Context Matters: Enhancing Metaphor Recognition in Proverbs

Gamze Goren, Carlo Strapparava
University of Trento, FBK-irst
gamze.goren@studenti.unitn.it, strappa@fbk.eu

Abstract

Despite the remarkable achievements of Large Language Models (LLMs) in various Natural Language Processing tasks, their competence in abstract language understanding remains a relatively under-explored territory. Figurative language interpretation serves as ideal testbed for assessing this as it requires models to navigate beyond the literal meaning and delve into underlying semantics of the figurative expressions. In this paper, we seek to examine the performance of GPT-3.5 in zero-shot setting through word-level metaphor detection. Specifically, we frame the task as annotation of word-level metaphors in proverbs. To this end, we employ a dataset of English proverbs and evaluated its performance by applying different prompting strategies. Our results show that the model shows a satisfactory performance at identifying word-level metaphors, particularly when it is prompted with a hypothetical context preceding the proverb. This observation underscores the pivotal role of well-designed prompts for zero-shot settings through which these models can be leveraged as annotators for subjective NLP tasks.

Keywords: metaphors, proverbs, large language models

1. Introduction

Recently, Large Language Models (LLMs) has been the focus of attention for their remarkable achievements in sophisticated and complex NLP tasks (Brown et al., 2020; Radford et al., 2019; Dasgupta et al., 2022). One interesting domain that has yet to receive comprehensive scrutiny is the models' abilities in understanding abstract language constructs, particularly within the domain of figurative language. Figurative language is very common in everyday discourse, and they require to surpass straightforward literal interpretations. Proverbs are one of the most widely used source of figurative language, and they can be defined as fixed expressions conveying a well-established truth or a moral lesson in a short manner (Charteris-Black, 1995). They can also be considered as condensed expressions of cultural wisdom since they encapsulate generations of collective knowledge, offering insights into social values and shared experiences (Mieder, 1985).

A noteworthy characteristic of proverbs is their pervasive use of metaphors. Accurate interpretation of proverbs is rooted in metaphorical mappings, mapping the experiences from concrete domains onto abstract domains (Lakoff and Johnson, 1980). Furthermore, understanding the figurative meaning in proverbial expressions also assisted by several types of reasoning including analogical and cause-and-effect reasoning (Gibbs and Beitel, 1995). Therefore, successful identification of metaphors present in the proverbs serves as valuable testbed for evaluating language models' capabilities and limitations in abstract language understanding.

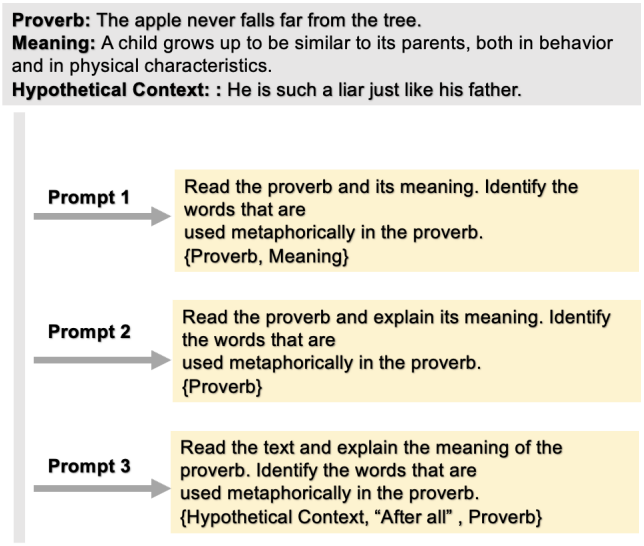

Figure 1: Prompt types

It is also significant to highlight that due to the intrinsic subjectivity and inherent ambiguity of metaphor interpretation, metaphor annotation tasks often exhibit lower levels of inter-annotator agreement, even amongst trained annotators (Sandri et al., 2023). The evaluation of models' performance in metaphor identification offers valuable insights into the feasibility of their potential to serve as automated annotators alternative to human annotators, particularly in subjective NLP tasks with large-scale or time-sensitive annotation requirements. Furthermore, deployment of automated annotators holds promise in mitigating potential biases inherent to human annotation process which

LREC-COLING 2024, pages 3825–3830
20-25 May, 2024.

can be influenced by personal experiences or cultural backgrounds, thereby enhancing the overall reliability and replicability of the annotation process.

