Probing and Generalization 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. In this paper, we investigate this hypothesis for PLMs, by probing metaphoricity information in their encodings, and by measuring the cross-lingual and cross-dataset generalization of this information. We present studies in multiple metaphor detection datasets and in four languages (i.e., English, Spanish, Russian, and Farsi). Our extensive experiments suggest that contextual representations in PLMs do encode metaphorical knowledge, and mostly in their middle layers. The knowledge is transferable between languages and datasets, especially when the annotation is consistent across training and testing sets. 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 have sometimes been referred to as “foundation models” (Bommasani et al., 2021) due to their significant impact on research and industry.

Metaphors are important aspects of human languages. In conceptual metaphor theory (CMT) (Lakoff and Johnson, 2008), metaphor is defined as a cognitive phenomenon associating two different concepts or domains. This phenomenon is built in cognition and expressed in language. The creativity and problem solving (i.e., generalization to new problems) depend on the analogies and metaphors a cognitive system, like our brain, relies on. Modeling metaphors is therefore essential in building human-like computational systems that can relate emerging concepts to the more familiar ones.

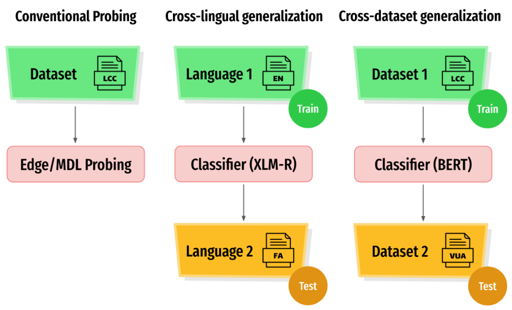

Figure 1: An illustration of our probing and generalization scenarios for metaphorical knowledge.

So far, there has been no comprehensive analysis of whether and how PLMs represent metaphorical information. We intuitively assume that PLMs must encode some information about metaphors due to their great performance in metaphor detection and other language processing tasks. Confirming that experimentally is a question that we address here. Specifically, we aim to know *whether generalizable metaphorical knowledge is encoded in PLM representations or not*. The outline of our work is presented in Figure 1.

We first do *probing* experiments to answer questions such as: (i) with which accuracies and extractablities do different PLMs encode metaphorical knowledge? (ii) how deep is the metaphorical knowledge encoded in PLM multi-layer representations? We take two probing methods, edge probing (Tenney et al., 2019b) and minimum description length (Voita and Titov, 2020), and apply them to four metaphor detection datasets, namely LCC (Mohler et al., 2016), TroFi (Birke and Sarkar, 2006), VUA pos, and VUA Verbs (Steen, 2010).

Since annotation biases and artifacts are common in human-labeled NLP datasets, we might overestimate the *generalization* of metaphorical knowledge in PLMs while performing our probing analysis. To overcome that, we design two general setups in which testing comes from a different distribution than training: cross-lingual and cross-dataset metaphor detection. Each setup can reveal important information on whether or not the metaphorical knowledge is encoded consistently in PLMs. Four languages (English, Farsi, Russian and Spanish) and four datasets (LCC, TroFi, VUA pos, and VUA Verbs) are considered in these generalization experiments.

Overall, our contributions could be summarized as follows: (i) For the first time, and through careful probing analysis, we confirm that PLMs do encode metaphorical knowledge. (i) We show that metaphorical knowledge is encoded better in the middle layers of PLMs. (ii) We evaluate the generalization of metaphorical knowledge in PLMs across four languages and four dataset sources, and find out that there is considerable transferability for the pairs with consistent data annotation even if they are in different languages.

2 Related Work

Metaphor detection using PLMs. The metaphor detection task (Mason, 2004; Birke and Sarkar, 2007; Shutova et al., 2013) is a good fit for analyzing the metaphorical knowledge. Using PLMs for metaphor detection has been common in recent years, setting new state-of-the-art results, indicating implicitly that PLMs represent metaphorical information. Choi et al. (2021) introduce a new architecture that integrates metaphor detection theories with BERT. They use the definitions as well as example usages of words jointly with PLM representations. 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 earlier work 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). Note that our goal is not to compete with these models, but to probe and analyze the relevant knowledge in PLMs.

Tsvetkov et al. (2014) present **cross-lingual metaphor detection** models using linguistic features and word embeddings. Bilingual dictionaries map different languages. Their datasets are quite small ($\tilde{1}000$ training and $\tilde{2}00$ testing examples), making them unsuitable for a robust evaluation. However, this paper still remains as the only cross-lingual evaluation of metaphor detection, to the best of our knowledge. Here, using multilingual PLMs, we perform zero-shot cross-lingual transfer for metaphor detection. Our goal is to test if PLMs represent metaphorical knowledge transferable across languages.

