LDK 2023

MEAN: Metaphoric Erroneous ANalogies dataset for PTLMs metaphor knowledge probing

Lucia Pitarch
Universidad de Zaragoza
lpitarch@unizar

Jorge Bernad
Universidad de Zaragoza
jbernad@unizar.es

Jorge Gracia
Universidad de Zaragoza
jogracia@unizar.es

Abstract

Despite significant progress obtained in Natural Language Processing tasks thanks to Pre-Trained Language Models (PTLMs), figurative knowledge remains a challenging issue. This research sets a milestone towards understanding how PTLMs learn metaphoric knowledge by providing a novel hand-crafted dataset, with metaphoric analogy pairs where per correct analogy pair, other three erroneous ones are added controlling for the semantic domain and the semantic attribute. After using our dataset to fine-tune SoTa PTLMs for the multiclass classification task we saw that they were able to choose the correct term to fit the metaphor analogy around the 80% of the times. Moreover, thanks to the added erroneous examples on the dataset we could study what kind of semantic mistakes was the model making.

1 Introduction

Metaphors are not only very common devices but also key elements in language. They both help us express ourselves and shape the way we think by using a concept to reference and delimit another (Lakoff and Johnson, 1980).

For instance, let's look at the following example extracted from The Guardian:

> The intriguing echo of Eliza in thinking about ChatGPT is that people regard it as **magical** even though they know how it works – as a **"stochastic parrot"** (in the words of Timnit Gebru, a well-known researcher) or as **a machine for "hi-tech plagiarism"** (Noam Chomsky). (Naughton, 2023)

In the same paragraph three views about ChatGPT[1] are compared: either it is conceived as a magical device, as a 'stochastic parrot' meaning it only repeats statistical patterns, or as a plagiarism tool. The metaphors and narratives we use to talk about Artificial Intelligence tools such as GPT have a huge impact on the sentiment we have towards them, being an already flagged concern at the European Parliament (Boucher, 2021).

Despite the pervasiveness and impact of metaphors in language and culture, processing them remains challenging for Natural Language Processing. Approaches taken towards them have shifted from pattern and statistical-based discovery since Shutova et al. (Shutova, 2015), towards Language Model exploitation for their discovery and interpretation (Ge et al., 2022). While the second approach is providing more efficient models and accurate results, in comparison to pattern-based methods it lacks interpretability. Moreover, it has been stated that PTLMs lack figurative knowledge (Liu et al., 2022) and have trouble processing it (Czinczoll et al., 2022). Though uncovering the kind of knowledge PTLMs encode has been a major concern since their origins (Petroni et al., 2019), attention to the figurative knowledge they keep has just gained attention in the last year. And if interpretability is a major concern in the Artificial Intelligence community (Bender et al., 2021) it should be even more relevant when treating metaphors, as they are especially sensitive devices that can be used to change the way we perceive the world (Semino et al., 2017).

At the moment, questions such as the following ones are being researched:

1. Do PTLMs encode figurative knowledge? (Liu et al., 2022; Aghazadeh et al., 2022)

2. Do PTLMs have figurative analogical reasoning? (Czinczoll et al., 2022; Chen et al., 2022)

3. What kind of figurative knowledge is the most challenging one? (Liu et al., 2022)

[1] Open AI's generative large language model

Our work follows the goal of understanding how PTLMs process figurative language, particularly the one dealing with metaphors, it adds a new research question to the ones already addressed in the literature, namely: 'How do PTLMs acquire figurative knowledge?', and contributes towards it in the following ways:

1. We provide MEAN, a novel manually curated dataset[2] with selected metaphoric analogies from MetaNet (Dodge et al., 2015) enriched with erroneous examples. Its main aim is to uncover what aspects of the metaphor PTLMs learn.

2. We test our dataset on the metaphoric analogy completion task and provide novel baselines for it.

3. We obtain promising results in the metaphor analogy task, suggesting PTLMs after fine-tuning can acquire semantic inference abilities for metaphor interpretation tasks.