This study aims to evaluate the inherent abstract language understanding capabilities of GPT-3.5, particularly in the context of word-level metaphor identification within proverbs. To achieve this, we designed and implemented different prompting strategies in zero-shot setting, and assessed their effectiveness in metaphor detection.

2. Related Work

Prior research in figurative language has encompassed a range of topics, including metaphor detection and generation (Leong et al., 2020; Pramanick et al., 2018; Gao et al., 2018; Yu and Wan, 2019; Chakrabarty et al., 2021). Additionally, certain studies have delved into cognitively-inspired methodologies for identification of metaphors (Turney et al., 2011; Tekiroğlu et al., 2015; Mykowiecka et al., 2018). For instance, Shutova et al. (2016) integrated combination of linguistic and visual embeddings while Swarnkar and Singh (2018) leveraged the contrast between the target word and its context. Mao et al. (2018) tackled the task of word-level metaphor detection by comparing a target word's context with synonyms and hypernyms retrieved from WordNet. Although metaphor generation remains relatively under-studied area, Chakrabarty et al. (2021) introduced an approach centered on the substitution of relevant words within a literal expression.

Regarding figurative language interpretation and comprehension, the main approach has involved framing it as paraphrasing task (Bizzoni and Lappin, 2018; Mao et al., 2018). However, recent advances in Large Language Models (LLMs) have ushered in a fresh wave of investigations. Recently, Prystawski et al. (2022) analyzed metaphor understanding in GPT-3 using chain-of-thought prompts inspired by psychological models while Wachowiak and Gromann (2023) probed GPT-3 in identifying the source domain of the metaphors. Liu et al. (2022) designed a Winograd-style task to evaluate the nonliteral language understanding and reasoning capacities of both auto-regressive and masked language models. In the most related recent work, authors introduced a dataset of proverbs paired with narrative context and benchmarked several LLMs on proverb recommendation given the narrative to test their abilities on abstract language understanding and analogical reasoning (Ghosh and Srivastava, 2022).

3. Dataset

We evaluate model's performance of word-level metaphor detection in proverbs on PROMETHEUS dataset (Özbal et al., 2016). The dataset consists of 1054 English proverbs and their equivalents in Italian. Proverbs are annotated with word-level metaphors, overall metaphoricity degree together with the meaning of the proverb. Annotation was carried out by multiple annotators and agreement among annotators reported as 0.76 and 0.74 for token-level metaphors and metaphoricity degree respectively. The authors of the dataset collected the English proverbs from a dictionary of proverbs while the meanings of proverbs were gathered from various sources including Wiktionary[1] and Free dictionary by Farlex[2]. As focus of our work is only English, we discarded Italian proverbs.

3.1. Dataset Expansion

A line of work in cognitive psychology and psycholinguistics has shown that illustrating proverbs with hypothetical context is a widely used strategy in humans and it facilitates the process of making connection between the proverb and its underlying meaning (Pasamanick, 1983; Gibbs and Beitel, 1995). In order to evaluate impact of providing a contextual illustration of the metaphorical mapping on the LLM's performance in detecting word-level metaphors, the current dataset was expanded with hypothetical context sentences that are appropriate to precede the proverb. For expanding the dataset with context sentences, first we collected example usages of proverbs in sentences from Free dictionary by Farlex. Proverbs which were not present in the dictionary therefore didn't have example sentences were discarded from the dataset. The remaining 891 proverbs, their meanings, and example sentences were divided into two and provided to two proficient English speaking NLP researchers for context sentences creation. They were asked to modify the example sentences in a way that it would create the hypothetical context for the proverb that is aligned with the meaning of the proverb and suitable to precede the proverb. After context sentences were created by one annotator, they were verified and modified whenever necessary by the other annotator.

