

Abstract

Metaphor is a widespread linguistic and cognitive phenomenon that is ruled by mechanisms which have received attention in the literature. Transformer Language Models such as BERT have brought improvements in metaphor-related tasks. However, they have been used only in application contexts, while their knowledge of the phenomenon has not been analyzed. To test what BERT knows about metaphors, we challenge it on a new dataset that we designed to test various aspects of this phenomenon such as variations in linguistic structure, variations in conventionality, the boundaries of the plausibility of a metaphor and the interpretations that we attribute to metaphoric expressions. Results bring out some tendencies that suggest that the model can reproduce some human intuitions about metaphors.

1 Introduction

Metaphor is a blooming affair. Since the publication of Lakoff and Johnson (1980) *Metaphors we live by*, it has been shown that metaphors represent a core cognitive mechanism, identified as a process that aids human beings in the comprehension of abstract concepts. Metaphors could be described as a process to endow linguistic expressions with new meaning: A concept of a *target* domain A is understood in terms of a *source* domain B.

Metaphors are pervasive in language, and they are a complex phenomenon to describe and categorize. We can distinguish metaphors for their degree of conventionalization, their linguistic structure (e.g., "A is B", "A of B", etc.), and for the semantic effect they create (Newmark, 1980; Charteris-Black, 2004), namely concretizing metaphors, animating and personifying metaphors, and synaesthetic metaphors. All these categories can tell us something about the degree of metaphoricity that an expression conveys. Many psycholinguistic studies have been carried out to understand how metaphors are perceived by humans and to what extent a metaphor is recognized as such (Lai et al., 2009; Glucksberg, 2003; Al-Azary and Buchanan, 2017).

A key aspect is surely the distinction between **conventional** metaphors (e.g., *Her lawyer is a shark*) and **novel** (or **creative**, Birdsell (2018)) metaphors (e.g., *Her mouth was a fountain of delight.*). These types of metaphors are processed differently by humans (Glucksberg et al., 1982; Gentner and Bowdle, 2008). Moreover, the latter constitutes an open class, as we have the ability to create new metaphors and make sense out of them. However, this ability is not erratic, but it is ruled by cognitive mechanisms: For example, metaphors have been shown to be justified by analogies between conceptual domains. This ability has also clear boundaries: Due to the same mechanisms we can evaluate the plausibility of a given expression, and distinguish creative meanings from nonsense expressions. For instance, even if both *An ambassador is a peacock* and *An ambassador is a fish* constitute semantic violations, we accept only the first as a plausible metaphor. Another example are the sentences *The wind was a howling wolf* and *The wind was a jumping seat* (McGregor et al., 2019).

Transformer Language Models such as BERT (Devlin et al., 2019) have brought important improvements in metaphor-related tasks (see section 2). However, such models have been used only in application contexts, while analysis aimed at investigating directly what the models capture about this phenomenon has not been conducted. In the first part of our work, we tested whether BERT predicts metaphors differing for linguistic structure and conventionality, with a particular focus on whether the model is able to identify the boundaries of metaphoric creativity. Of particular relevance for the second aspect is the comparison of novel, unconventional metaphors with nonsense expressions.

Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, pages 192–204

Our ability with figurative language also allows us to assign new meanings to words when they are used metaphorically, independently from the conventionality of a metaphor. Therefore, we expect to find information about these meanings in the model's representations of different types of metaphors.

To test these aspects, we collected a dataset of conventional and creative metaphors, matched with control literal and nonsense sentences, evaluated by English native speakers for their degree of metaphoricity and semantic plausibility. We carried out two experiments. In **Experiment 1** we modeled the dataset of human judgments using BERT as a **language model**, to evaluate whether its pseudo-log-likelihood values correlate with the human semantic plausibility scores. Our results show that BERT seems to distinguish the literal, metaphorical and nonsense expressions. In **Experiment 2**, we used the **landmark method** introduced by Kintsch (2000) to test the representation of metaphorical meanings in BERT contextualized embeddings. Despite limitations given by the size of our dataset, we show that some consistent trends about how the model processes metaphors can be identified by analyzing the model's representations. We observed that several factors such as the layer and the representations analyzed could influence the model's performance.

To sum up, we collected various evidence related to what BERT learns about metaphorical language. Such results pave the way for future investigation, and can be of interest for those who use Transformers Language Models in NLP applications or in metaphor-related tasks.

2 Related works

The computational literature on metaphors has focused on two distinct tasks. Metaphor identification involves deciding whether a sequence or a single word is an instance of a metaphor or not. On the other hand, Metaphor interpretation concerns the description of the meaning of metaphorical expressions, and is typically cast as a paraphrasing (Shutova, 2010) or a classification (Bizzoni and Lappin, 2017) task. Even if some recent work has approached the interpretation task (Su et al., 2016; Mao et al., 2018; Rosen, 2018), much of the literature in the last years has been devoted to metaphor identification (Leong et al., 2018; Gao et al., 2018; Dankers et al., 2019, 2020; Leong et al., 2020). The use of deep learning for this task has become widespread and has contributed to advance the state-of-the-art. While the most recent models proposed for this task differ with respect to the information they exploit, most of them use Transformers Neural Language Models like BERT in order to obtain an initial representation of the processed sequence. This strategy is now very common and has led to general improvements in performance: Four out of the six best systems that participated in the 2020 shared task on the VUA corpus (Steen et al., 2010) used it, and a system based only on BERT and an additional layer outperformed many other systems which were based on explicit linguistic information (Leong et al., 2020). These results strongly suggest that models like BERT already possess some knowledge that is relevant to the detection of metaphors. However, to the best of our knowledge, there is no study directly investigating what these models know about metaphors and if this knowledge shares some aspects with that of humans.