2 Related Work

Probing language models to understand what linguistic and common ground knowledge they encode has been a major research line since 2019 with the arrival of Pre-Trained Language Models with transformer architecture (PTLMs) (Devlin et al., 2019). Simultaneously, computational metaphor processing has also benefited from such PTLMs and regained attention, leading to huge advances in metaphor identification, interpretation, and generation tasks (Ge et al., 2022; Rai and Chakraverty, 2020). Yet, just very recently, in 2022, these two interests are being aligned (PTLMs probing and computational metaphor processing), resulting in works such as (Liu et al., 2022; Chen et al., 2022; Czinczoll et al., 2022; Aghazadeh et al., 2022), where researchers try to uncover the figurative knowledge encoded in PTLMs.

When conducting probing tests in metaphor detection tasks, Aghzadeh et al. (2022), came to the conclusion that PTLMs do encode figurative knowledge, particularly in their middle layers, yet other authors (Liu et al., 2022) when experimenting with probing in metaphor generation and interpretation tasks highlight the inability of PTLMs to capture figurative language. The mentioned works probe PTLMs in fill in the mask tasks. This kind of setting has as limitation that several words can correctly fill in the gap in the sentence, and if just one or two options are given as gold standard the possibilities of not having a match between the predicted token and the gold one are high. The solutions they apply to minimize this effect are using Mean Reciprocal Ranking metrics and (Chen et al., 2022; Czinczoll et al., 2022) also search if the synonyms of the predicted tokens match their gold standard. Additionally, the fill-in-the-mask setting, has trouble dealing with multi-words, as only one token is selected to fill in the mask, yet metaphoric expressions are usually multi-words. Thus, the experimental setting we choose is more similar, though still different to the one proposed by Liu et al. (2022) who instead of conducting a fill-in-the-mask task, perform classification experiments. Particularly they provide as the first part of the sentence a verbalized metaphor and as the second part of the sentence the verbalized explanation of the metaphor. Given the metaphor and two possible explanations, the model has to select the best fit between both. In their experiment, they claim that even if in zero-shot environment figurative language understanding is extremely challenging for PTLMs, they can in fact learn it after some fine-tuning. Moreover, by annotating the kind of background knowledge needed to understand the inputted metaphors, they observe object and commonsense metaphors were easier to interpret while sarcastic metaphors were the most difficult ones. The later research is the most similar to our own one, as it focuses on probing the knowledge of figurative language in PTLMs through a metaphor interpretation task, while they focus on paraphrasing we focus on metaphoric inference by the completion of metaphoric analogies. Moreover, we explore where the semantic challenge relies (either on the semantic domain or attribute) by manually selecting the errors.

3 MEAN Dataset

If we understand metaphor as a linguistic device used to express something in terms of another thing (Lakoff and Johnson, 1980), this means two conceptual domains are involved, the source domain is the one that the speaker is using in the text and the target domain is the implicit one, trying to be expressed.[3] Source domain is expressed in the

[2] Our code and dataset are openly available at https://github.com/sid-unizar/MEAN.git

[3] In metaphor literature conceptual domains are understood as the background knowledge needed to understand

text by particular lexical entries which make reference to different elements involved in the source domain. These elements have their corresponding elements in the target domain, which is implicitly referenced through the explicit expression of the source domain elements. Such process of drawing correspondences between the source and target domain in a metaphor through the expression of the individual elements involved is called *metaphor mapping* (Kövecses, 2016). A natural way of representing such correspondences and inputting them to PTLMs is via analogical reasoning as in (Czinczoll et al., 2022). That is, we can rewrite the metaphor mapping as "source domain is to target domain what source element is to target element".