4. Experimental Setup

In this section, we describe the model, types of prompts, and the metrics employed for evaluating the model's performance.

[1] https://en.wiktionary.org/
[2] http://idioms.thefreedictionary.com/

4.1. Model

In our experiments, we used OpenAI's GPT-3.5 model. In particular, we selected the most advanced DaVinci model, text-davinci-003 as subject of our analysis. While the precise parameter count is not disclosed by OpenAI, it is known that the model is fine-tuned with Reinforcement Learning from Human Feedback (RLHF). DaVinci is one of the most capable models in the GPT-3.5 family as it reported to outperforms other models on common benchmarks (Chen et al., 2023). The model was prompted using OpenAI's official API. We set the maximum number of tokens to 256 and the temperature parameter was set to 0 to have more precise results.

4.2. Prompt Types

Figure 1 shows the prompt types we employed. We designed and tested three different types of prompts specific to our task. In all prompts, the main instruction was "Identify the words that are used metaphorically in the proverb." Based on the type of the prompt, the model received a different instruction before this fixed instruction. To mirror human annotation process, we initially presented the proverb along with its meaning. Here, the model is first instructed to read the proverb and its meaning; then identify the metaphorical words in the proverb.

Subsequently, we drew inspiration from the zero-shot chain-of-thought (CoT) prompting approach, to implicitly force the model to engage in reasoning. Kojima et al. (2022) introduced zero-shot-CoT prompts which is constructed by adding the phrase "Let's think step by step" after the input to decompose the complex tasks and extract step-by-step reasoning without few-shot demonstrations. It has been shown that LLMs prompted with zero-shot-CoT yields to better performance on several reasoning benchmarks compared to zero-shot LLMs. In our case, we first tasked the model with providing the meaning of the proverb before proceeding to metaphor identification. By introducing this intermediate step, we anticipate the model to exhibit an improvement in word-level detection of metaphors, as asking to provide the meaning potentially guide the model towards utilizing reasoning.

Finally, we explored the prompts involving hypothetical contexts, as described in section 3.1. The objective of this investigation is to assess whether providing a contextual illustration of metaphorical mapping would enhance the model's ability to detect metaphorical words. This involved presenting the context and proverb as separate sentences, with the proverb immediately following the context sentence. Additionally, we included the phrase "*After all*" before each proverb to ensure a coherent and meaningful connection between two sentences, and to signal the proverb. Similar to the approach taken in the second prompt, the initial instruction to the model was to provide the meaning of the proverb.

Prompt Type	GTC	HTC	LTC
Proverb + Meaning	0.177	0.176	0.201
Only Proverb	0.371	0.363	**0.596**
Context + Proverb	0.565	0.484	**0.651**

Table 1: Word-level metaphor detection results: Ratio of overlapping token count to Ground Truth Token Count (GTC), Highest Token Count (HTC) and Lowest Token Count (LTC) among both annotations.

	Cohen's Kappa
Proverb + Meaning	0.009
Only Proverb	0.226
Context + Proverb	0.099

Table 2: Agreement between GPT-3.5 and human annotations for overall metaphoricity

4.3. Metrics

To evaluate the model's performance in word-level metaphor identification across different prompt types, we employed three distinct metrics, all of which incorporate a measure of overlap with human annotations.

The initial metric, denoted as *Ground Truth Token Count* (GTC), involves computing the ratio of detected words both by the model and humans to the number of ground truth words annotated by humans. This metric specifically evaluates the alignment between the model and human judgments, and offers a direct comparison point for the model's capacities. To ensure a balanced comparison and account for potential variation in the number of labeled words, *Highest Token Count* (HTC) and *Lowest Token Count* (LTC) metrics are also computed. While HTC measures the ratio of overlapping tokens to the maximum count of labeled words among the model's answers and human annotation sets, LTC metric computes such ratio relative to the minimum number of words from both sets.