Probing methods in NLP. Probing is an analytical tool used for assessing linguistic knowledge in language representations. In probing, the information richness of the representations is inspected by the quality of a supervised model in predicting linguistic properties based only 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; Belinkov, 2021). Here, we apply probing to perform our study on whether metaphorical knowledge is present in PLM representations, and whether that is generalizable across languages and datasets.

A popular probing method introduced by Tenney et al. (2019b) is *edge probing* (Figure 2). They propose a suite of span-level tasks, including POS tagging and coreference resolution. Despite the widespread use of edge probing and other conventional probes, the question of whether the probing classifier is learning the task itself rather than identifying the linguistic knowledge raises concerns. 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 in addition to the final quality of the probe, showing that this approach is more informative and robust than normal probing methods. We employ both edge 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,b; Choenni and Shutova, 2020). Choenni and Shutova (2020) introduce probing tasks for typological features of multiple languages in multilingual 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 probe the generalization of metaphorical knowledge in XLM-R (Conneau et al., 2020), a notable multilingual PLM.

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

3 Inspecting Metaphorical Knowledge in PLMs

Metaphors are used frequently in our everyday language to convey our thoughts more clearly. There are related theories in linguistics and cognitive science. Following linguistic theories, metaphoricity is mostly annotated using metaphor identification procedure (MIP). MIP identifies a word in a given context as a metaphor if it has a basic or literal meaning that contrasts with its context words. Based on conceptual metaphor theory (CMT) (Lakoff and Johnson, 2008), one target domain (e.g., ARGUMENT) is explained using a source domain (e.g., WAR). The source domain is usually more concrete or physical, while the target is more abstract. Relating these two theories, metaphors are expressed in language connecting two contrasting domains. For example, in “We won the argument”, the domain of ARGUMENT is linked to the domain of WAR by using the word “won”. The word “won” is a “metaphor” here since its primary domain contrasts with its contextual domain. The same word “won” in a sentence like “The Allies won the war” refers to its literal meaning and therefore is not a metaphor. The task of metaphor detection is defined to do this classification of “literal” and “metaphor”.

Accordingly, when designing a metaphor detection system, to figure out if a token is a metaphor in a particular context, we assume following a process like: (i) finding if the token has multiple meanings in different domains, including a more basic, concrete, or body-related meaning. For example, “fight”, “win” and “mother” satisfy this condition. (ii) finding if the source domain of the token contrasts with the target domain. Here the contrast is important and finding the exact domains might not be necessary. The source domain, in which its literal / basic meaning resides, is a non-contextual attribute, while the target domain is mainly found using the contextual clues (WAR and ARGUMENT for “won” in the above example).

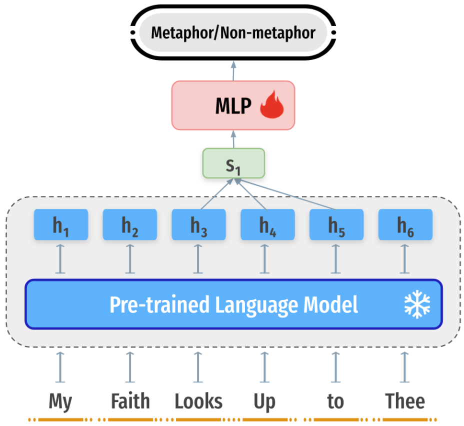

Figure 2: Probing architecture for metaphors employed in edge probing and MDL probing.

Here, we use the metaphor detection datasets annotated based on these theories and analyze the PLM representations to see if they encode metaphorical knowledge and if the encoding is generalizable. To do so, we first probe PLMs for their metaphorical information, generally and also across layers. This gives us intuition on how well metaphoricity is encoded and how local or contextual that is. Then, we test if the knowledge of metaphor detection can be transferred across languages and if multilingual PLMs capture that. Finally, the generalization of metaphorical knowledge across datasets is examined to see if the theories and annotations followed by different datasets are consistent, and if PLMs learn generalizable knowledge rather than dataset artifacts.

3.1 Probing

Here, we aim to answer general questions about metaphors in PLMs: do PLMs encode metaphorical information and, if so, how it is distributed in their layers. We do not attempt to achieve the best metaphor detection results but to analyze layers of PLMs to test if they contain 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 by freezing the PLM parameters and training classifiers on top.