For instance in this quote from an article in Nature: ‘Although OpenAI has tried to put guard rails on what the chatbot will do, users are already finding ways around them.’[4] The metaphor being expressed there would be: ‘Artificial Intelligence is a moving vehicle’, the source domain would be ‘moving vehicle’ and the target domain ‘Artificial Intelligence’, the lexical entries being used metaphorically in the text (or in other words, the source element) are ‘putting guard rails around’ and ‘them’ in ‘users are already finding ways around them’ the metaphoric mapping from this lexical entry to its correspondent one in the Artificial Intelligence domain would be ‘firewall’ or ‘security measures’ to avoid things such as bias or missusage of the tool.

Our dataset consists of analogy pairs where the first part of the analogy contains the metaphor source and target domains and the second part consists of the individual lexical entries that could serve as instances in the text of the metaphor. Both the source and target domains and the first set of lexical entries proposed in the dataset are a subsample extracted from MetaNet (Dodge et al., 2015). MetaNet is a repository of metaphors and frames containing almost 700 conceptual metaphors, design to aid the computational exploration of corpora. From them we just selected the ones which had assigned one or more metaphor mappings between the different frame entities and which had the pattern ‘A are B’. We extend MetaNet data by adding curated erroneous endings to the analogy. The three erroneous target elements per analogy were manually selected following linguistic criteria to control what the model is learning and to which semantic aspect of the metaphor it is paying attention to. If the criteria for a target element to properly fit the analogy is that it has to share the semantic domain with the target domain and the semantic attribute with the source element, then erroneous examples are when one of these criteria fails. We consider as semantic domain the general category to which the target domain and target element belong. Semantic attribute is the specific role that an individual element within that domain might play; for instance the semantic domain of ‘hospital’ would be ‘healthcare’ and the role it plays inside the healthcare domain would be ‘location’. In our dataset, an element is added per analogy for each of the three erroneous possibilities found when these criteria are not met. Namely:

1. the target element fits the same semantic domain as the target domain of the metaphor, but has a different attribute than the proposed source element (shortened as sDdA in Tables 1 and 4);
2. the target element shares the same attribute as the source element, but does not share the semantic domain with the target domain (shortened as dDsA in Tables 1 and 4);
3. or it has both different semantic domains and attributes from the needed ones (shortened as dDdA in Tables 1 and 4).

The resulting dataset contains 166 analogies (composed of a source domain, a target domain, a source element, a four target element candidates withing which just one is correct) made for 71 different metaphors (composed by a source and target domain pair) and 100 different source and target metaphor domains. At the moment the dataset exists just for English. A sample of our dataset can be found in Table 1.

4 Experiments

In this section we describe the different choices taken for fine tuning the model and testing our approach.

4.1 Multiple choice task

We fine-tune and test BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) models both in their large and base versions for multiple choice

a text (Clausner and Croft, 1999).

[4] In https://tinyurl.com/NatureAnon2023

Source Domain	Target Domain	Source Element	Target Element (Gold)	Target Element (Erroneous)		
				sDdA	dDsA	dDdA
anger	fire	anger level	fire intensity	wood	flood magnitude	unicorn
taxation	punishment	taxer	punisher	prison	vampire	amusement

Table 1: Sample of the dataset.

classification using hugging's face library[5]. The task consists of providing the model with a beginning and four possible endings of a sentence, among which just one is correct. The first part of the sentence is the verbalized first pair of the analogy with the source and target domains of the metaphor. The second part of the sentence contains the individual source and target elements of the metaphor, where the last element (target element) varies to cover the four possible choices of our dataset. In Table 2 the different templates to verbalize the analogies are summarized. We experiment with three different verbalization which range from minimal templates with just punctuation to larger templates with more complex phrasings, following previous literature on prompting (Schick and Schütze, 2022).

Start template	End template	id
’ W1 ’ : ’ W2 ’	’ W3 ’ : ’ W4	T1
’ W1 ’ is to ’ W2 ’	what ’ W3 ’ is to ’ W4 ’.	T2
If ’ W1 ’ is like ’ W2 ’,	then ’ W3 ’ is like ’ W4 ’.	T3

Table 2: Templates and identifiers used along the paper to identify them. In order to create an input sequence for a language model, the start and end templates are joined with the sep token, and, in the case of BERT models, the tokens of the start and end templates have a different token type.