In addition, we estimated the agreement between the model and human annotations for overall metaphoricty. Specifically, we considered a proverb labeled as metaphorical by the model if the model identified at least one token as metaphorical, and not metaphorical if it identified none as such. Given the absence of binary annotations from humans regarding overall metaphoricity, we deemed proverbs labeled with a metaphoricity degree of 'slightly metaphorical' and 'very metaphorical' as

metaphorical for the purposes of this assessment. The agreement between the model and humans was determined by Cohen's kappa coefficient.

5. Results

Table 1 presents the evaluation of DaVinci's performance in word-level metaphor detection across all prompt types using the metrics outlined in section 4.3. Notably, when the prompt included both the meaning and the proverb, the model exhibited a weak performance in identifying metaphorical tokens in comparison to human annotators. It poses a challenge for the model to interpret underlying figurative meaning of the proverb when prompted with a condition similar to those provided to humans. Conversely, the model's performance significantly improves when instructed to provide the meaning of the proverb before identifying metaphorical words($p < .05$). By including this intermediate step in the task, the model appears to be able to make more meaningful connections between the words and their metaphorical usage in the proverb. The inclusion of hypothetical context preceding the proverb seems to be the most effective prompt type as we obtained the best results under this condition across all metrics. The model is able to identify metaphorical tokens more precisely when the metaphors are contextually illustrated.

Our evaluation of overall metaphoricity reaffirms the model's limited capacity in identifying the figurative meaning of proverbs when prompted with their corresponding meanings (See Table 2). Intriguingly, introduction of hypothetical context, while leading to a satisfactory performance for token-level metaphor identification, does not enhance the performance in identifying whether a given proverb is metaphorical. One possible explanation of this observation could be attributed to the domain of metaphorical mapping; in cases where metaphorical mappings require higher level understanding of the physical world, the model may encounter difficulty in establishing links between the hypothetical context and the proverb. Further in-depth analysis are essential for understanding the interaction between the nature of metaphor domains and the performance of the model.

6. Conclusion

The goal of this paper was to evaluate the abstract language understanding capacity of large language models, specifically OpenAI's GPT-3.5 DaVinci model, in the context of word-level metaphor detection within English proverbs. Our findings reveal nuanced insights into the model's performance under different prompting conditions. We find that when prompted with the proverb and its corresponding meaning, the model faced challenges to identify metaphors. On the other hand, a notable improvement was observed when the model was instructed to provide the meaning of the proverb prior to identifying metaphorical tokens. This intermediate appeared to foster the model to utilize reasoning and make more meaningful connections. Furthermore, introduction of contextual illustration preceding the proverb in the prompt proved to be most effective prompt type for token-level metaphor detection. While this approach is promising for word-level detection, it did not enhance model's performance for detection of overall metaphoricity. We leave for future work the in-depth analysis of interaction between the model's performance and the metaphor domain. These insights also pave the way for future research in utilizing large language models for streamlined and efficient annotation process in subjective NLP tasks, particularly those encompassing figurative language.

Acknowledgements

We acknowledge the support of the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU.

7. Bibliographical References

Yuri Bizzoni and Shalom Lappin. 2018. Predicting human metaphor paraphrase judgments with deep neural networks. In *Proceedings of the Workshop on Figurative Language Processing*, pages 45–55, New Orleans, Louisiana. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Tuhin Chakrabarty, Xurui Zhang, Smaranda Muresan, and Nanyun Peng. 2021. MERMAID: Metaphor generation with symbolism and discriminative decoding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4250–4261, Online. Association for Computational Linguistics.

Jonathan Charteris-Black. 1995. Proverbs in communication. *Journal of Multilingual & Multicultural Development*, 16(4):259–268.