The application of this last question to other aspects of language has characterized the field of study known as BERTology (Rogers et al., 2020). Researchers in this field intrinsically evaluate BERT and its variants on challenge sets (Belinkov and Glass, 2019), annotated datasets of various sizes that target specific linguistic phenomena. Part of the literature on this subject made use of methods which we will adopt in our work. A first method is to study the probabilities assigned by the language model to the instances of a specific linguistic phenomenon, to establish whether the model is able to predict such phenomenon (Goldberg (2019); Ettinger (2020) among others). This methodology can provide interesting insights when applied to figurative language. Since these phenomena can be seen as exceptions to general linguistic rules, they require the model to apply some special ability, which is independent and often in contrast with the signals it has been exposed to during training.

Another possibility is to directly investigate the model's embeddings which are then used for tasks like metaphor identification. Since these representations have been shown to be transferable to a high number of linguistic tasks (Liu et al., 2019), they must encode some sort of general knowledge about linguistic expressions. Previous work (Bommasani et al., 2020; Chronis and Erk, 2020) hypothesizes

that they encode knowledge about meaning, and therefore BERT and its variants can be seen as Distributional Semantics Models (**DSMs**), (Lenci, 2018). Consequently, they applied the same methods used in distributional semantics to BERT, for example comparing the model's representations via cosine similarity to see whether they reproduce human judgments of similarity. However, BERT as a DSM has a crucial property which previous DSMs lack: The ability to produce **contextualized embeddings** for each word token. As such, it can positively contribute to the modeling of the meaning of figurative expressions like metaphors, where words acquire new senses which are highly context-dependent. This is why BERT can be compared with previous distributional models of metaphors, and the evaluation methods used for these models can be adapted to test BERT. In this work, we use the landmark method proposed by Kintsch (2000), which tested whether distributional representations of metaphors reflect human intuitions about the meaning of the expressions they model. We will describe the method in more detail in section 3.2.

3 Data and Experiments

3.1 Dataset

We found that existing datasets of metaphors (see Parde and Nielsen (2018) for a review) are not particularly well-suited to test how the model deals with expressions with different structures and different degrees of metaphoricity and plausibility, and more specifically with their interpretation. In fact, to the best of our knowledge none of the existing datasets presents, for different types of structures, annotation regarding both the conventionalization of a metaphor and its interpretation. Moreover, nonsense sentences have never been included in datasets used for computational modeling of such phenomenon.

We therefore decided to create a new dataset. Our dataset contains 47 conventional metaphors and 53 creative metaphors. Metaphors have different linguistic structures: Attributive, or "A is B" like in *An ambassador is a peacock* (Cacciari and Glucksberg, 1994), Genitive or "A of B", like in *There was an invisible necklace of now* (Bambini et al., 2014), Complex like in *He planted the seeds of doubt in my mind* (Newmark, 1980) and Single-Word or one-word-metaphors, like in *The mother broke the silence* (Newmark, 1980). The conventional metaphors were selected from BNC and the English Web Corpus (2015, 2018) while the creative metaphors from Katz et al. (1988), Rasse et al. (2020), poetries, and the Web. We matched each sentence with two control items (literal and nonsense) with the same structure as the metaphors (an example of a sentence with the corresponding control items can be seen in Table 1).

	Test sentence
Met	I could almost taste victory.
Lit	I can taste ginger in this cake.
Nonsense	I could almost wash victory.

Table 1: Example of test sentences from the dataset.

The final dataset includes 300 items that have been rated for their degree of **semantic plausibility** and **metaphoricity** by human subjects recruited through the Prolific crowdsourcing platform.

Semantic Plausibility test To assess the semantic plausibility of our sentences, we submitted a survey to 20 English speakers. The participants were asked to judge how meaningful a given sentence was (metaphoric, literal, or nonsense), on a Likert scale from 1 (meaningless) to 7 (meaningful). Pairwise Wilcoxon comparisons showed that each group was significantly different from the others (p-value < 0.001). The participants rated conventional metaphors as less meaningful (M= 4.86) than literal expressions (M=5.45), but more plausible than creative metaphors. Creative metaphors were judged (M= 3.69) as more meaningful than nonsense expressions (M= 2.7). Crucially, this test shows that, on the one hand, subjects perceive that metaphorical sentences somehow "deviate" from literal ones, but on the other hand, they recognize that metaphorical sentences, even the most creative ones, differ from purely nonsense structures.

Metaphoricity test To assess the metaphoricity of our sentences, we submitted a second survey to 20 English speakers. The participants were asked to judge how metaphoric a given metaphorical or literal expression was, on a Likert scale from 1 (literal) to 7 (metaphorical). Significance values for the pairwise comparisons were corrected for multiple comparisons using the Bonferroni correction. The results showed that creative metaphors were rated as more metaphoric (M= 5.59) than conventional metaphors (M= 4.64, p-value < 0.001), and conventional metaphors obtained a significantly higher score of metaphoricity than literal expres-

sions (M= 1.94, p-value < 0.001). Therefore, this test reveals that the conventionalized nature of conventional metaphors does not alter their figurative power with respect to literal sentences.

3.2 Models and Experiments

We carried out our experiments with the base (number of layers = 12, hidden size = 768) and the large (number of layers = 24, hidden size = 1024) cased versions of BERT. We used the pretrained model that is provided by the HuggingFace library Transformers (Wolf et al., 2020).

Experiment 1: Sentence plausibility scores To get an estimate of how much BERT considers an expression as plausible, we computed a probability score for each sentence in our dataset. Then we examined whether the scores vary with the sentence types (metaphorical conventional, metaphorical creative, literal, nonsense), and whether they mirror human plausibility judgments.

