A big ant or a small elephant: metaphor interpretation on large language models

Luiz Matos, Aline Paes

lfmatosmelo@id.uff.br, alinepaes@ic.uff.br

Machine Learning and Learning from Language (MeLLL IC-UFF)

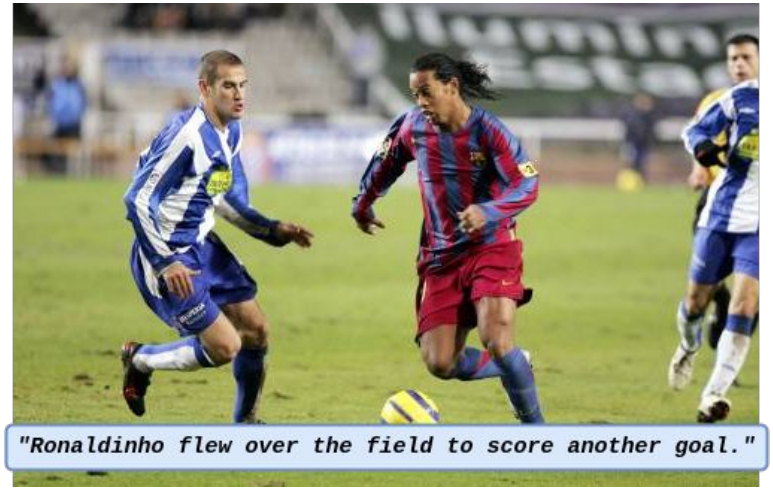
Universidade Federal Fluminense

Instituto de Computação

LXAI

Problem

- Figurative language (FL) is crucial for communication
- Metaphor is one of the most creative manifestation of FL



"Ronaldinho flew over the field to score another goal."

"The birds flew over the fields."

Code available at: https://github.com/lfmatosm/metaphor-interpretation-on-llms

Problem

- Previous work identified that LMs still face challenges to deal with metaphors[1]
- Although they also identified that LMs are aware of them[1] [2] [3]
- **What is the extension of their knowledge regarding metaphors? And how LMs of different size handle them?**
    a. Theory: conceptual metaphors (CM)
    b. LLM flavors: bigger closed-source vs. smaller open-source
- The linguistic theory of conceptual metaphors define them as mappings between source and target domains[4]

1. Aghazadeh et al., 2022
2. Jang et al., 2023
3. Wachowiak and Gromann, 2023
4. Lakoff and Johnson, 1980

Datasets and tasks

- Metaphor interpretation and detection datasets used are TroFi[5], VUA Verbs, VUA POS[6], Metaphor List and LCC's English version[7]
- TroFi, VUA Verbs/POS: include sentence and metaphor presence indicator pairs
- Metaphor List, LCC: include sentences and conceptual metaphor (CM) information

Classification

Sentence: I 'm going to sleep on it.
Question: Is the sentence metaphoric?
Answer: *Yes*

Source/Target domain inference

Context: In linguistics, conceptual metaphors consists of understanding a given concept in terms of another
Task: Extract the {source,target} domain from the sentence
Sentence: Out of the lap of luxury into the armpit of poverty.
{Target,Source} domain: human body
Answer: *poverty*

Source/Target lexeme inference

Context: In linguistics, conceptual metaphors consists of understanding a given concept in terms of another
Task: Extract the {source,target} lexeme from the sentence
Sentence: Out of the lap of luxury into the armpit of poverty.
{Target,Source} lexeme: poverty
Answer: *armpit*

5. Birke and Sarkar, 2006
6. Steen et al., 2010
7. Mohler et al., 2016

Experiments

- Llama2-7B is the baseline, and comparisons are made to GPT-3 as-is to determine the ideal performance to be achieved by the former
- Fine-tuning performed for Llama2-7B for each task and dataset
- Low-resource environment -> fine-tuning with low rank adapters and quantization
- For fine-tuning: only prompt samples alongside desired answers, no few-shot examples are prefixed
- For testing: few-shot examples were prepended on input prompts
    a. [2, 12] examples prefixed for each sample
    b. Number determined on a (dataset, model, task) basis
    c. No global number of samples was defined for tasks