Xuanting Chen, Junjie Ye, Can Zu, Nuo Xu, Rui Zheng, Minlong Peng, Jie Zhou, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. How robust is gpt-3.5 to predecessors? a comprehensive study on language understanding tasks. *arXiv preprint arXiv:2303.00293*.

Ishita Dasgupta, Andrew K Lampinen, Stephanie CY Chan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. 2022. Language models show human-like content effects on reasoning. *arXiv preprint arXiv:2207.07051*.

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*, pages 607–613, Brussels, Belgium. Association for Computational Linguistics.

Sayan Ghosh and Shashank Srivastava. 2022. ePiC: Employing proverbs in context as a benchmark for abstract language understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3989–4004, Dublin, Ireland. Association for Computational Linguistics.

Raymond W Gibbs and Dinara Beitel. 1995. What proverb understanding reveals about how people think. *Psychological Bulletin*, 118(1):133.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

George Lakoff and Mark Johnson. 1980. Metaphors we live by. *University of Chicago, Chicago, IL*.

Chee Wee 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*, pages 18–29.

Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. 2022. Testing the ability of language models to interpret figurative language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4437–4452, Seattle, United States. Association for Computational Linguistics.

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)*, pages 1222–1231, Melbourne, Australia. Association for Computational Linguistics.

Wolfgang Mieder. 1985. Popular views of the proverb. *Proverbium*, 2(1985):109–143.

Agnieszka Mykowiecka, Aleksander Wawer, and Malgorzata Marciniak. 2018. Detecting figurative word occurrences using recurrent neural networks. In *Proceedings of the Workshop on Figurative Language Processing*, pages 124–127, New Orleans, Louisiana. Association for Computational Linguistics.

Gözde Özbal, Carlo Strapparava, and Serra Sinem Tekiroğlu. 2016. PROMETHEUS: A corpus of proverbs annotated with metaphors. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3787–3793, Portorož, Slovenia. European Language Resources Association (ELRA).

Judith Pasamanick. 1983. Talk does cook rice: Proverb abstraction through social interaction.

Malay Pramanick, Ashim Gupta, and Pabitra Mitra. 2018. An lstm-crf based approach to token-level metaphor detection. In *Proceedings of the Workshop on Figurative Language Processing*, pages 67–75.

Ben Prystawski, Paul Thibodeau, Christopher Potts, and Noah D Goodman. 2022. Psychologically-informed chain-of-thought prompts for metaphor understanding in large language models. *arXiv preprint arXiv:2209.08141*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Jezek. 2023. Why don't you do it right? analysing annotators' disagreement in subjective tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2428–2441, Dubrovnik, Croatia. Association for Computational Linguistics.

Ekaterina Shutova, Douwe Kiela, and Jean Maillard. 2016. Black holes and white rabbits: Metaphor identification with visual features. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 160–170, San Diego, California. Association for Computational Linguistics.

Krishnkant Swarnkar and Anil Kumar Singh. 2018. Di-LSTM contrast : A deep neural network for metaphor detection. In *Proceedings of the Workshop on Figurative Language Processing*, pages 115–120, New Orleans, Louisiana. Association for Computational Linguistics.

Serra Sinem Tekiroğlu, Gözde Özbal, and Carlo Strapparava. 2015. Exploring sensorial features for metaphor identification. In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 31–39, Denver, Colorado. Association for Computational Linguistics.

Peter Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 680–690, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Lennart Wachowiak and Dagmar Gromann. 2023. Does GPT-3 grasp metaphors? identifying metaphor mappings with generative language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1018–1032, Toronto, Canada. Association for Computational Linguistics.

Zhiwei Yu and Xiaojun Wan. 2019. How to avoid sentences spelling boring? towards a neural approach to unsupervised metaphor generation. 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 861–871, Minneapolis, Minnesota. Association for Computational Linguistics.