As a measure of sentence plausibility, we used the **pseudo-log-likelihood score** (PLL) (Wang and Cho, 2019). The probability of a sentence cannot be computed using autoencoding models like BERT. In fact, these models are inherently bidirectional, that is, they are trained to predict a word given all the other words in the left and the right context. Therefore, they cannot be used for estimating probabilities of sequences via the chain rule, since this requires to compute the probability of any word given the *previous* words in a sequence. The PLL of a sentence W is obtained by masking one token w at a time, calculating the token's probability given all the other context words, and summing the log-probabilities for all the tokens as in Equation 1. Salazar et al. (2020) showed that the PLL score, even if strictly speaking it is not a probability, outperforms scores from autoregressive models in a variety of tasks related to the acceptability of sentences. This is probably due to the fact that the PLL eliminates the discrepancy between the probabilities of the first elements of a sequence and those of the last elements.

$$PLL(W) = \sum_{t=1}^{|W|} \log P(w_t \mid W_{\setminus t}) \qquad (1)$$

Experiment 2: The landmark method To determine whether the model's representations of metaphorical expressions reproduce the shift from a source literal domain to a new target one, we applied the landmark method, which was proposed by Kintsch (2000) to test a model producing distributional representations of predicate-argument metaphors (e.g., *My lawyer is a shark*). His aim was to test to which extent these vectors encoded information about the meaning of the argument (*lawyer*), the inappropriate literal sense of the predicate (*shark*) and the appropriate metaphorical sense of the predicate (e.g., *ferocious*). To this end, he selected three different sets of words, where the words in each set were related to only one of those meanings based on the author's intuitions. For example, *justice* and *crime* were selected because they are related to *lawyer*, *fish* is related to the literal sense of *shark* and *viciousness* is related to the metaphorical sense of *shark*. These words were used in Kintsch (2000) as **landmarks**, in that their vector representations were compared to the model's representations of metaphors.

The landmark method gives us a straightforward way of examining BERT's internal representations (i.e., the contextualized embeddings generated by its layers) with regards to metaphorical meaning. Since the essence of BERT is that each word token is associated with a context sensitive embedding, if the model is somehow "aware" of the figurative interpretation of a metaphorical sentence, this should be reflected in the embeddings of its words. We performed two versions of our experiment, which mainly differ for the representations that are examined. In the **MetSentences** version, we tested the global metaphorical interpretation of a sentence, by representing it with an embedding obtained by summing the embeddings of its words. In the **MetWords** version, we examined the representations of specific words that undergo the metaphorical shift. For example, in *This fighter is a lion*, the word *lion* is used metaphorically, while the other words keep their literal meaning. In cases where the shift involves more than one word (e.g., *The heart is the legendary muscle that wants and grieve*), we summed the embeddings of the metaphorical words to obtain a representation that capture the information that is common to both words.

We tested these two types of representations with respect to the literal and metaphorical sense of the word(s) that carry the metaphorical meaning. Such information is crucial in determining whether BERT interprets metaphorical expressions successfully. We therefore defined two sets of landmarks for each metaphorical sentence of our dataset

Sentence	metaphorical landmarks literal landmarks
A smile is an **ambassador**.	message confident express official embassy envoy
The flowers **nodded** in the wind	movement wave sway assent head greeting

Table 2: Landmarks for two sentences from the dataset

(two examples of sentences from our datasets with the corresponding sets of landmarks can be seen in Table 2). The metaphorical landmarks are elements of the target conceptual domain to which the metaphorical words in a sentence point. For example, the words *want* and *grieve* in the sentence *The heart is the legendary muscle that wants and grieve* point to the target domain of emotion. The literal landmarks can be synonyms, associates or words morphologically derived from the literal sense of one of the metaphorical words in the sentence. If a sentence contains more than one metaphorical word, we defined the set of literal landmarks so that it contains at least one related word for each metaphorical word in the sentence.

To obtain BERT contextualized embeddings of the landmarks, we extracted a sentence from the landmark's lexical entry in WordNet or the Cambridge Dictionary and we passed it to BERT. Each landmark set was then represented with the averaged embeddings of its words.

Given the limited size of our dataset, we chose not to use the representations of the model as input for another layer specifically optimized for choosing the right set of landmarks. Therefore, we analyzed directly the model's output. Specifically, we compared the similarity (via cosine similarity) of the representations of the metaphorical expressions and the representations of the metaphorical/literal landmarks. To see whether conclusions can be derived from a simple analysis of the representations, we evaluated the model in two ways. First, we checked whether the similarity between the expressions and the metaphorical landmarks is in general significantly higher than the literal landmarks. We also evaluated the model on a binary classification task, measuring the accuracy in producing embeddings that are more similar to the metaphorical landmarks than to the literal ones. To estimate the added value of BERT contextualized embeddings with respect to traditional static embeddings, we compared BERT performance with a baseline computed with GloVe vectors (Pennington et al., 2014) (context window=2, dimension=300).

Since we are interested in exploring the inner behavior of the model, we compared the representations for each BERT layer. Recent studies (Liu et al., 2019; Jawahar et al., 2019) have shown that upper-intermediate layers produce general purpose representations which are less dependent on the language modeling objective and approximate linguistic knowledge better. We surmise that this ability could also reflect on the interpretation of metaphors.

4 Results

Experiment 1 We calculated the Spearman ρ correlation between the PLL and the semantic plausibility human ratings. Both models show a fair correlation: BERT base $\mathbf{0.504}$ and BERT large $\mathbf{0.496}$.

We then verified whether BERT mirrors the human perceived contrast between conventional metaphorical and literal sentences, and between creative metaphorical and nonsense sentences.

The models' results (first two boxplots of figure 4.1) show that the creative metaphors (BERT base: M= -4.1; BERT large: M= -4.04) were rated by the models with significantly higher scores than nonsense expressions (BERT base: M= -4.96; BERT large: M= -4.93, p-value < 0.003). Furthermore, the conventional metaphors (BERT base M= -3.62 ; BERT large M= -3.46) were considered less likely than literal ones (BERT base M=-3.01; BERT large M= -3.04), but they were rated with higher scores than creative metaphors, though these two differences are not statistically significant (p-value > 0.05). These results partially reflect human judgments (right boxplot of figure 4.1).

