Probing, Generalization and Application of Metaphorical Knowledge in Pre-trained Language Models

Anonymous ACL-IJCNLP submission

Abstract

Human languages are full of metaphorical expressions. Metaphors help people understand the world by connecting new concepts and domains to more familiar ones. Large pre-trained language models (PLMs) are therefore assumed to encode metaphorical knowledge useful for NLP systems when processing language. In this paper, we investigate this hypothesis for PLMs by probing the metaphoricity knowledge in their encodings, by measuring the cross-lingual and cross-dataset generalization of this knowledge, and by analyzing the application of this knowledge when generating metaphorical expressions. We present studies in multiple metaphoricity detection datasets and four languages (i.e., English, Spanish, Russian, and Farsi). Our extensive experiments suggest that contextual representations in PLMs do encode metaphoricity information, and mostly in their middle layers, and the knowledge is transferrable between languages and datasets in most cases. Finally, we show that PLMs face more challenges in generating metaphors, especially as their novelty increases. Our findings give helpful insights for both cognitive and NLP scientists.

1 Introduction

Pre-trained language models (PLMs) (Peters et al., 2018; Devlin et al., 2019), are now used in almost all NLP applications, e.g., machine translation (Li et al., 2021), question answering (Zhang et al., 2020), dialogue systems (Ni et al., 2021) and sentiment analysis (Minaee et al., 2020). They are the foundation models of NLP systems (Bommasani et al., 2021), causing huge impact in research and industry.

Metaphors are important aspects of human languages. In conceptual metaphor (CM) theory (Lakoff and Johnson, 2008), metaphor is mapping a concept in one domain (target) to a concept in another domain (source). Modeling metaphors is essential in building human-like computational systems that can relate new concepts to the old and familiar ones. The creativity and problem solving (i.e., generalization to new problems) also depend on the analogies and metaphors a cognitive system relies on.

We intuitively guess that PLMs must encode some information about metaphors due to their great performance in language processing tasks. Confirming that experimentally is a question that we try to focus on in this paper. So far, there has been no comprehensive analysis of how PLMs represent metaphorical information. The recent works in metaphor detection using PLMs, e.g., Choi et al. (2021), are related, but those are focused on achieving the best final performance.

We follow three paths in our study of metaphors in PLMs. First, we apply probing methods to understand the distribution of encoded metaphorical knowledge in different layers of PLMs. We employ edge probing (Tenney et al., 2019b) and minimum description length (Voita and Titov, 2020) techniques.

Second, to evaluate the metaphorical knowledge in their transferability and generalization, we design zero-shot cross-lingual and cross-dataset experiments. Four languages and four datasets are considered in this evaluation.

Finally, we explore how PLMs apply their metaphorical knowledge to generate or score a metaphorical expression. We take the novelty of metaphors and observe the differences in the fill-in-the-blank generation with that respect.

The LCC dataset (Mohler et al., 2016) is our main resource, containing annotations in four languages: English, Russian, Spanish and Farsi. We also experiment with three other English metaphor datasets, TroFi (Birke and Sarkar, 2006), VUA pos, and VUA verbs (Steen, 2010).

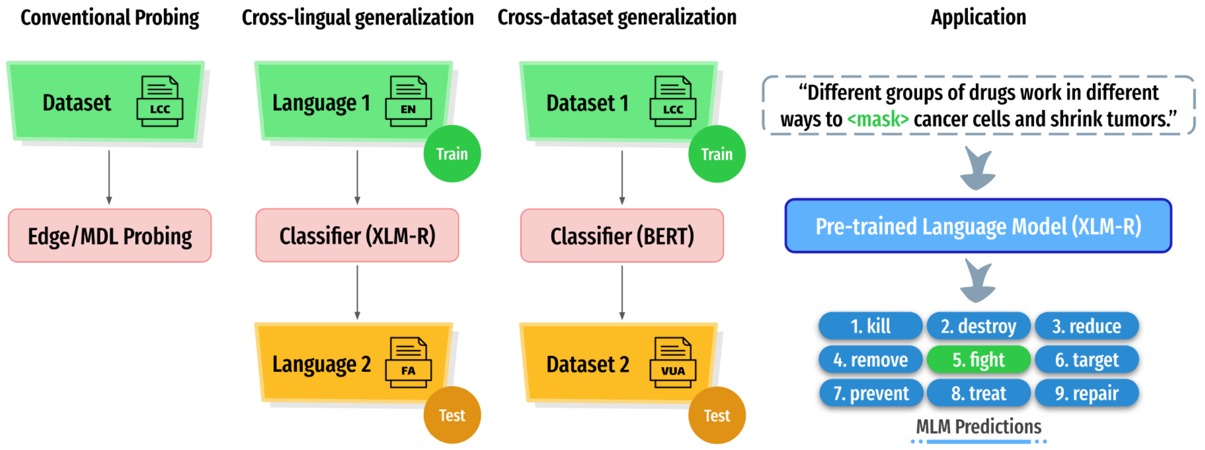

Figure 1: An illustration of our probing, generalization and application scenarios.

To summarize, we aim to answer questions such as (i) How do different layers of Transformers contribute to distinguishing literal from metaphorical usages of words? (ii) Do different PLMs encode metaphorical knowledge differently? (iii) Is the encoded information about metaphoricity transferrable across languages and datasets? (iv) How PLMs generation of a word in a context depends on its metaphoricity?

Our findings and contributions could be summarized as: (i) we show that metaphors are encoded differently across the layers of PLMs, but the trend is similar in three popular PLMs: BERT, RoBERTa, and ELECTRA. This is the first probing study of metaphors in PLMs. (ii) We run generalization analysis for PLMs in their out-of-distribution transferability of metaphorical knowledge. Our experiments confirm this generalization in most cases of cross-lingual and cross-dataset setups. (iii) We show that generating metaphors, especially more novel ones, is challenging for PLMs, confirming that PLMs behave similarly to humans in this respect.

2 Related Work

Metaphor detection using PLMs. The metaphoricity detection task (Mason, 2004; Birke and Sarkar, 2007; Shutova et al., 2013) is a good fit for most of our studies, even though we focus on analyzing the metaphorical knowledge and not on achieving the best detection results. Using PLMs for metaphoricity detection is normal in recent years, resulting in state-of-the-art results, indicating implicitly that PLMs do represent metaphoricity information. Choi et al. (2021) introduces a new architecture that integrates metaphor theories with BERT. Similarly, Song et al. (2021) presents a new perspective on metaphor detection task by framing it as relation classification, focusing on the verbs. These approaches beat other earlier works of using PLMs (Su et al., 2020; Chen et al., 2020; Gong et al., 2020), RNN-based (Wu et al., 2018; Mao et al., 2019) and feature-based approaches (Turney et al., 2011; Shutova et al., 2016).

Tsvetkov et al. (2014) present **cross-lingual metaphor detection** models using linguistic features and word embeddings. They make use of bilingual dictionaries to map between languages. The datasets they employ are quite small (1̃000 training and 2̃00 testing examples), making them unsuitable for a statistically robust evaluation.

Probing methods in NLP. Probing is an analytical tool used for assessing word representations linguistic knowledge. In probing, the information richness of the representations is determined by the quality of a supervised model in predicting linguistic properties solely based on the representations (Köhn, 2015; Gupta et al., 2015; Yaghoobzadeh and Schütze, 2016; Conneau et al., 2018; Tenney et al., 2019b,a; Hewitt and Manning, 2019).

A popular probing method introduced by Tenney et al. (2019b) is *edge probing* (Figure 2). They propose a suite of span-level tasks, including Part-of-speech tagging and coreference resolution. They demonstrate that BERT understands core NLP linguistic knowledge better than its uncontextualized counterparts and indicate the approximate position where each linguistic knowledge is encoded in BERT layers (Tenney et al., 2019a).

Despite the widespread use of edge probing and

VUA Verbs	He **[finds]**$_1$ it hard to communicate with people , not least his separated parents .→ 1
	He finds it hard to **[communicate]**$_1$ with people , not least his separated parents . → 0
VUA POS	They picked up power from a **[spider]**$_1$'s web of unsightly overhead wires . → 1
	They picked up power from a spider 's web of unsightly overhead **[wires]**$_1$. → 0
TroFi	" Locals **[absorbed]**$_1$ a lot of losses , " said Mr. Sandor of Drexel → nonliteral
	Vitamins could be passed right out of the body without being **[absorbed]**$_1$ → literal
LCC	Lawful **[gun ownership]**$_2$ is not a **[disease]**$_1$. → 3.0
	But the Supreme Court says it's not a way to **[hurt]**$_1$ the **[Second Amendment]**$_2$ → 2.0
	Is he angry that **[gun rights]**$_2$ **[progress]**$_1$ has been done without him? → 1.0
	I mean the **[2nd amendment]**$_2$ **[suggests]**$_1$ a level playing field for all of us. → 0.0

Table 1: Examples of sentences, spans, and target labels for each probing dataset.

other conventional probes, the question of whether the probing classifier is learning the task itself or it is actually identifying the linguistic knowledge in the representations raises concerns, see Belinkov (2021) for more discussions. An Information theoretic view can solve this issue (Voita and Titov, 2020) by reformulating probing as a data transmission problem. They consider the effort needed to extract linguistic knowledge as well as the final quality of the probe, showing that this approach is more informative and robust than normal probing methods. We employ both edge probing and MDL probing in this work.

Probing multi-lingual PLMs. The application of probing methods in NLP is extended to multilingual PLMs as well (Pires et al., 2019; Eichler et al., 2019; Ravishankar et al., 2019a; **?**; Choenni and Shutova, 2020). Choenni and Shutova (2020) introduce probing tasks for typological features of multiple languages in multi-lingual PLMs. Ravishankar et al. (2019a,b) extend the probing tasks of Conneau et al. (2018), to few other languages. Pires et al. (2019) study the generalization of multilingual-BERT across languages when performing cross-lingual downstream tasks. Here, as part of our study, we evaluate the generalization of multi-lingual PLMs in the representation of metaphoricity. We employ this setting as one of the main ways to understand the quality of encoded metaphoricity information in PLMs.

Out-of-distribution generalization. There has been no earlier work on studying or evaluating the out-of-distribution generalization in metaphor detection systems. Out-of-distribution generalization refers to scenarios where testing data comes from a different distribution from training data (Duchi and Namkoong, 2018; Hendrycks et al., 2020a,b). Here, we have scenarios where testing data is in different languages or in different domain / dataset. These are challenging evaluation scenarios for the generalization of encoded information (metaphoricity in our case).

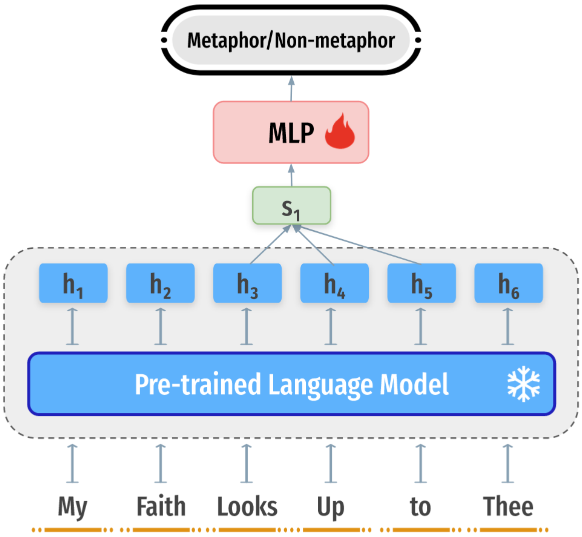

Figure 2: Probing architecture for metaphors employed in edge probing and MDL probing.

3 Metaphorical knowledge in PLMs

3.1 Probing

Based on conceptual metaphor theory, one domain (e.g., ARGUMENT) is explained using another domain (e.g., WAR). For example, in "We won the argument", ARGUMENT is linked to WAR, and the word "won" is used as its "non-literal" or "metaphorical" meaning. The same word "won" in a sentence like "China won the war" refers to its "literal" meaning. Here, our goal is to distinguish the metaphorical and literal words in given contexts. In other words, we want to detect if, given a context and a token span, the span refers to a metaphor or not.

As these examples show, metaphoricity is a highly contextualized phenomenon. The clues in the context might give us some hints. Here, our datasets take a sentence as the context. This might not be enough for many cases where larger contextual information is needed, but it is a reasonable simplification of the task to obtain large datasets.

Contextualized representations obtained by PLMs are shown to encode the contextual meanings of words relatively well (Devlin et al., 2019; Zhao et al., 2020). One word might have one or more metaphorical meanings and, likewise, one or more literal meanings. Our task is to distinguish between only two classes: metaphorical and literal.

We aim to answer a general question about metaphors in PLMs: do PLMs understand metaphors. We do not attempt to achieve the best metaphoricity detection results but to analyze and study how PLMs represent the necessary information to perform this task. In trying to answer this question, we apply probing methods, discussed as follows, to focus on the representation itself and not the fine-tuning task learning.

Methods We employ edge probing (Tenney et al., 2019b) and MDL (Voita and Titov, 2020). Edge probing consists of a classifier in which word representations obtained from PLMs are fed to it as inputs[1]. The quality of the classifier illustrates how well the representations encode a specific linguistic knowledge. This method is designed for span-level tasks, i.e., the classifier can only access the representations of a limited part of the input sentence specified in the dataset. The Edge Probing has two pooler sections for making fixed-sized vectors; one pools representations across the words in the span and the other pools representations across the layers.

The MDL probe is based on information theory and combines the quality of the classifier and the amount of effort needed to achieve this quality. Voita and Titov (2020) propose two methods for computing MDL: ”*variational coding*” and ”*online coding*.” The former computes the complexity of the classifier with a Bayesian model. In the latter, the classifier is trained gradually on different portions of the dataset and the code length will be the sum of the cross-entropies, each for a data portion. Voita and Titov (2020) show that the two methods' results are consistent with each other. Accordingly, we opted for the ”*online coding*” method since it is more straightforward in implementation. Since the code length is related to the size of the dataset, we report the compression, which is equal to 1 for a random classifier and more than 1 for better models. See extra details in (Voita and Titov, 2020)

3.2 Generalization

To see if PLM representations encode metaphoricity well, we evaluate them in settings where testing data comes from a different distribution from training. We explore transferability analysis across both languages and datasets. From the perspective of generality, these explorations also show how well PLMs generalize metaphorical information and how well a classifier can detect them across distributions. These important and helpful experiments are understudied in metaphor detection research. We adopt the term “domain” to refer to both “languages” and “datasets”.

3.2.1 Cross-lingual

Multi-lingual encoders project the representations in multiple languages into a shared space so that semantically similar words and sentences across languages end up close to each other. To answer questions about the transferability of metaphor information across languages, if we use a multi-lingual PLM model for representations, and our classifier shows that representations in language S are informative about metaphoricity, what happens if we apply this classifier to the representations in language T? If the representations are rich in both languages, information is accessible similarly across them, and metaphoricity can be transferred, then the classifier would be able to predict metaphoricity in language T from what it leans in S.

3.2.2 Cross-dataset

The great performance of PLMs is repeatedly related to their success in learning the existing heuristics in the training datasets, rather than the actual tasks (McCoy et al., 2019). By testing on a test set drawn from a different dataset compared to the training set, we can better measure the generalization of the knowledge encoded in PLMs. Therefore, another generalization dimension we consider is cross-dataset transfer, i.e., when training on dataset S and testing on dataset T. S and T are annotated in different groups with possibly different goals

[1] The representations are first projected to 256-dimensional vectors.

	Baseline		**BERT**		**RoBERTa**		**ELECTRA**	
Dataset	F1	Comp.	F1	Comp.	F1	Comp.	F1	Comp.
LCC (en)	75.78	1.08_8	88.24	1.93_6	88.03	2.02_6	**89.49**	**2.11_6**
TroFi	65.79	1.05_9	67.21	1.12_3	**67.87**	**1.16_3**	65.92	1.14_3
VUA POS	48.11	1.79_6	67.58	2.40_5	68.14	2.39_4	**68.79**	**2.46_5**
VUA Verbs	52.33	1.24_5	70.54	1.50_4	69.29	**1.57_6**	**72.14**	1.55_5

Table 2: Edge probing f1 results for various metaphoricity datasets in BERT, RoBERTa, and ELECTRA. The edge probing results are the average of three runs. The compression result is the best across layers, and the subscript denotes the best layer.

in mind and their raw sentences could come from different domains, both.

In our case, our datasets differ in their distribution of labeled examples of metaphors and literals (c.f., Table 3, as well as the distribution of the candidate spans (e.g., TroFi is only verbs, but LCC is not). Further, the annotation process is different as each follows their own guidelines. However, the basic task of metaphoricity detection, i.e., distinguishing metaphor and literal usages, is the same for all.

3.3 Application

Amanzio et al. (2008) find that novel metaphors introduce challenges in language comprehension for people with Alzheimer's. Here, we study if PLMs also struggle with more novel metaphors. We do so by inspecting their generation of metaphorical words given context. We relate this generation to the usage of metaphorical knowledge. We leverage the LCC dataset metaphoricity score associated with each example which indicates various levels of metaphoricity, including none, conventional, or clear metaphors for comparing the models' generative ability.

By using the fill-in-the-blank methodology, after acquiring the ranked list of the predicted tokens from MLM probabilities, we calculate the $recall@k$ for multiple values of k. The source span of the LCC dataset is masked for the model to predict, and only the entries which span a single token are considered in this experiment. An example is shown in Figure 1, where the real output is "fight" and MLM predicted that as its fifth ranked token.

Dataset	$\%M$	Size
LCC (en)	45.2 / 46.8	39,769 / 4,465
LCC (fa)	43.7 / 44.8	18,496 / 2,070
LCC (es)	34.6 / 34.4	29,002 / 3,339
LCC (ru)	45.1 / 34.4	17,492 / 1,932
TroFi	42.6 / 42.3	5,770 / 666
VUA Verbs	27.8 / 30.0	17,240 / 5,873
VUA POS	15.2 / 17.9	72,611 / 2,2196

Table 3: Statistics of the datasets. Percentage of metaphors (%M) and number of instances for train / test sets are given for each dataset.

4 Experimental setup and results

4.1 Datasets and setup

Datasets We use three metaphoricity detection datasets in our study. One of these datasets, i.e., LCC, contains annotations in four languages: English, Russian, Spanish, and Farsi. The other two, TroFi and VUA, are in English only. The statistics of the datasets are shown in Table 3. Some example sentences with the metaphoricity annotation can be seen in Table 1.

The annotations of LCC (Mohler et al., 2016) are done according to Conceptual Metaphor theory, with source and target domains, in four languages. TroFi (Birke and Sarkar, 2006) TroFi dataset consists of metaphoric and literal usages of 51 English verbs from the Wall Street Journal. VU Amsterdam (Steen, 2010) corpus consists of words in the academic, fiction, and news subdomains of the British National Corpus (BNC), annotated with five annotators as figurative (specifying metaphor/personification/other) or literal.

Setup In implementing the edge probe, following Tenney et al. (2019b), we use 32 as the batch size, learning rate of 5e-5, and 256 projection dimensions.

For the MDL probe, the same structure of edge probing is employed. We apply a logarithm to the base in cross-entropy loss to have all the obtained code lengths in bits. See extra details in Voita and Titov (2020)

4.2 Probing

We use BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and ELECTRA (Clark et al., 2020) to represent our PLMs. Due to resource limitations, we conduct our experiments formed on the *base* version of the models (12 layers, 768 hidden size, 110M parameters) implemented in HuggingFace's Transfomers library (Wolf et al., 2020). We employ edge probing for evaluating overall metaphorical knowledge in our selected PLMs. However, as discussed by Belinkov (2021) and Fayyaz et al. (2021), edge probing is not reliable for layer-wise cross-model experiments, and so we leverage the MDL probe to this end.

Table 2 shows the edge probing F1 and MDL probing compression results for our three PLMs. Accordingly, RoBERTa and ELECTRA are shown to encode metaphorical knowledge better than BERT on both metrics. This can be attributed to the two models' better performance on various tasks, acquired by having better pre-training objectives and enjoying more extensive pre-training data.

MDL probing compression across layers is demonstrated in Figure 3. Except for RoBERTa results in TroFi and VUA Verbs, we see the numbers increase at the first 3 to 6 layers, depending on the dataset, but it decreases afterwards. In other words, metaphorical information is more concentrated in the middle layers. The representations in the middle layers are relatively contextualized but not as much as higher layers. This indicates that detecting metaphoricity of a span is a contextualized task but it usually needs only a few layers of contextualization.

4.3 Generalization

As our PLMs, we use XLM-RoBERTa (Conneau et al., 2020) for cross-lingual and BERT for cross-dataset experiments. We apply the edge probing architecture as in the probing experiments. As we mentioned in Section 3.2, we sometimes refer to both language and dataset as domain for simplicity.

For each case of a source domain S and a target domain T, we run two experiments: one with the PLM and one with a randomized version of the PLM where weights are set to random values. Randomly initialized PLM that did not experience the pre-training process is a commonly used baseline in the community. So, the difference between the two gives evidence about the helpfulness of the encoded knowledge in PLMs for metaphoricity detection. When $S = T$, this effect is measured for in-domain generalization and when $S \neq T$, for out-of-domain generalization. Comparing results of in-domain (e.g., training and testing on English data) and out-of-domain (e.g., training on Spanish and testing on English) setups demonstrates how generalizable the metaphoricity knowledge is across domains.

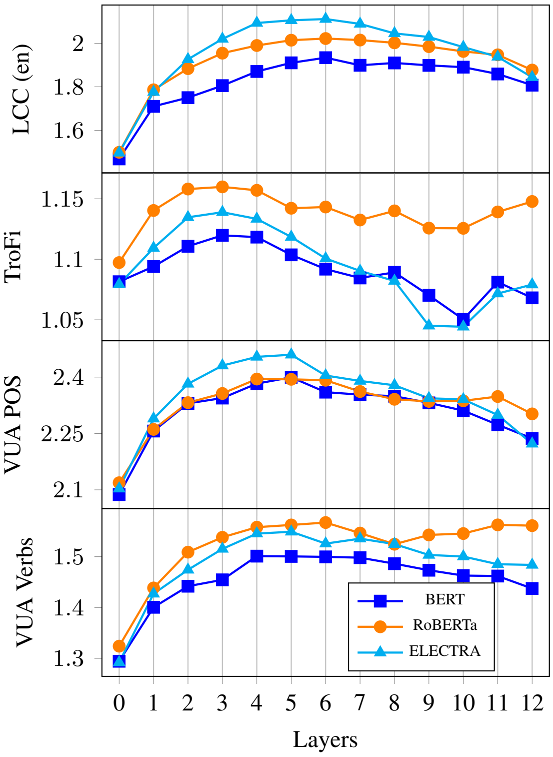

Figure 3: Metaphoricity detection MDL probing compression across layers.

4.3.1 Cross-lingual

The LCC datasets in four languages are used here. We train on one source (e.g., English) and test on a target (e.g., Spanish) language. We subsample from the datasets to have the same number of examples in each dataset (17,492 which is the size of Russian dataset).

The results of our cross-lingual experiments are shown in Table 4. The random baseline is acquired using a randomly initialized XLM-RoBERTa. We observe that in all cases, the pre-trained PLM

		Train Language			
		en	es	fa	ru
Test Lang.	en	**85.4** (72.3)	76.6 (41.5)	78.1 (63.7)	80.9 (63.7)
	es	75.8 (46.4)	**82.4** (64.8)	75.5 (51.2)	75.8 (47.3)
	fa	71.4 (61.9)	65.2 (44.7)	**81.4** (72.0)	76.8 (55.3)
	ru	81.9 (61.2)	77.19 (51.3)	79.9 (59.2)	**88.9** (71.7)

Table 4: Cross-lingual metaphoricity detection F1 results. The multi-lingual PLM, i.e., XLM-RoBERTa, is better than random by large margins in most cases. For each test language, we bold its in-domain (e.g., en → en), and underline the best out-of-domain (e.g., ru → en) numbers.

		Train Dataset			
		LCC(en)	TroFi	VUA POS	VUA Verbs
Test Dataset	LCC(en)	**83.7** (62.7)	63.2 (63.4)	64.2 (63.1)	61.2 (42.3)
	TroFi	58.6 (39.1)	**66.2** (63.6)	58.3 (24.5)	58.3 (59.3)
	VUA POS	39.9 (26.3)	31.6 (29.0)	**53.18** (28.8)	49.9 (29.6)
	VUA Verbs	38.5 (34.5)	47.0 (46.0)	59.8 (29.8)	**66.4** (48.0)

Table 5: Cross dataset edge probing F1 results on BERT shown in pairs: pre-trained model and, in the parenthesis, the randomly initialized model. We set the training size to the minimum among datasets, i.e., TroFi. For each test dataset, we bold its in-domain (e.g., TroFi → TroFi), and underline the best out-of-domain (e.g., LCC(en) → TroFi) numbers.

outperforms the random. This shows that some metaphorical knowledge, learned during the pre-training phase, is transferable across languages.

Further, some language pairs (English & Russian and Russian & Farsi) seem to have higher transferrability than others. In some cases, interestingly, one way is much better than the other (e.g., Farsi → English is 78.1 but English → Farsi is 71.4).

Finally, we can also compare the in-domain random PLM with the out-of-domain pre-trained PLM results. For English as test, the best out-of-domain result, i.e., 80.9 for Russian → English, is better than the in-domain random result of 72.3.

4.3.2 Cross-dataset

Similar to the cross-lingual evaluations, here we have four datasets as sources and targets. We set the size of all datasets to the minimum, i.e., 5770. For each pair, we run two experiments: one with random and one with XLM-RoBERTa as our PLM. In total, here we run $4 \cdot 4 \cdot 2$ (random and pre-trained PLM) $= 32$ experiments. Results are shown in Table 5.

As expected, VUA Verbs and VUA POS achieve the best results when mutually tested since they are from fairly the same distribution. Similar distributions seem to be impactful when testing on the VUA Verbs dataset as well, where we can see that training on the TroFi dataset outperforms LCC. This can be attributed to the fact that TroFi only has verbs as metaphorical spans, just like VUA Verbs, whereas LCC has both verbs and nouns. This observation can be validated where we test on the VUA POS dataset. In this setting, the opposite happens, and the LCC dataset outperforms TroFi, which can be related to having both nouns and verbs like VUA POS rather than just verbs.

4.4 Application

LCC dataset provides metaphoricity scores including 0 as no metaphoricity, 2 likely/conventional metaphor, and 3 clear metaphor.[2] We leverage these scores to study PLM's understanding of the commonality of the metaphors and their resemblance to human annotators as the novelty changes.

Table 6 shows the average predicted probability for the masked spans in different metaphoricity classes across four languages. Our results demonstrate that, as the metaphoricity score increases, the model's prediction probability for the desired word decreases consistently over languages.

[2] 1 is possible/weak metaphor and as Mohler et al. (2016) describes metaphors with $0.5 \leq score < 1.5$ as unclear we ignore this score.

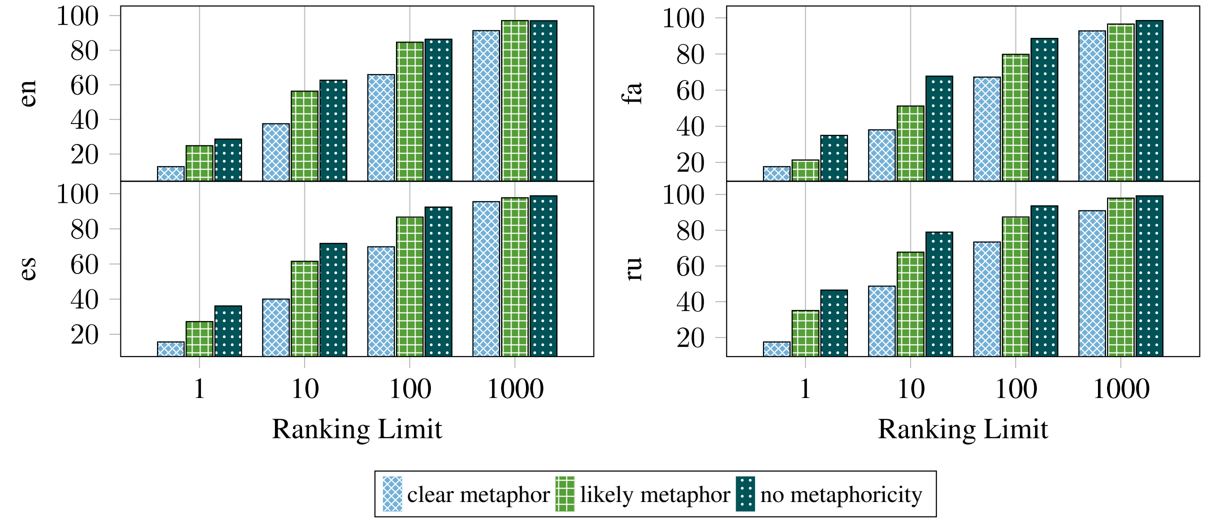

Figure 4: XLM-RoBERTa MLM prediction recall in LCC dataset over four languages. Recall is the proportion of the entries predicted within the ranking limit to all the entries. Scores are from the LCC dataset including no metaphoricity, likely/conventional metaphor, and clear metaphor.

	Metaphoricity Score		
Language	0	2	3
English	.20	.17	.08
Spanish	.27	.19	.12
Farsi	.28	.15	.13
Russian	.37	.27	.15

Table 6: XLM-RoBERTa MLM average probability of the source concept in the LCC datasets over four languages. Metaphoricity scores are from the LCC dataset, and 0: no metaphoricity, 1: possible/weak metaphor, 2: likely/conventional metaphor, and 3: clear metaphor. The probabilities are acquired after applying softmax.

We also validate these findings by calculating $recall@k$ for different values of k. Figure 4 demonstrates the MLM predictions recall for different metaphoricity levels. We consider different ranking limits k to show what fraction of the examples were correctly predicted by the language model in the top k predictions. The results imply that metaphorical sentences in general, and clear metaphors in particular, result in less recall than normal use of language. This means that the model can predict the masked token in typical sentences with higher probabilities. In contrast, a figurative use of language causes difficulties for the model to predict the desired word.

We show that the trend and conclusion are consistent in Farsi, Spanish, and Russian as well.

5 Conclusion

In this paper, we shed light on how metaphorical knowledge is encoded in PLMs, through probing, cross-dataset and cross-language analysis and generation. We ran novel scenarios on metaphor detection and generation and presented findings helpful for both NLP and cognitive sciences.

We showed that metaphorical knowledge in PLMs is somewhat generalizable across languages. This could be an interesting direction to follow for further investigations within both linguistics and NLP. Our evaluation also demonstrated that PLMs generate metaphors and even novel ones but with more hardship than literal expressions. This is an indication that they might face issues understanding them as well, and more work is needed to equip them with better metaphorical information. Metaphors are important in human cognition, and if we seek to build cognitively inspired or plausible language understand systems, we need to work more in their integration in the future.

References

Martina Amanzio, Giuliano Geminiani, Daniela Leotta, and Stefano Cappa. 2008. Metaphor comprehension in alzheimer's disease: Novelty matters. *Brain and language*, 107(1):1–10.

Yonatan Belinkov. 2021. Probing classifiers: Promises, shortcomings, and alternatives. *CoRR*, abs/2102.12452.

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*, Trento, Italy. Association for Computational Linguistics.

Julia Birke and Anoop Sarkar. 2007. Active learning for the identification of nonliteral language. In *Proceedings of the Workshop on Computational Approaches to Figurative Language*, pages 21–28, Rochester, New York. Association for Computational Linguistics.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie S. Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, and et al. 2021. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258.

Xianyang Chen, Chee Wee (Ben) Leong, Michael Flor, and Beata Beigman Klebanov. 2020. Go figure! multi-task transformer-based architecture for metaphor detection using idioms: ETS team in 2020 metaphor shared task. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 235–243, Online. Association for Computational Linguistics.

Rochelle Choenni and Ekaterina Shutova. 2020. What does it mean to be language-agnostic? probing multilingual sentence encoders for typological properties. *CoRR*, abs/2009.12862.

Minjin Choi, Sunkyung Lee, Eunseong Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. 2021. Melbert: Metaphor detection via contextualized late interaction using metaphorical identification theories. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 1763–1773. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136. Association for Computational Linguistics.

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.

John C. Duchi and Hongseok Namkoong. 2018. Learning models with uniform performance via distributionally robust optimization. *CoRR*, abs/1810.08750.

Max Eichler, Gözde Gül Şahin, and Iryna Gurevych. 2019. LINSPECTOR WEB: A multilingual probing suite for word representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 127–132, Hong Kong, China. Association for Computational Linguistics.

Mohsen Fayyaz, Ehsan Aghazadeh, Ali Modarressi, Hosein Mohebbi, and Mohammad Taher Pilehvar. 2021. Not all models localize linguistic knowledge in the same place: A layer-wise probing on bertoids' representations.

Hongyu Gong, Kshitij Gupta, Akriti Jain, and Suma Bhat. 2020. IlliniMet: Illinois system for metaphor detection with contextual and linguistic information.

In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 146–153, Online. Association for Computational Linguistics.

Abhijeet Gupta, Gemma Boleda, Marco Baroni, and Sebastian Padó. 2015. Distributional vectors encode referential attributes. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 12–21, Lisbon, Portugal. Association for Computational Linguistics.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. 2020a. The many faces of robustness: A critical analysis of out-of-distribution generalization. *arXiv preprint arXiv:2006.16241*.

Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. 2020b. Pretrained transformers improve out-of-distribution robustness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2744–2751, Online. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. 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 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Arne Köhn. 2015. What's in an embedding? analyzing word embeddings through multilingual evaluation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2067–2073, Lisbon, Portugal. Association for Computational Linguistics.

George Lakoff and Mark Johnson. 2008. *Metaphors we live by*. University of Chicago press.

Junyi Li, Tianyi Tang, Wayne Xin Zhao, and Ji-Rong Wen. 2021. Pretrained language models for text generation: A survey. *CoRR*, abs/2105.10311.

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.

Rui Mao, Chenghua Lin, and Frank Guerin. 2019. End-to-end sequential metaphor identification inspired by linguistic theories. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3888–3898, Florence, Italy. Association for Computational Linguistics.

Zachary J. Mason. 2004. CorMet: A computational, corpus-based conventional metaphor extraction system. *Computational Linguistics*, 30(1):23–44.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2020. Deep learning based text classification: A comprehensive review. *CoRR*, abs/2004.03705.

Michael Mohler, Mary Brunson, Bryan Rink, and Marc T. Tomlinson. 2016. Introducing the LCC metaphor datasets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*. European Language Resources Association (ELRA).

Jinjie Ni, Tom Young, Vlad Pandelea, Fuzhao Xue, Vinay Adiga, and Erik Cambria. 2021. Recent advances in deep learning based dialogue systems: A systematic survey. *CoRR*, abs/2105.04387.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Vinit Ravishankar, Memduh Gökırmak, Lilja Øvrelid, and Erik Velldal. 2019a. Multilingual probing of deep pre-trained contextual encoders. In *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*, pages 37–47, Turku, Finland. Linköping University Electronic Press.

Vinit Ravishankar, Lilja Øvrelid, and Erik Velldal. 2019b. Probing multilingual sentence representations with X-probe. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 156–168, Florence, Italy. 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.

Ekaterina Shutova, Simone Teufel, and Anna Korhonen. 2013. Statistical metaphor processing. *Computational Linguistics*, 39(2):301–353.

Wei Song, Shuhui Zhou, Ruiji Fu, Ting Liu, and Lizhen Liu. 2021. Verb metaphor detection via contextual relation learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 4240–4251. Association for Computational Linguistics.

Gerard Steen. 2010. *A method for linguistic metaphor identification: From MIP to MIPVU*, volume 14. John Benjamins Publishing.

Chuandong Su, Fumiyo Fukumoto, Xiaoxi Huang, Jiyi Li, Rongbo Wang, and Zhiqun Chen. 2020. DeepMet: A reading comprehension paradigm for token-level metaphor detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 30–39, Online. Association for Computational Linguistics.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019a. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019b. What do you learn from context? Probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pages 248–258. The Association for Computer 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.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi 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 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.

Chuhan Wu, Fangzhao Wu, Yubo Chen, Sixing Wu, Zhigang Yuan, and Yongfeng Huang. 2018. Neural metaphor detecting with CNN-LSTM model. In *Proceedings of the Workshop on Figurative Language Processing*, pages 110–114, New Orleans, Louisiana. Association for Computational Linguistics.

Yadollah Yaghoobzadeh and Hinrich Schütze. 2016. Intrinsic subspace evaluation of word embedding representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 236–246, Berlin, Germany. Association for Computational Linguistics.

Zhuosheng Zhang, Hai Zhao, and Rui Wang. 2020. Machine reading comprehension: The role of contextualized language models and beyond. *CoRR*, abs/2005.06249.

Mengjie Zhao, Philipp Dufter, Yadollah Yaghoobzadeh, and Hinrich Schütze. 2020. Quantifying the contextualization of word representations with semantic class probing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1219–1234, Online. Association for Computational Linguistics.

A Appendices

Language	Sentence	Annotations
fa	اما امریکا در افغانستان، از همان آغاز، با [سلاح]$_1$ [دموکراسی]$_2$ آمده است .	Score: 3.0 Src Concept: WAR(3.0) Target Concept: DEMOCRACY Polarity: NEUTRAL Intensity: 1.0
es	[atorado]$_1$ en la [deuda]$_2$ pública y sin avances en Estado de Derecho	Score: 3.0 Src Concept: BARRIER(3.0) Target Concept: DEBT Polarity: NEGATIVE Intensity: 2.0
ru	Мировые [деньги]$_2$ [мечутся]$_1$, не зная , куда вложиться .	Score: 3.0 Src Concept: MOVEMENT(3.0) Target Concept: MONEY Polarity: NEGATIVE Intensity: 2.0

Table 7: Examples of sentences, spans, and annotations for LCC dataset in Farsi, Spanish, and Russian.

	English				**Spanish**			
R\Score	0	1	2	3	0	1	2	3
10	1.77	1.93	1.95	2.32	1.59	1.77	1.87	2.21
100	11.01	10.13	12.43	18.07	8.57	9.49	11.43	18.70
1000	44.62	37.08	51.63	113.01	26.13	24.63	42.35	92.81
5000	96.43	78.53	97.54	264.92	47.09	39.34	80.70	168.75
	Farsi				**Russian**			
R\Score	0	1	2	3	0	1	2	3
10	1.57	1.89	2.09	2.15	1.20	1.50	1.55	1.95
100	9.56	11.91	14.26	18.96	5.95	7.66	9.50	14.46
1000	40.86	51.98	70.23	115.45	23.28	24.90	40.71	88.97
5000	65.31	79.55	123.31	227.19	40.41	33.77	77.72	297.92

Table 8: XLM-RoBERTa MLM ranking average of the source concept in LCC dataset over four languages. Scores are from the LCC dataset, 0 indicating no metaphoricity, 1 possible/weak metaphor, 2 likely/conventional metaphor, and 3 clear metaphor.