On the Cusp of Comprehensibility: Can Language Models Distinguish Between Metaphors and Nonsense?

Bernadeta Griciūtė[1,2], Marc Tanti[1], and Lucia Donatelli[2]
[1]University of Malta
[2]Saarland University
{bernadeta.griciute.21, marc.tanti}@um.edu.mt
donatelli@coli.uni-saarland.de

Abstract

Creative texts can sometimes be difficult to understand as they balance on the edge of comprehensibility. However, good language skills and common sense can allow advanced language users to both interpret creative texts and reject some linguistic input as nonsense. The goal of this work is to evaluate whether current language models can make the distinction between creative language use and nonsense. To test this, we have computed the mean rank and pseudo-log-likelihood score (PLL) of metaphorical and nonsensical sentences. We have also fine-tuned RoBERTa for binary classification between the two categories. There was a significant difference in the mean ranks and PLL scores of the categories, and the classifier reached around 75-88% accuracy. The results raise interesting questions on what could have led to such satisfactory performance.

1 Introduction

The ultimate goal of Natural Language Understanding (NLU) models is to reach a human-like level of language comprehension. However, a good command of language manifests itself not only in being able to interpret advanced usages of a language, but also in discriminating the uninterpretable, erroneous cases. While automatic grammar checkers are already in place, semantic incongruity is more difficult to trace. The task is further complicated by the existence of figurative language, where a listener is required to go an extra step (when compared to literal language) in order to decode the meaning. The borderline between creative, but still understandable text, and nonsense can be seen as the cusp of comprehensibility.

One of the types of figurative language is metaphors, which are convenient to research due to their ubiquity. Linguistic **metaphors** can be defined as expressions of an understanding of one concept in terms of another, where there is some similarity between the two. While metaphor per se signifies a shift in meaning, they do vary in the degree of metaphoricity and creativity. The most threadbare metaphors which are so commonly used that they become unnoticeable are **conventional metaphors**, for example, "he *takes* a few moments to reply". On the other side of the scale of metaphoricity there are **creative metaphors**, where a novel meaning emerges in a sentence, for example, "the ATM *coughed up* my card" (Cardillo et al., 2010). However, even when it comes to novel metaphors, language users should still be able to infer the meaning - otherwise, they are just **nonsense**.

Professor Irving Massey has suggested that distinguishing between a metaphor and nonsense could be a new Turing test (Massey, 2021). The professor claims that switching between literal and metaphorical senses is an aesthetic gesture inaccessible for computers, and that "the ability to experience metaphor is the very definition of the human". While admittedly for the time being there is no way to track aesthetic experiences of a computer, the (in)ability of computational models to make the distinction between a metaphor and mere nonsense might be worth looking at.

While we sometimes deify metaphors as "a hallmark of human intelligence" (Cardillo et al., 2010), and assume that the interpretation of metaphors, especially of novel metaphors, demands human cognitive skills and real world experiences, it is also possible that there are enough clues encoded at the linguistic level that they would help a non-human to distinguish between metaphors and nonsense.

In order to test whether the ability to demystify metaphors is a skill exclusively possessed by mortals, we are going to measure and compare the PLL scores of metaphors and nonsense, as well as use mean ranks of predictions on masked words to test how well the nonsensicality can be explained by plausibility (language model probability). Finally, a binary classifier based on a pre-trained lan-

Proceedings of the 3rd Workshop on Figurative Language Processing (FLP), pages 173 - 177

guage model is going to be trained in order to check whether the current language models are able to distinguish between metaphors and nonsense.

2 Related Works

A study by Pedinotti et al. (2021) hints that language models might already have acquired a human-like intuition of sentence plausibility. The authors of the study have found out that the pseudo-log-likelihood scores (PLL) of sentences obtained using BERT (Devlin et al., 2019) correlated with the plausibility ratings of human annotators. The best performing model in the Corpus of Linguistic Acceptability (CoLA) (Warstadt et al., 2019) task in the GLUE benchmark (Wang et al., 2018), ERNIE, surpasses even the human baseline (75.5 vs. 66.4 MCC), discriminating linguistically unacceptable sentences better than human participants.