Experiment 2 In the **MetSentences** setting, the embeddings of both models (base, large) for the metaphorical sentences have significantly higher cosine similarity with the embeddings of the metaphorical landmarks (e.g. layer 10 (base) M= 0.83582; layer 20 (large) M = 0.58177) than with the embeddings of the literal landmarks (e.g. layer 10 (base) M= 0.82691, p-value < 0.03 ; layer 20 (large) M = 0.56087, p-value < 0.01) in the upper intermediate layers (base: 9,10,11; large: 11-22, p-value < 0.05). The accuracy of the models (which is shown in figure 4.2) varies along the layers with the same tendency.

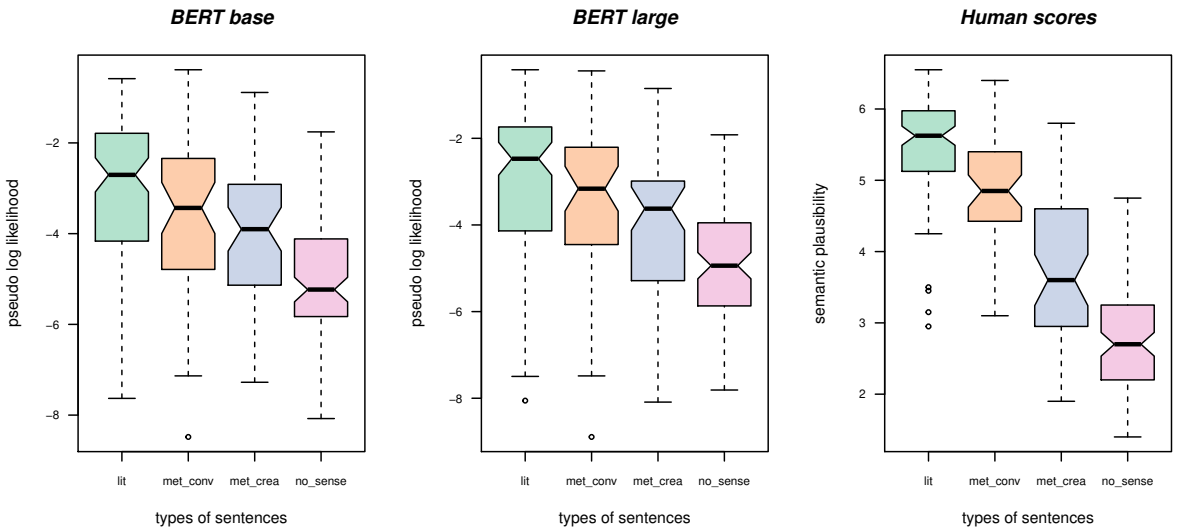

Figure 4.1: Boxplots of BERT pseudo-log-likelihood scores and the human ratings of semantic plausibility. For each distribution, the boxplot indicates the median, the quartiles, the maximum and the minimum of the distribution and the outliers when they are present.

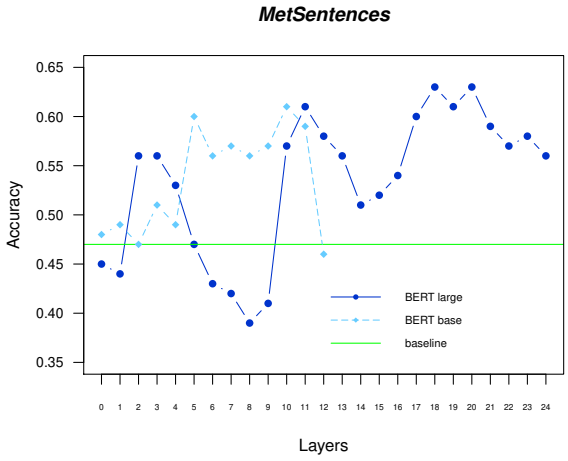

Figure 4.2: Models' accuracy for **MetSentences**

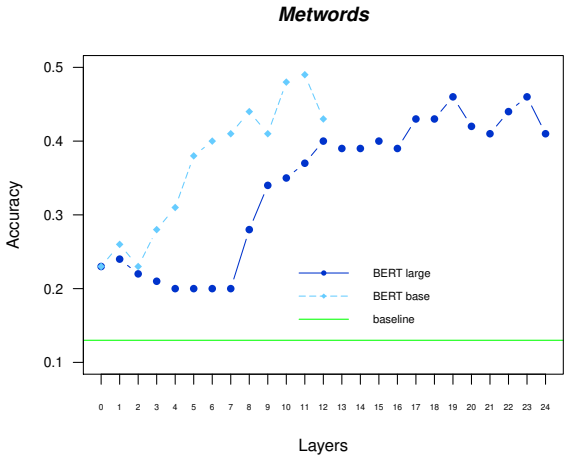

Figure 4.3: Models' accuracy for **MetWords**

As can be seen in Figure 4.2, BERT accuracy outperforms the GloVe baseline (0.47) in the layers 0-1, from 3 to 11 (BERT base), from 2 to 4 and from 10 onward (BERT large). This baseline is outperformed by more than 10 percentage points in some layers (10-11 BERT base; 10-12,17-21,23 BERT large). The layers 10 (base) and 20 (large) produce the highest accuracy values: 0.61 (base) and 0.63 (large). Examples of sentences that are classified correctly (the model's representations are more similar to the metaphorical landmarks than to the literal ones) and sentences that are classified wrongly by the best BERT large layer are reported in Table 3.