Preliminary results

- Classification: FT Llama is better for most metrics, but results on TroFi indicate strong misclassification

Model	Dataset											
	TroFi				VUA Verb				VUA POS			
	f1	prec	rec	acc	f1	prec	rec	acc	f1	prec	rec	acc
GPT-3	0.59	**0.57**	0.61	**0.58**	0.59	0.57	0.61	0.57	0.56	0.53	**0.60**	0.54
Llama2-7B	0.61	0.55	0.70	0.56	**0.60**	0.57	0.62	0.58	0.55	0.52	0.58	0.52
+ FT	**0.67***	0.50*	**1.00***	0.50*	**0.65**	**0.64**	**0.67**	**0.65**	**0.58**	**0.59**	0.57	**0.59**

Preliminary results

- CM domain inference: target domain was easier to infer
- CM lexeme inference: GPT-3 with best results overall across tasks
- FT Llama2-7B responses often included gold label followed by an excerpt of the input prompt context prefix

Model	**Dataset**					
	Metaphor List		**LCC (en)**		**LCC (en)**	
	SD	**TD**	**SD**	**TD**	**SL**	**TL**
	$\cos\theta \pm \sigma$	$\cos\theta \pm \sigma$	$\cos\theta \pm \sigma$	$\cos\theta \pm \sigma$	$\cos\theta \pm \sigma$	$\cos\theta \pm \sigma$
GPT-3	0.51 ± 0.21	$\mathbf{0.60} \pm 0.24$	0.65 ± 0.26	$\mathbf{0.84} \pm 0.27$	$\mathbf{0.84} \pm 0.27$	$\mathbf{0.88} \pm 0.26$
Llama2-7B	0.49 ± 0.12	0.55 ± 0.14	0.55 ± 0.13	0.64 ± 0.13	0.58 ± 0.18	0.63 ± 0.17
+ FT	$\mathbf{0.52} \pm 0.11$	0.58 ± 0.15	$\mathbf{0.70} \pm 0.17$	0.71 ± 0.13	0.69 ± 0.13	0.66 ± 0.18

Conclusion and future works

- Metaphor knowledge do exist in LLMs, varying between models
- Models achieved results at least slightly above chance on all tasks
- Improvement areas: hallucinations on responses and poor performance on TroFi
- Mapping available metaphor knowledge inside the model architecture is ongoing work
- Exploration of smaller and linguistically aware LM architectures
- External knowledge bases for interpretability and explainability are also avenues to explore

Acknowledgments

- Research financed by
  - CNPq (National Council for Scientific and Technological Development)
  - FAPERJ - Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro
  - LatinX in AI organizers provided a registration grant
- We are also deeply thankful for the anonymous reviewers' comments and suggestions

References

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

[Aghazadeh et al., 2022] Ehsan Aghazadeh, Mohsen Fayyaz, and Yadollah Yaghoobzadeh. 2022. Metaphors in pre-trained language models: Probing and generalization across datasets and languages. arXiv preprint arXiv:2203.14139.

[Jang et al., 2023] Hyewon Jang, Qi Yu, and Diego Frassinelli. 2023. Figurative language processing: A linguistically informed feature analysis of the behavior of language models and humans. In Findings of the Association for Computational Linguistics: ACL 2023, pages 9816–9832.

[Wachowiak and Gromann, 2023] 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.

[Birke and Sarkar, 2006] Julia Birke and Anoop Sarkar. 2006. A clustering approach for nearly unsupervised recognition of nonliteral language. In 11th Conference of the European Chapter of the Association for Computational Linguistics, pages 329–336.

[Steen et al., 2010] Gerard Steen, Aletta G Dorst, J Berenike Herrmann, Anna Kaal, Tina Krennmayr, Trijntje Pasma, et al. 2010. A method for linguistic metaphor identification. Amsterdam: Benjamins.

[Mohler et al., 2016] Michael Mohler, Mary Brunson, Bryan Rink, and Marc Tomlinson. 2016. Introducing the lcc metaphor datasets. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 4221–4227.

Thank you! Questions?

lfmatosmelo@id.uff.br, alinepaes@ic.uff.br

Code, slides and paper found here: