LREC-COLING 2024

Context Matters: Enhancing Metaphor Recognition in Proverbs

Gamze Goren and Carlo Strapparava

Motivation

Large Language Models (LLMs) have made significant progress:

- Great language proficiency.
- Can generalize to new tasks with few-shot and zero-shot learning.

However, can they navigate **beyond literal language?**

Motivation

Large Language Models (LLMs) have made significant progress:

- Great language proficiency.
- Can generalize to new tasks with few-shot and zero-shot learning.

However, can they navigate **beyond literal language?**

For exploring this, we evaluated abstract language understanding abilities of GPT-3.5 in zero-shot setting through word-level metaphor annotation in proverbial expressions by applying different prompting strategies.

This evaluation also offers valuable insights into the feasibility of their potential to serve as automated annotators.

Nature of Proverbial Expressions

Proverbs are:

- Fixed expressions conveying a well-established truth or a moral lesson in a short manner (Charteris-Black, 1995).
- Mainly characterized by their pervasive use of metaphors; they are commonly metaphorical.

Therefore, accurate interpretation of proverbs rely on:

Nature of Proverbial Expressions

Proverbs are:

- Fixed expressions conveying a well-established truth or a moral lesson in a short manner (Charteris-Black, 1995).
- Mainly characterized by their pervasive use of metaphors; they are commonly metaphorical.

Therefore, accurate interpretation of proverbs rely on:

- Metaphorical mapping of experiences from concrete domains onto abstract domains (Lakoff and Johnson, 1980).
- Analogical and cause-and-effect reasoning processes for drawing parallels between concepts (Gibbs and Beitel, 1995).

Related Work

Figurative Language in Pre-LLM Era:

- Several techniques were explored for metaphor detection and interpretation, including combining different embeddings of different modalities (Shutova et al., 2016), as well as paraphrasing metaphors into their literal counterparts (Mao et al., 2018).

Related Work

Figurative Language in Pre-LLM Era:

- Several techniques were explored for metaphor detection and interpretation, including combining different embeddings of different modalities (Shutova et al., 2016), as well as paraphrasing metaphors into their literal counterparts (Mao et al., 2018).

LLMs and Figurative Language Understanding:

- Ghosh and Srivastava (2022) benchmarked several LLMs on proverb recommendation given a narrative context.
- More recently,Prystawski et al. (2022) analyzed metaphor understanding in LLMs with chain-of-thought prompts inspired by psychological models, while Wachowiak and Gromann (2023) focused on identifying metaphor source domains.

Dataset

The model was evaluated on dataset of English proverbs, *PROMETHEUS* which consists of 1054 English proverbs annotated with word-level metaphors and overall metaphoricity degree (Özbal et al., 2016).

[1] Free dictionary by Farlex

Dataset

The model was evaluated on dataset of English proverbs, *PROMETHEUS* which consists of 1054 English proverbs annotated with word-level metaphors and overall metaphoricity degree (Özbal et al., 2016).

For the purposes of prompt manipulation in the current experiment, *PROMETHEUS* is expanded with hypothetical context sentences that are appropriate to precede the proverb through additional data collection [1] and annotation.

[1] Free dictionary by Farlex

Dataset

The model was evaluated on dataset of English proverbs, *PROMETHEUS* which consists of 1054 English proverbs annotated with word-level metaphors and overall metaphoricity degree (Özbal et al., 2016).

For the purposes of prompt manipulation in the current experiment, *PROMETHEUS* is expanded with hypothetical context sentences that are appropriate to precede the proverb through additional data collection [1] and annotation.

The final version of *PROMETHEUS* includes 891 proverbs, their meanings, word-level metaphor annotations, and their hypothetical context sentences.

[1] Free dictionary by Farlex

Model

- In the experiments, we utilized OpenAI's GPT-3.5 DaVinci model (text-davinci-003).
- The model was prompted via OpenAI's official API, with a maximum token limit of 256 and a temperature parameter set to 0 for precision.

Prompt Types

Prompt Types

1. **Proverb + Meaning**
   - Presented the proverb together with its meaning to simulate human annotation process.

Prompt Types

1. **Proverb + Meaning**
   - Presented the proverb together with its meaning to simulate human annotation process.

2. **Only Proverb**
   - Instructed the model to provide the proverb's meaning before identifying metaphorical words for encouraging reasoning.

Prompt Types

1. **Proverb + Meaning**
   - Presented the proverb together with its meaning to simulate human annotation process.

2. **Only Proverb**
   - Instructed the model to provide the proverb's meaning before identifying metaphorical words for encouraging reasoning.

3. **Context + Proverb**
   - The proverb is presented after the hypothetical context to provide contextual illustration of the metaphorical mapping.

Prompt Types

Proverb: The apple never falls far from the tree.
Meaning: A child grows up to be similar to its parents, both in behavior and in physical characteristics.
Hypothetical Context: : He is such a liar just like his father.

Prompt 1 →
Read the proverb and its meaning. Identify the words that are
used metaphorically in the proverb.
{Proverb, Meaning}

Prompt 2 →
Read the proverb and explain its meaning. Identify the words that are
used metaphorically in the proverb.
{Proverb}

Prompt 3 →
Read the text and explain the meaning of the proverb. Identify the words that are
used metaphorically in the proverb.
{Hypothetical Context, “After all” , Proverb}

Metrics

We employed three metrics incorporating human annotations for **word-level metaphor** detection:

Metrics

We employed three metrics incorporating human annotations for **word-level metaphor** detection:

1. Ground Truth Token Count (GTC): Ratio of detected words by model and humans to total human annotations.

Metrics

We employed three metrics incorporating human annotations for **word-level metaphor** detection:

1. Ground Truth Token Count (GTC): Ratio of detected words by model and humans to total human annotations.
2. Highest Token Count (HTC): Ratio of overlapping tokens to maximum count of labeled words between model and human sets.

Metrics

We employed three metrics incorporating human annotations for **word-level metaphor** detection:

1. Ground Truth Token Count (GTC): Ratio of detected words by model and humans to total human annotations.
2. Highest Token Count (HTC): Ratio of overlapping tokens to maximum count of labeled words between model and human sets.
3. Lowest Token Count (LTC): Ratio of overlapping tokens to minimum number of words between model and human sets.

Metrics

We employed three metrics incorporating human annotations for **word-level metaphor** detection:

1. Ground Truth Token Count (GTC): Ratio of detected words by model and humans to total human annotations.
2. Highest Token Count (HTC): Ratio of overlapping tokens to maximum count of labeled words between model and human sets.
3. Lowest Token Count (LTC): Ratio of overlapping tokens to minimum number of words between model and human sets.

In addition, we estimated the agreement between the model and human annotations for **overall metaphoricity** using *Cohen's kappa coefficient*.

Results

The inclusion of hypothetical context led to a significant improvement in word-level metaphor detection whereas its impact on overall metaphoricity was less pronounced.

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.

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

Table 2: Agreement for overall metaphoricity.

Conclusions

- Our findings 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.
- While a notable gap still exists between the figurative language abilities of humans and the model, careful prompt design offers a promising avenue for narrowing this disparity in zero-shot settings. The model shows potential as an automated annotator for word-level metaphor annotation.
- Future work could investigate the interaction between the metaphor domains and model's performance as overall metaphoricity remains a challenge for the model.

References

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

Ghosh, S. and Srivastava, S. (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.

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

Lakoff, G. and Johnson, M. (1980). Metaphors we live by. *University of Chicago, Chicago, IL*.

Mao, R., Lin, C., and Guerin, F. (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.

Özbal, G., Strapparava, C., and Tekiroğlu, S. S. (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).

Prystawski, B., Thibodeau, P., Potts, C., and Goodman, N. D. (2022). Psychologically-informed chain-of-thought prompts for metaphor understanding in large language models. *arXiv preprint arXiv:2209.08141*.

Shutova, E., Kiela, D., and Maillard, J. (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.

Wachowiak, L. and Gromann, D. (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.