This kind of task in comparison with fill-in-the-mask settings, benefits from being able to deal with whole sequences of tokens, facilitating dealing with multiword expressions. Moreover, as the answer is selected from a closed set of items we can better control the model output and what it is learning by biasing each of the possible answers with a particular linguistic restriction (in our case different domain and attribute selection).

As our dataset is very small, the provided results for the PTLMs consist of the mean accuracy of a 10-fold cross-validation and a 95% confidence interval for the mean accuracy calculated by bootstrapping (Efron, 1979).

Fine tuning setting. To fine-tune the models, we used the following hyperparameters: batch size of 8, Adam optimizer with weight decay of 0.01 and learning rate of 2e-5, no warm-up, and training during 5 epochs.

4.2 Baselines

To compare whether fine-tuning with the metaphors provided in our dataset improved the model's output we compare the results obtained to the static 300-dimensional embeddings from three different models: GloVe (Pennington et al., 2014), word2vec (Mikolov et al., 2013), and fastText (Bojanowski et al., 2017). All models were retrieved via Gensim (Rehurek and Sojka, 2011). To avoid Out of Vocabulary words the following strategy, similar to the one in (Speer et al., 2017), was followed: for a word, remove the last character until the word is found in the model. To deal with multiword expressions, the mean of the word embeddings were calculated.

Since the problem is posed as an analogy task, the cosine similarity is used to discover the best target element from a set of predefined ones following (Mikolov et al., 2013). That is, given a source and target domain word embeddings, s_d and t_d, a source element s_e, and a set of target elements $T = \{t_{e_1}, \ldots, t_{e_k}\}$, solve the following equation:

$$argmax_{t_e \in T}\{\cos(s_e + t_d - s_d, t_e)\}$$

4.3 Error analysis

Additionally to analyse with which semantic feature the model is having more trouble (attribute or domain distinction) when choosing the correct analogy we report percentages of the different error types made by the model.

5 Results and discussion

Table 3 shows the accuracy of RoBERTa and BERT models for each of the provided templates and compares them to GloVe, word2vec, and fastTest baselines. A huge improvement can be observed when finetuning the model and shifting from static to contextual embeddings. The high results obtained

[5] Original code, setup and documentationfrom hugging face at: `https://huggingface.co/docs/transformers/tasks/multiple_choice`

point to the ability of PTLMs to learn metaphorical analogy inference, coincidentally with the conclusions obtained by (Liu et al., 2022).

	Acc.	CI
Baselines		
GloVe	32.5	-
word2vec	33.7	-
fastText	45.8	-
BERT large		
T1	85.5	(81.5, 89.7)
T2	69.3	(52.5, 83.6)
T3	87.3	(83.7, 91.0)
RoBERTa large		
T1	84.9	(72.3, 93.3)
T2	86.7	(83.2, 90.8)
T3	74.7	(65.6, 83.8)
BERT base		
T1	84.3	(76.5, 91.1)
T2	78.9	(68.5, 87.2)
T3	84.9	(78.6, 90.8)
RoBERTa base		
T1	75.3	(61.7, 85.1)
T2	80.1	(74.4, 86.2)
T3	88.0	(81.7, 93.5)

Table 3: Results for baselines and fine-tuned PTLMs. The reported accuracy for PTLMs is the mean of a 10-fold cross-validation. For these latter cases, it is also reported a 95% confidence interval (CI) calculated by bootstrapping.

In Table 4 the percentages per error type in the classification are shown. On all models and templates, the most errors were made by predicting a target element that shared the same domain as the source element but had a different attribute than the target domain. This could point to a lesser knowledge of PTLMs regarding semantic roles. Further research should be done on this line. In future work, we will experiment with injecting this kind of linguistic knowledge into PTLMs models for metaphor interpretation tasks.