In the **MetWords** setting and in both models, the similarities to the two landmark sets are not significantly different (p-value > 0.05). The only exceptions are the first layers where, in contrast with our expectations, the metaphorical expression embeddings are significantly more similar to the literal landmarks embeddings than to the metaphorical ones. As can be seen from figure 4.3, the accuracy generally increases along the layers apart from the final ones, in the same way as we saw for **MetSentences**.

In **MetWords** (cf. figure 4.3), BERT accuracy overcomes the correspondent GloVe baseline (acc= 0.13) by a margin ranging from 10 (layers 0,2) to

Best perfomance Acc.0.63	Conventional Creative
Correct	What is the source of your anger? Laughter is the mind sneezing.
Wrong	Lawyers are real sharks. Prejudice is the child of ignorance.

Table 3: Examples of correct and wrong sentences from the layer with the best performance (20 BERT large)

36 (layer 11) points in BERT base and from 07 (layers 4-7) to 33 (layers 19,23) points in BERT large. Instead, BERT accuracy almost never goes beyond the **MetSentences** version of the baseline, except the layers 10-11 of BERT base with a difference no higher than 2 points.

We also evaluated the influence of the conventionality of a metaphor on the success of the models' interpretation[1]. Data (see the right plot in figure .4) show that, in BERT-large, the proportion of the correct conventional metaphors is higher than the proportion of the correct creative ones up to a margin of 10 points in **MetSentences** (layers 2-5, 10, 16-17, 24 BERT large). In this setting, this disparity decreases along the layers of the model and it disappears in the layers close to 20 (BERT large). In BERT base (cf. the left plot in figure .4), the proportion of correct creative metaphors overcomes the proportion of correct conventional metaphors at the layers 0, 3-4, 6-12. On the other hand, the wrong cases in **MetSentences** (see figure .7) are mostly creative metaphors. As well as for the correct representations, the disparity tends to decrease along the layers of both models and in BERT large it disappears at the layer 23, in BERT large at the layer 7). On the other hand, these differences persist in the **MetWords** version at all layers and in both models (see figures .5 and .6).

5 Discussion

Likelihood of metaphors with respect to other expressions. The likelihood scores that the models assign to metaphorical expressions lie halfway between those of literal and nonsense sentences, in line with human judgments of semantic plausibility (compare the first two plots in Figure 4.1 with the last one). Most importantly, BERT discriminates unconventional creative metaphors from nonsense sentences and seems to discriminate highly conventionalized metaphors from literal expressions, even though the difference is not significant in the second case.

This result raises some interesting questions that need to be further investigated. Where does BERT's ability to discriminate metaphorical expressions come from? How does BERT know that, for example, the sentence *A smile is an ambassador* (PLL=−6.02) is more plausible than the sentence *A smile is a fishing man* (PLL=−6.89)? It is likely that the model has not received explicit training on this, since it has probably never or rarely encountered any of such expressions during learning. Does this ability share some aspects with the human ability of producing novel metaphors? A possible answer is that the creative metaphors that are in our dataset are grounded on associations that manifest themselves more frequently in language. For example, the creative metaphor *The sea whispered on the sand* represents a personifying mapping which is expressed by other more frequent expressions such as *whispering wind*. The model might learn these mappings from training data in the form of associations between groups of words, and then extend them to similar constructions.

An analysis of the model's representations, although it does not allow us to draw firm conclusions about BERT capability to capture our intuitions about metaphorical meaning, brings out fairly clear trends with respect to the model's interpretation of metaphors, some of which are consistent with findings in the previous literature.

General evaluation of BERT. Our second experiment reveals that in some cases BERT representations of metaphorical expressions are significantly more similar to the metaphorical landmarks than to the literal ones. This result is not stable but varies considerably across configurations (see below). Moreover, in both versions of Experiment 2, BERT is almost always above the GloVe baseline (as can be seen from the plots in Figures 4.2 and 4.3). For **MetWords** this is not surprising, since in that case we examined only the static vector of a word, which is obviously more similar to the literal than to the metaphoric landmarks. However, it is less obvious for **MetSentences**, where the baseline is outperformed by 16% in some layers (BERT classified correctly 16 more sentences out of 100).

Analysis of internal representations. At upper-intermediate layers, we find significantly more metaphorical than literal information in the model's representations, and the accuracy in the

[1]These data are reported in the Appendix due to space constraints

landmark task typically rises until it reaches its peak (see the plots in Figures 4.2 and 4.3). This behavior is probably explained by the fact that, at higher layers, the model produces more contextualized embeddings (Ethayarajh, 2019), and contextual information is relevant to understanding metaphorical meaning shifts. This also explains why BERT large achieves higher scores than the base version: The deeper the model, the stronger the contextualization.

This result is consistent with the previous findings about the distribution of linguistic abilities in different layers of the model. It has been shown (Liu et al., 2019; Jawahar et al., 2019) that upper-intermediate layers are the ones that perform best in probing tasks involving different types of linguistic knowledge. Our results are also consistent with findings from metaphor identification (Liu et al., 2020). The error analysis of our results shows that the first layers of the model do not encode much information that is useful for the task and is not present in the subsequent layers. In the **MetSentences** setting, 27 out of the 37 errors made from the layer 20 of BERT large (the layer with the highest accuracy) are common to layer 1.

Sentence vs. word representations. In **MetWords** (where we examined only the embeddings of the specific words that are used metaphorically), BERT representations are never significantly biased in favor of the metaphorical reading. As can be seen by comparing the plots in Figures 4.2 and 4.3, results are almost always worse than in the **MetSentences** setting and below the additive GloVe baseline. In **MetWords**, the performance is affected by the fact that the BERT embeddings of a single word token encode a significant amount of information about its literal sense. The salience of this information diminishes as the representation is fed to the higher layers of the models, but it is not enough for the models to achieve performances comparable to those in the **MetSentences** version.