However, another study conducted by Gupta et al. (2021) found that the BERT family of models are easily susceptible to adversarial examples and fail to even recognize incoherent, ungrammatical utterances, giving similarly confident scores to input that was perturbed to be nonsensical as to its meaningful counterpart. Findings like this are evidence that, when discriminating between meaningful and nonsensical sentences, the models might be relying on some spurious correlations or annotator artifacts rather than the targeted divergence in comprehensibility.

3 Data and Experiments

3.1 Dataset

To the best of our knowledge, there's only one dataset that is annotated for both metaphors and nonsense - the one by Pedinotti et al. (2021), which the authors have kindly agreed to share. The dataset consists of 300 matched sentences, 100 for each of the three categories: metaphors (47 conventional and 53 creative), literal sentences, and nonsensical sentences.

In order to have more input sentences for the experiments, the dataset was further extended by adding 200 pairs of matched metaphorical and literal sentences from Cardillo et al. (2010) and Cardillo et al. (2016). These datasets were originally aimed at aiding the research of human metaphor comprehension, and contains 400 pairs (280 in Cardillo et al. (2010) and 120 in Cardillo et al. (2016)) of matched literal and metaphorical sentences, which had been carefully normalized along a number of dimensions, including length, naturalness, and figurativeness.

Type	Example
Met-Ped	I could almost taste victory.
Non-Ped	I could almost wash victory.
Met-Car	Her orders were a sharp bark.
Non-Gen	His orders were a sharp crust.
Non-BEL	Our homework buys more sky.

Table 1: Metaphor and nonsense examples from *Ped* (Pedinotti et al., 2021), *Car* (Cardillo et al., 2010), *BEL* (O'Neill et al., 2020) and the automatically *Gen*erated datasets.

Since the Cardillo et al. datasets do not include nonsense sentences, to have a balanced dataset, 200 nonsensical sentences were added. 100 of them were automatically generated (and manually handpicked from several options) by shuffling either subjects (for nominal metaphors) or subject complements (for predicate metaphors) across the sentences. Another 100 were generated with the help of BackTranslationAugmenter perturbation technique from the TextAttack framework (Morris et al., 2020), or by swapping places of verb arguments in a sentence.[1]

By generating the nonsense sentences from the metaphorical ones, we hoped to create a normalized dataset where the sentences between the categories would have similar syntactical structures and similarly plausible words. However, part of our experiments was also repeated on an extended dataset where we added the rest of the sentences (200 pairs) from the Cardillo et al. datasets, and randomly picked 200 nonsensical sentences from a corpus of sentences "without semantic context" by O'Neill et al. (2020). See Table 1 for example sentences from each dataset.

3.2 Experiments

With the chosen set of data, several experiments have been conducted. The first two explore properties of the dataset and whether the plausibility of the data can be a sufficient indicator for nonsense classification, and in the third set of experiments, a binary classifier has been trained.

3.2.1 Experiment 1: Plausibility

Following the Pedinotti et al. (2021) study, a pseudo-log-likelihood score (PLL) has been com-

[1]Our code and data are available at https://github.com/bgriciute/Metaphors-vs-Nonsense.

puted for every sentence in the picked datasets. This was done in order to check whether the same tendencies as pointed in the Pedinotti et al. (2021) paper, could be observed on a larger scope, as well as for comparing the datasets.

Since models like BERT are bidirectional, they cannot be used for computing sentence probability. An alternative way to get a probability-like score is to use PLL (Wang and Cho, 2019). The PLL score is computed by masking one token at a time, calculating its probability given all the other context words, and then summing the log-probabilities of all the words in the sentence. For the scoring, an *MLM* Python library by Salazar et al. (2020) has been used.

3.2.2 Experiment 2: Mean Ranks

Another strategy chosen to test how probable a string is according to a language model was to see, what rank a masked target word would get among the predictions of a model.

In the sentences from the Pedinotti et al. (2021) dataset, the target words (single words that are used metaphorically or nonsensically) were masked. The masked sentences were then fed to the BERT (Devlin et al., 2019) language model. To compare the predictability of the target words, we looked at which ranking position the target word that was masked would appear when sorted by probability.