6 Conclusions and Future Work

By experimenting with our novel dataset with selected erroneous answers: MEAN, we conclude PTLMs can learn, through fine tuning, metaphoric analogical reasoning, improving the baselines stated by static embeddings. We also observed most errors were made by confusing the needed attribute of the word to meet the metaphor analogy restrictions and thus we propose the injection of such linguistic features as a possible research line for future work. Additionally, in further iterations of this research line, we would like to expand our dataset with more analogies and to other languages such as Spanish.

	sDdA	dDsA	dDdA
Model			
BERT base	76.2	20.0	3.8
BERT large	58.8	33.8	7.5
RoBERTa base	84.5	12.7	2.8
RoBERTa large	56.0	29.9	14.2
Templates			
T1	64.0	27.2	8.8
T2	72.3	23.5	4.2
T3	63.6	24.8	11.6
Total (all models and templates)			
	66.6	25.2	8.2

Table 4: Percentage of errors per error type, calculated for each model, template and totals.

Aknowledgements

Supported by the Spanish project PID2020-113903RB-I00 (AEI/FEDER, UE), by DGA/FEDER, by the *Agencia Estatal de Investigación* of the Spanish Ministry of Economy and Competitiveness and the European Social Fund through the “Ramón y Cajal” program (RYC2019-028112-I), and by the EU research and innovation program HORIZON Europe through the “4D PICTURE" project under grant agreement 101057332.

References

Ehsan Aghazadeh, Mohsen Fayyaz, and Yadollah Yaghoobzadeh. 2022. Metaphors in pre-trained language models: Probing and generalization across datasets and languages. In *Annual Meeting of the Association for Computational Linguistics*.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Philip Boucher. 2021. What if we chose new metaphors for artificial intelligence?

Weijie Chen, Yongzhu Chang, Rongsheng Zhang, Jiashu Pu, Guandan Chen, Le Zhang, Yadong Xi, Yijiang Chen, and Chang Su. 2022. Probing simile knowledge from pre-trained language models. *arXiv preprint arXiv:2204.12807*.

Timothy C. Clausner and W. Bruce Croft. 1999. Domains and image schemas*. *Cognitive Linguistics*, 10(1):1–31.

Tamara Czinczoll, Helen Yannakoudakis, Pushkar Mishra, and Ekaterina Shutova. 2022. Scientific and creative analogies in pretrained language models. *arXiv preprint arXiv:2211.15268*.

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.

Ellen K Dodge, Jisup Hong, and Elise Stickles. 2015. Metanet: Deep semantic automatic metaphor analysis. In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 40–49.

B. Efron. 1979. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1 – 26.

Mengshi Ge, Rui Mao, and Erik Cambria. 2022. A survey on computational metaphor processing techniques: From identification, interpretation, generation to application. Preprint.

Zoltán Kövecses. 2016. Conceptual metaphor theory. In *The Routledge handbook of metaphor and language*, pages 31–45. Routledge.

George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. The University of Chicago Press.

Emmy Liu, Chen Cui, Kenneth Zheng, and Graham Neubig. 2022. Testing the ability of language models to interpret figurative language. *arXiv preprint arXiv:2204.12632*.

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. *arXiv preprint arXiv:1907.11692*.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR*.

John Naughton. 2023. Chatgpt isn't a great leap forward, it's an expensive deal with the devil.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*.

Sunny Rai and Shampa Chakraverty. 2020. A survey on computational metaphor processing. *ACM Computing Surveys (CSUR)*, 53(2):1–37.

Radim Rehurek and Petr Sojka. 2011. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2).

Timo Schick and Hinrich Schütze. 2022. True few-shot learning with prompts—a real-world perspective. *Transactions of the Association for Computational Linguistics*, 10:716–731.

Elena Semino, Zsófia Demjén, Andrew Hardie, Sheila Payne, and Paul Rayson. 2017. *Metaphor, cancer and the end of life: A corpus-based study*. Routledge.

Ekaterina Shutova. 2015. Design and evaluation of metaphor processing systems. *Computational Linguistics*, 41:579–623.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 4444–4451. AAAI Press.