We hypothesize that metaphorical information does exist in PLM layers and more in the middle layers since: (i) as we discussed earlier, metaphor detection depends on contrast prediction between source and target domains of a token. We assume that this prediction is done mainly based on the initial layers of PLM representations of either the token itself or its context or both. A few layers of processing is still helpful to capture different clues well. (ii) in higher layers of PLMs, the representations are dominated by the contextual information, making it hard to retrieve the source domain, and so reasoning about the contrast of the source and target domains becomes difficult.

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 as inputs after projecting to 256-dimensional vectors first. 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. 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 Minimum Description Length (MDL) probing is based on information theory and combines the quality of the classifier and the amount of effort needed to achieve this quality. The goal is to send the labels of given inputs with the minimum description length. 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 N, we report the "*compression*", which is equal to 1 for a random classifier and larger for better models, and is defined as: $compression = \frac{N \cdot \log_2(K)}{\text{MDL}}$ See extra details in Voita and Titov (2020).

3.2 Generalization

The excellent performance of PLMs is repeatedly related to their success in learning heuristics in the data rather than the actual tasks. When training and testing on the same distribution, any learning model often uses heuristics and annotation biases. The consequence is the recurring overestimation of the capabilities of PLMs in doing hard tasks.

To see if PLMs encode generalizable metaphorical knowledge, we evaluate them in settings where testing and training data are in different distributions. We explore transferability analysis across languages and datasets as two sources of distribution. We explain each in the following sections.

3.2.1 Cross-lingual

Multilingual 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. If we use a multilingual PLM model, 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? We hypothesize that if the representation is rich in both languages, the annotation of metaphor is consistent, and the concept of metaphor is transferable across languages, then the classifier would be able to predict metaphoricity in language T from what it learns in S.

When testing cross-lingual generalization, the linguistic and cultural differences of metaphoricity is important as well. We assume that metaphors are conceptualized in a similar process across languages, and metaphor detection is defined consistently. The lexicalization is, of course, different, but that is something that multi-lingual PLMs are supposed to handle to some extent.

3.2.2 Cross-dataset

Another generalization dimension we consider is cross-dataset transfer, i.e., training on dataset S and testing on dataset T. S and T could be annotated by different people with possibly different goals in mind, and their raw sentences could come from different domains. However, they must be annotated for the same task of metaphor detection.

In our case, the four datasets discussed more in §4.1 differ in their distribution of the candidate POS types (e.g., TroFi is only verbs, but LCC is not). Further, the annotation process is different as each follows its own guidelines. However, the

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

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

dataset	POS	Sizes
LCC (en)	ALL	28,096 / 4,014 / 8,028
LCC (fa)	ALL	12,238 / 1,802 / 3,604
LCC (es)	ALL	12,238 / 2,236 / 4,474
LCC (ru)	ALL	12,238 / 1,748 / 3,498
TroFi	V	3,838 / 548 / 1,096
VUA Verbs	V	9,176 / 1,310 / 2,622
VUA POS	ALL	21,036 / 3,006 / 6,010

Table 2: Statistics of the datasets. We label-balance each to have 50% metaphor. Number of instances for train / dev / test sets and the types of POS are given as well. N: Noun, V: Verb, ALL: Noun, Verb, Adjective, Adverb.

essential task of metaphor detection, i.e., distinguishing metaphor and literal usages, is the same for all. Therefore, we expect some transferability across datasets but with differences aligned with their mismatches.

4 Experimental Setup and Results

4.1 Datasets and Setup

Datasets We use four metaphor detection datasets in our study. The annotations of LCC (Mohler et al., 2016) are done mostly on Web crawled data as well as news corpora. It provides metaphoricity scores including 0 as no , 2 as conventional, and 3 as clear metaphor.[1] We use the examples with score 0 as literal, and others as metaphor. TroFi dataset (Birke and Sarkar, 2006) consists of metaphoric and literal usages of 51 English verbs from WSJ. VUA (Steen, 2010) corpus consists of words in the academic, fiction, and news subdomains of the British National Corpus (BNC). The authors published two versions: VUA POS and VUA Verbs.

LCC contains annotations in four languages: English, Russian, Spanish, and Farsi. The other three datasets, TroFi, VUA Verbs and VUA POS, are in English only. We have label-balanced all the datasets to get a more straightforward interpretation of results (the accuracy of a fair-coin random baseline is 50% in all cases) and have splitted the datasets to train / dev / test sets with ratios of 0.7 / 0.1 / 0.2. [2] The statistics of the datasets are shown in Table 2. Example sentences with the corresponding annotations can be seen in Table 1.