Another major difference between the two approaches concerns metaphor conventionality. In the **MetSentences** setting, the improvements that we observe when using higher layers mainly concern creative metaphors (this can be seen from the plots in .4 and .7). While in the first layers the models generally classify correctly more conventional than creative metaphors, the difference tends to reduce as we climb up the layers. In other words, the interpretation of creative metaphors seems to benefit the most from the process of contextualization. On the other hand, this does not occur in the **MetWords** version (see .5 and .6), where the number of conventional metaphors among the items that were correctly classified by the model is always higher than the number of creative metaphors.

An important difference between conventional and creative metaphors can account for both these results. Since creative metaphors are idiosyncratic and context-dependent, it is more likely that the model needs global information about all sentence components to "understand" metaphorical aspects of their meaning. This ability manifests itself more clearly in the later layers of the model, where a larger amount of contextual information has been processed through the repetition of the self-attention mechanism. Accordingly, the information crucial to the interpretation of these metaphors will more likely be a feature of the representation of the entire sequence. In contrast, conventional metaphors occur more systematically in language and eventually become lexicalized. As such, they should be already encoded in the general representations that the model creates for single words, since these representations account for the behavior of a word in all the contexts in which it has been encountered. The process of interpreting highly conventionalized metaphors is thus akin to disambiguation, since the relevant meaning is already encoded in the original embedding of a word and the model only needs to recover it by using contextual information. Therefore, the process is likely to be successfully accomplished earlier by the model.

6 Conclusion

In this paper, we adopted different methods used in the literature on the analysis of neural language models to explore how BERT captures various human evidence about a linguistic phenomenon that is relatively underexplored in this field, namely metaphors. In **Experiment 1**, starting from the assumption that metaphorical sentences are more plausible than nonsense sentences (even if they are "deviant" from literal ones), we tested whether BERT can make a distinction between literal, nonsense and metaphorical expressions with various degrees of conventionality. We show that BERT can distinguish between these types of expressions, assigning them different degrees of plausibility in much the same way as humans do.

In **Experiment 2**, we wanted to test BERT's

ability to understand the figurative meaning of metaphorical sentences and the lexical meaning shifts these sentences activate, in comparison with their literal meaning. Despite the limited size of our dataset, we can derive some trends. In particular, we observed that the shift is captured by the model's representations (without fine-tuning and when compared via cosine similarity) to a fair extent, especially when it can exploit more contextual information (in upper layers and at the sentence level). We take the results of this experiment as the starting point for future work aimed at investigating how the model can perform when it receives explicit training on this specific task.

There are many directions for future works, including testing the model using landmarks which do not derive from subjective intuitions, but which are obtained from human judgments. Some of the hypotheses formulated here (for example, BERT ability to derive metaphorical mappings from training) need to be verified on a larger dataset of metaphors, where different conceptual domains involved in the phenomenon can be explored. Moreover, the use of datasets made of longer portions of texts could help BERT in improving its ability to encode metaphorical meanings, since we have shown that contextualization improves the model's ability, with positive effects in particular on the interpretation of creative metaphors.

References

H. Al-Azary and L. Buchanan. 2017. Novel metaphor comprehension: Semantic neighbourhood density interacts with concreteness. *Memory Cognition*, 45:296–307.

Valentina Bambini, Donatella Resta, and Mirko Grimaldi. 2014. A Dataset of Metaphors from the Italian Literature: Exploring Psycholinguistic Variables and the Role of Context. *PLoS ONE*, 9(9).

Yonatan Belinkov and James Glass. 2019. Analysis Methods in Neural Language Processing: A Survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Brian J. Birdsell. 2018. Conceptual Wandering and Novelty Seeking: Creative Metaphor Production in an L1 and L2. *Journal of Cognitive Science*, 19(1):35–67.

Y. Bizzoni and S. Lappin. 2017. Predicting human metaphor paraphrase judgments with deep neural networks. *Memory Cognition*, 45:296–307.

Rishi Bommasani, Kelly Davis, and Claire Cardie. 2020. Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings. In *Proceedings of ACL*.

Cristiana Cacciari and Sam Glucksberg. 1994. Understanding figurative language. In *Handbook of psycholinguistics*, page 447–477. Academic Press.

J. Charteris-Black. 2004. *Corpus Approaches to Critical Metaphor Analysis*. Palgrave Macmillan.

Gabriella Chronis and Katrin Erk. 2020. When Is a Bishop Not Like a Rook? When It's Like a Rabbi! Multi-prototype BERT Embeddings for Estimating Semantic Relationships. In *Proceedings of CONLL*.

Verna Dankers, Karan Malhotra, Gaurav Kudva, Volodymyr Medentsiy, and Ekaterina Shutova. 2020. Being neighbourly: Neural metaphor identification in discourse. In *Proceedings of the Second Workshop on Figurative Language Processing*, page 227–234.

Verna Dankers, Marek Rei, Martha Lewis, and Ekaterina Shutova. 2019. Modelling the interplay of metaphor and emotion through multitask learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, page 2218–2229.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL*.

Kawin Ethayarajh. 2019. How Contextual Are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. In *Proceedings of EMNLP*.

Allyson Ettinger. 2020. What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models. *Transactions of the Association for Computational Linguistics*, 8.

Ge Gao, Eunsol Choi, Yejin Choi, and Luke Zettlemoyer. 2018. Neural Metaphor Detection in Context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, page 607–613.

Dedre Gentner and Brian Bowdle. 2008. Metaphor as structure-mapping. In Raymond W.Editor Gibbs, Jr., editor, *The Cambridge Handbook of Metaphor and Thought*, Cambridge Handbooks in Psychology, page 109–128. Cambridge University Press.

Sam Glucksberg. 2003. The psycholinguistics of metaphor. *Trends in Cognitive Sciences*, 7(2):92–96.