3.2.3 Experiment 3: Classifier

For the classification experiments, we chose to fine-tune a pre-trained RoBERTa (Liu et al., 2019) language model. It has been chosen after conducting some primary experiments where it did perform better than BERT or MultiBERT. The `roberta-base` version by HuggingFace(Wolf et al., 2020) was fine-tuned with Adam optimizer and a learning rate of 1e-6 for 8 epochs, picking afterwards a model from the best epoch for testing. The classification was performed on different combinations between metaphorical, literal, and nonsensical sentences.

Additionally, we have also trained a Naive Bayes classifier in order to validate that the classification task on the target dataset requires a more complex method than a bag-of-words approach.

4 Results

Experiment 1

Table 2 indicates average PLL scores of each type of sentences (where applicable) for each of the aforementioned datasets that have been chosen for the final training. Additionally, Figure 1 illustrates the distribution of the scores within each category. The PLL scores reveal, in accordance with the results of the Pedinotti et al. (2021) experiments, a difference between the three categories, the literal sentences being most plausible, followed by the metaphors, and nonsense sentences, meaning that the RoBERTa model finds nonsense sentences the least plausible.

	Pedinotti	Cardillo	O'Neill
Literal	-17.8	-17.8	-
Metaphor	-26.4	-23.5	-
Nonsense	-33.1	-30.13	-44.7

Table 2: Average PLL score of the different categories across datasets (the nonsense sentences in the Cardillo column are automatically generated).

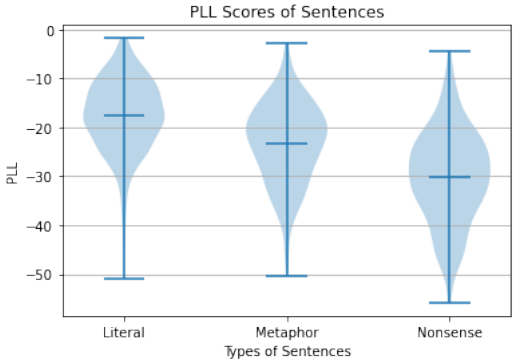

Figure 1: PLL scores of the literal and metaphorical sentences from Cardillo et al. (2010, 2016) datasets, and nonsensical sentences automatically generated from them.

It is interesting to note that the metaphorical sentences from the Pedinotti et al. (2021) dataset were on average less probable than the ones from Cardillo et al. (2010) (-26.4 versus -23.5 PLL), even though both conventional and creative metaphors were scored together. On the other hand, the nonsensical sentences manually created by Pedinotti et al. (2021) were evaluated by the model as way more probable than the sentences from the O'Neill et al. (2020) dataset which have been created by automatically shuffling words in the sentences (-33.1 vs. -44.7 PLL).

Experiment 2

In Experiment 2, we could also observe a significant difference between the ranks of sentences from different categories. Figure 2 gives a violin plot of the ranks of sentences from different categories.

Categories	Accuracy
lit-non	92.5%
lit-met	85.0%
met-non	75.0%
met-non (ext.)	88.0%

Table 3: Accuracy of the fine-tuned RoBERTa classifier between the different categories: *lit* - literal, *met* - metaphorical, and *non* - nonsense. The last experiment was also repeated on an extended dataset.

One can observe that the median ranks of nonsensical sentences were way higher than the ones of target words in literal or metaphorical sentences, meaning that the target words were less predictable.

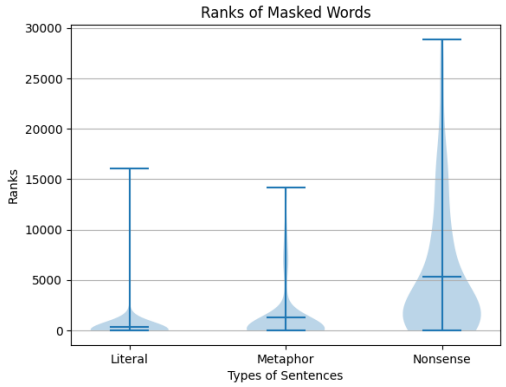

Figure 2: Ranks of target words among mask predictions sorted by probability of the sentences from Pedinotti et al. (2021) dataset.

Experiment 3

Table 3 summarizes the accuracy of the trained classifiers. We run several combinations categories. The three first numbers report the accuracy of models trained on the joined dataset consisting of 100 sentences for each category from Pedinotti et al. (2021) and 200 sentences from Cardillo et al. (2010) (or automatically generated) with 80/10/10 split for train/dev/test sets. The last experiment was conducted on a dataset with additional 200 metaphorical sentences from Cardillo et al. (2010) and 200 nonsensical from (O'Neill et al., 2020).

The Naive Bayes classifier received 22.5% accuracy when discriminating between metaphors and nonsense, suggesting that bag-of-words approach for the target classification task is not sufficient.

5 Discussion

The experiment results have demonstrated that language models can see the difference in plausibility between nonsense and metaphorical sentences. Such finding can be a useful probe when investigating what do models know about the language. The ability of models to distinguish between nonsense and metaphors (especially creative ones) suggest that the language models have an intuition that even highly unusual phrases/sentences can make sense.

The findings can also be brought up in a discussion about the nature of metaphors. While it could seem that, in order to understand some metaphor (and, it this way, to distinguish it from nonsense), extensive world-knowledge and an associative thinking is needed, our results suggest that, unless the models have also already acquired the aforementioned assets, metaphors can be distinguished from nonsense based on their linguistic form as well.

Furthermore, using a metaphor vs. nonsense classifier could be useful in ranking translated (literary) sentences, to see if the metaphors have been used correctly.

6 Future Work

While we were trying our best to ensure the training and testing data is free of unintended biases, further research would be needed to find out whether there really are no artifacts left. It is not clear whether the models are really relying on semantic acceptability in the case of our classifiers. It can also be that models are taking advantage of annotation artifacts when making decisions. One way to test for this would be to remove the target word from the sentences and try to train a classifier on the rest of the sentence.

7 Conclusion

The conducted experiments have shown that the current language models are able to pick the difference in plausibility between metaphorical and nonsensical sentences. The classifier between these two categories is also performing well, reaching about 75-88% accuracy (depending on the size of the training dataset). However, further research is needed to see whether this classification performance comes from distinguishing the semantic acceptability of the sentences, or if it is due to linguistic artifacts in the sentences that models can rely on when making the decision.

Limitations

For reliable results, the classification experiments should be repeated on a larger, more varied dataset, with extensive hyperparameter tuning and model comparison.

Acknowledgments

We would like to thank Reviewers 1 & 2 for taking the time and effort necessary to review the paper, and for their valuable suggestions.

References

Eileen Cardillo, Gwenda Schmidt-Snoek, Alexander Kranjec, and Anjan Chatterjee. 2010. Stimulus design is an obstacle course: 560 matched literal and metaphorical sentences for testing neural hypotheses about metaphor. *Behavior research methods*, 42:651–64.

Eileen Cardillo, Christine Watson, and Anjan Chatterjee. 2016. Stimulus needs are a moving target: 240 additional matched literal and metaphorical sentences for testing neural hypotheses about metaphor. *Behavior Research Methods*, 49.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ashim Gupta, Giorgi Kvernadze, and Vivek Srikumar. 2021. Bert amp; family eat word salad: Experiments with text understanding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12946–12954.

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. *CoRR*, abs/1907.11692.

Irving Massey. 2021. A new turing test: metaphor vs. nonsense. *AI & society*, 36(3):677–684.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126.

Erin O'Neill, Morgan Parke, Heather Kreft, and Andrew Oxenham. 2020. Development and validation of sentences without semantic context to complement the basic english lexicon sentences. *Journal of Speech, Language, and Hearing Research*, 63:3847–3854.

Paolo Pedinotti, Eliana Di Palma, Ludovica Cerini, and Alessandro Lenci. 2021. A howling success or a working sea? testing what BERT knows about metaphors. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 192–204, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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*, pages 2699–2712, Online. Association for Computational Linguistics.

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*, pages 30–36, Minneapolis, Minnesota. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

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