Setup In implementing the edge probe, following Tenney et al. (2019b), we use batch size = 32 and learning rate = 5e-5 and train for five epochs in all experiments. For the MDL probe, the same structure of edge probing is employed. We apply a logarithm to the base two instead of the natural logarithm in cross-entropy loss to have all the obtained code lengths in bits (see extra details in Voita and Titov 2020).

4.2 Probing Results

We use BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and ELECTRA (Clark et al., 2020) to represent our PLMs. Due to our resource limitations, we conduct all experiments on the *base* version of the models (12 layers, 768 hidden size, 110M parameters) implemented in HuggingFace's Transfomers (Wolf et al., 2020). We employ edge probing for evaluating overall metaphorical knowledge in our selected PLMs, and MDL for the layer-wise comparisons. MDL is shown to be more effective for layer-wise probing (Fayyaz et al., 2021).

Table 3 shows the edge probing accuracy 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 is consistent with their better performance on various tasks, acquired

[1] 1 is weak metaphor and as Mohler et al. (2016) describe metaphors with $0.5 \leq score < 1.5$ as unclear, we ignore it.

[2] The datasets and the implementation are available at https://anonymous.4open.science/r/paper_supplementary_material_codes_data-6531

Dataset	Baseline Acc.	Baseline Comp.	BERT Acc.	BERT Comp.	RoBERTa Acc.	RoBERTa Comp.	ELECTRA Acc.	ELECTRA Comp.
LCC (en)	74.86	1.05_2	88.25	1.85_6	88.06	1.96_5	**89.30**	$\mathbf{2.05_5}$
TroFi	67.34	1.01_4	**68.58**	1.07_4	68.46	$\mathbf{1.09_6}$	68.07	1.08_3
VUA POS	65.92	1.03_0	80.32	1.43_5	81.72	1.48_6	**83.03**	$\mathbf{1.51_4}$
VUA Verbs	65.97	1.04_9	78.29	1.28_9	78.88	$\mathbf{1.34_5}$	**79.96**	1.31_4

Table 3: Edge probing accuracy results for various metaphoricity datasets in BERT, RoBERTa, and ELECTRA. Baseline is a randomly initialized BERT. 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.

by having better pre-training objectives and / or enjoying more extensive pre-training data. The higher probing quality of ELECTRA's representations, is also consistent with Fayyaz et al. (2021) results on various linguistic knowledge tasks, including dependency labeling, named entity recognition, semantic role labeling, and coreference resolution.

MDL probing compression across layers is demonstrated in Figure 3. We see the numbers increase mostly at the first 3 to 6 layers, depending on the dataset, but it decreases afterwards[3]. In other words, metaphorical information is more concentrated in the middle layers, where the representations are relatively contextualized but not as much as higher layers. To put this in perspective, we can consider Tenney et al. (2019a) and Fayyaz et al. (2021) where the best layers for various linguistic knowledge tasks in BERT are within 4 and 9. This shows that metaphor detection in PLM representations can be resolved earlier than some basic linguistic tasks.

In §3.1, we elaborated a hypothesize that the process of detecting metaphors is not very deep since what it needs to do is mainly contrast prediction between source and target domains, and the deep layers do not represent the source domain well. Our reported probing results confirm that metaphor detection is not deep in PLM layers. To further evaluate our reasoning, we probe the domain knowledge in PLM representations across layers. We employ LCC's annotation of source domains, and run a similar MDL probing on different PLMs but for domain prediction. The obtained results, shown in Figure A.1 in appendix, demonstrate that the main domain information is represented in the initial layers (2-6), confirming that the source domain is dominated by other information in higher layers.

[3]For RoBERTa and in the case of TroFi and VUA Verbs, we see exceptional increases in the last layers.

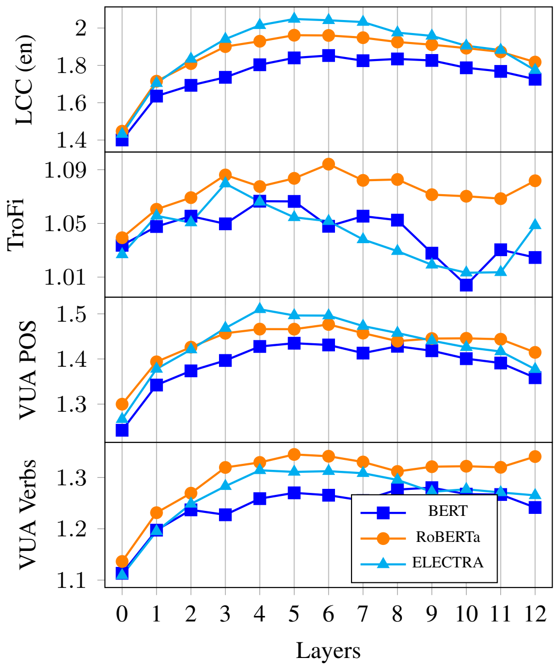

Figure 3: MDL compression across layers of three PLMs in four metaphor detection datasets. Higher number means better quality and extractability.