Sam Glucksberg, Patricia Gildea, and Howard B. Bookin. 1982. On understanding nonliteral speech: Can people ignore metaphors? *Journal of Verbal Learning and Verbal Behavior*, 21(1):85–98.

Yoav Goldberg. 2019. Assessing BERT's Syntactic Abilities. arXiv:1901.05287.

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*, page 3651–3657.

Albert N. Katz, Allan Paivio, Marc Marschark, and James M. Clark. 1988. Norms for 204 Literary and 260 Nonliterary Metaphors on 10 Psychological Dimensions, Metaphor and Symbolic Activity. *Metaphor and Symbolic Activity*, 3:4:191–214.

Walter Kintsch. 2000. Metaphor comprehension: A computational theory. *Psychonomic Bulletin Review*, 7:257–266.

VT Lai, T Curran, and L. Menn. 2009. Comprehending conventional and novel metaphors: an ERP study. *Brain Research*, 1284:145–55.

G. Lakoff and M. Johnson. 1980. *Metaphors We Live By*. University Of Chicago Press.

Alessandro Lenci. 2018. Distributional Models of Word Meaning. *Annual Review of Linguistics*, 4:151–171.

Chee Wee (Ben) Leong, Beata Beigman Klebanov, Chris Hamill, Egon Stemle, Rutuja Ubale, and Xianyang Chen. 2020. A Report on the 2020 VUA and TOEFL Metaphor Detection Shared Task. In *Proceedings of the Second Workshop on Figurative Language Processing*, page 18–29.

Chee Wee (Ben) Leong, Beata Beigman Klebanov, and Ekaterina Shutova. 2018. A Report on the 2018 VUA Metaphor Detection Shared Task. In *Proceedings of the Workshop on Figurative Language Processing*, page 56–66.

Jerry Liu, Nathan O'Hara, Alexander Rubin, Rachel Draelos, and Cynthia Rudin. 2020. Metaphor Detection Using Contextual Word Embeddings From Transformers. In *Proceedings of the Second Workshop on Figurative Language Processing)*, page 250–255.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic Knowledge and Transferability of Contextual Representations. 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)*, page 1073–1094.

Rui Mao, Chenghua Lin, and Frank Guerin. 2018. Word Embedding and WordNet Based Metaphor Identification and Interpretation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 1222–1231.

S. McGregor, K. Agres, K. Rataj, M. Purver, and G. Wiggins. 2019. Re-Representing Metaphor: Modeling Metaphor Perception Using Dynamically Contextual Distributional Semantics. *Frontiers in Psychology*, 10(765).

Peter Newmark. 1980. The Translation of Metaphor. *Babel*, 26(2):93–100.

Natalie Parde and Rodney Nielsen. 2018. A Corpus of Metaphor Novelty Scores for Syntactically-Related Word Pairs. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of EMNLP*, pages 1532–1543.

Carina Rasse, Alexander Onysko, and Francesca M. M. Citron. 2020. Conceptual metaphors in poetry interpretation: A psycholinguistics approach. *Language and Cognition*, 12(2):310–342.

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.

Zachary Rosen. 2018. Computationally Constructed Concepts: A Machine Learning Approach to Metaphor Interpretation Using Usage-Based Construction Grammatical Cues. In *Proceedings of the Workshop on Figurative Language Processing*, page 102–109.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked Language Model Scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 2699–2712.

Ekaterina Shutova. 2010. Automatic Metaphor Interpretation as a Paraphrasing Task. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics)*, page 1029–1037.

Gerard J. Steen, Aletta G. Aletta G. Dorst, J. Berenike Herrmann, Anna A. Kaal, Tina Krennmayr, and Trijntje Pasma. 2010. *A method for linguistic metaphor identification. From MIP to MIPVU.* John Benjamins, Amsterdam, Netherlands.

Chang Su, Shuman Huang, and Yijiang Chen. 2016. Automatic detection and interpretation of nominal metaphor based on the theory of meaning. *Neurocomputing*, 219.

Alex Wang and Kyunghyun Cho. 2019. BERT has a Mouth, and It Must Speak: BERT as a Markov Random Field Language Model. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, page 30–36.

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*, page 38–45.

Figure .4: Experiment 2. Distribution of correct answers with respect to the conventionality of the metaphor along the layers in BERT base (on the left) and BERT large (on the right) in MetSentences version.

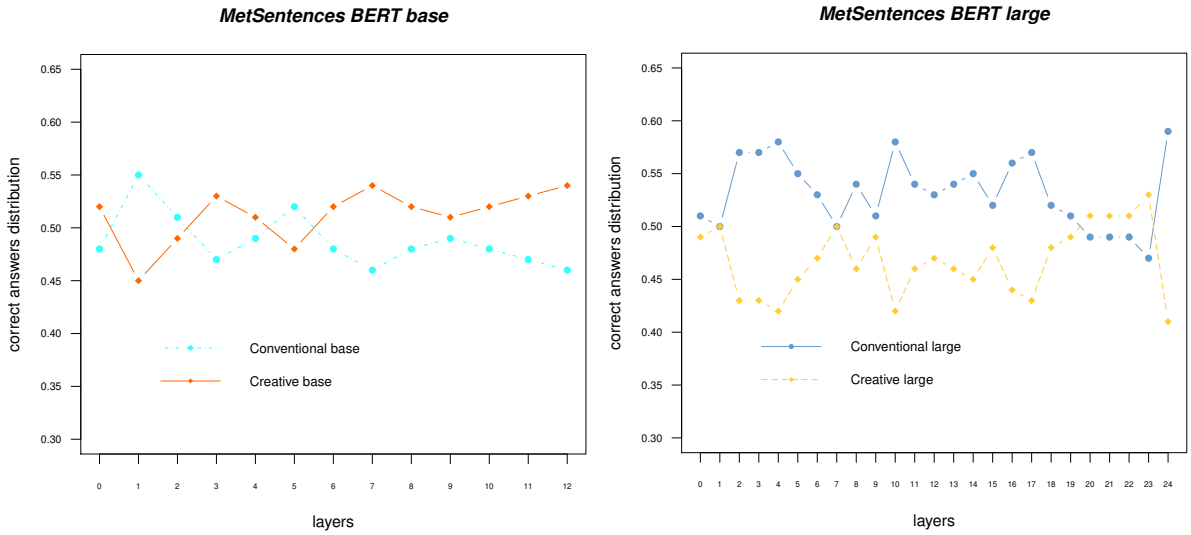

Figure .5: Experiment 2. Distribution of correct answers with respect to the conventionality of the metaphor along the layers in BERT base and BERT large in MetWords version.

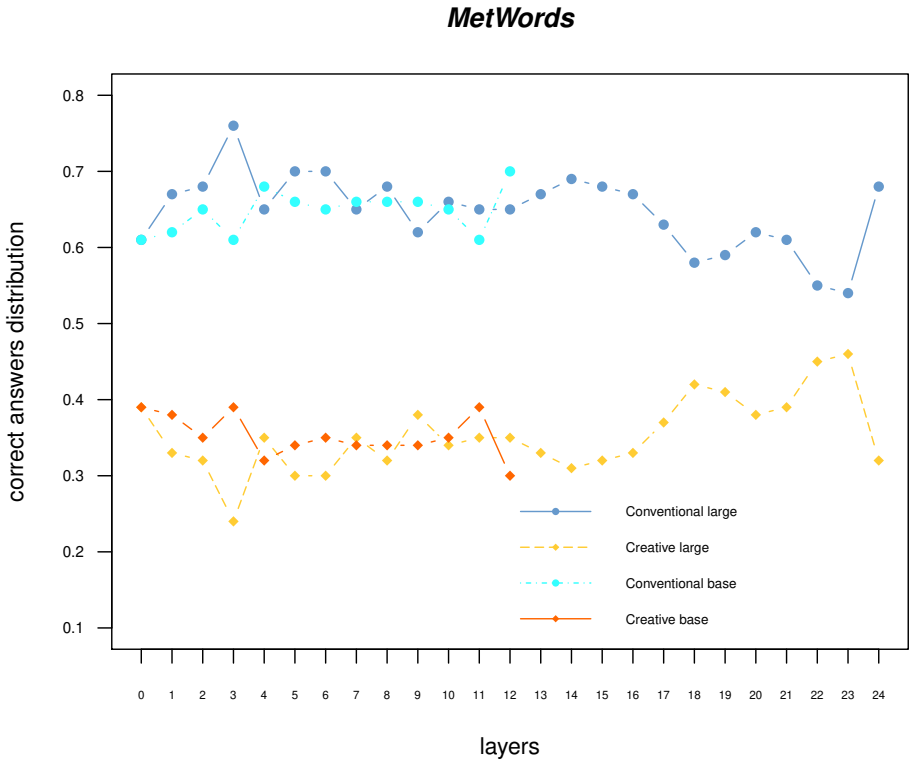

Figure .6: Experiment 2. Distribution of wrong answers with respect to the conventionality of the metaphor along the layers in BERT base and BERT large in MetWords setting.

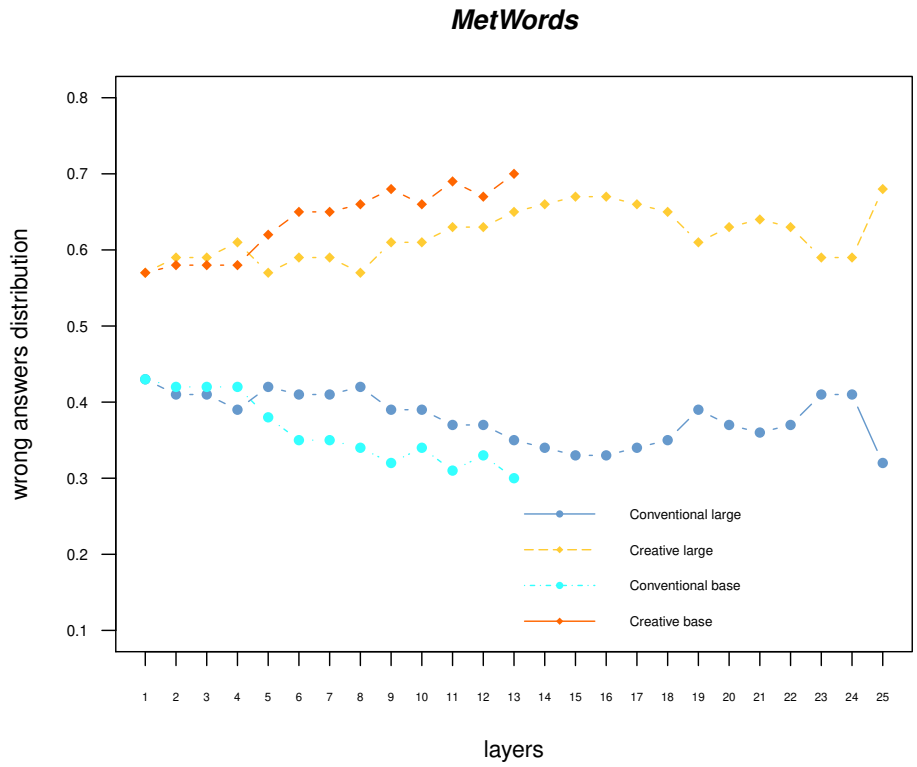

Figure .7: Experiment 2. Distribution of wrong answers with respect to the conventionality of the metaphor along the layers in BERT large and BERT base in MetSentences version.