4.3 Generalization Experiments

As our PLMs, we use XLM-R (Conneau et al., 2020) for cross-lingual and BERT for cross-dataset experiments. To compare the cross-lingual and cross-dataset transferability, in §4.3.3, we employ the same setup, including using XLM-R as PLM for both. The results in §4.3.1 and 4.3.2 are not comparable. We apply the same edge probing architecture as in the probing experiments. We sometimes refer to both language and dataset as *domain*.

We run two experiments for each case of a source domain S and a target domain T: one with the PLM and one with a randomized version of the PLM where weights are set to random values. Randomly initialized Transformers with the same architec-

		Train Language			
		en	es	fa	ru
Test Lang.	en	**85.14** (65.37)	79.31 (52.71)	77.59 (50.22)	<u>80.51</u> (52.40)
	es	79.40 (53.17)	**84.59** (66.09)	76.70 (50.32)	<u>79.68</u> (53.32)
	fa	75.70 (50.07)	75.29 (52.65)	**81.04** (65.91)	<u>77.14</u> (50.36)
	ru	<u>83.92</u> (53.25)	80.54 (51.48)	76.61 (51.05)	**88.36** (67.98)

Table 4: Cross-lingual metaphor detection accuracies after five epochs of training for XLM-R and (its random version). For each test language, we bold its in-domain (e.g., en $\rightarrow$ en), and underline the best out-of-domain (e.g., ru $\rightarrow$ en) numbers.

ture as PLMs are common baselines in the community. The difference between the two gives evidence about the helpfulness of the encoded knowledge in doing the task. When $S = T$, this effect is measured for in-domain 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 metaphorical knowledge in PLM is and how consistent the annotations are.

4.3.1 Cross-lingual

The four LCC datasets corresponding to four languages are used here. We subsample from the datasets to have the same number of examples in the training sets, i.e., 12,238 which is the size of the Russian training set. The results are shown in Table 4. The random baseline is acquired using a randomly initialized XLM-R.

We observe that XLM-R significantly outperforms the random, confirming that metaphorical knowledge learned during the pre-training is transferable across languages. This considerable transferability can be attributed to the ability of XLM-R to build language-universal representations useful for metaphority transfer. Besides that, the innate similarities of metaphors in distinct languages can contribute to higher transferability, despite the lexicalization differences. E.g., analogizing a concept to a tool (en) occurs the same way in other languages like instrumento (es), ابزار (fa) and инструмент (ru). Finally, the constraints of the dataset producers in, for instance, keeping the languages in relatively similar target and source domains, could be influential. (See Figures A.2 and A.3).

An interesting observation is that training on Russian shows the best out-of-domain results when testing on other languages. We analyze this further. First, we observe that LCC(ru) has almost the closest target domain distribution to all other languages (See Table A.3 in Appendix). Second, LCC(ru) includes the least number of examples annotated with the source concept of "OTHER" (See Figure A.2 in Appendix), and thus, contains a larger number of effective training instances for the listed source domains. Thirdly, the reported results can also be influenced by the amount of data from each of these languages in the pre-training data of XLM-R. (e.g., Russian has the second largest size after English (Conneau et al., 2020)) Finally, for English, the higher-resource language than Russian, we find out that there are considerable number of examples in the LCC(en) related to "GUNS" and "CONTROL_OF_GUNS". These domains are not covered in other LCC datasets (See Figure A.3 in Appendix).

4.3.2 Cross-dataset

Similar to the cross-lingual evaluations, here we have four datasets used as sources and targets. We set the train size of each to the minimum of all, i.e., 3,838. For each pair, we run two experiments: one with randomized and one with pre-trained BERT as our PLM. Results are shown in Table 5.

PLM is much better than random in all out-of-domain cases, suggesting the presence of generalizable metaphorical information. As expected, VUA Verbs and POS achieve the best results when mutually tested, because, apart from the POS, they have the same distribution. VUA datasets and LCC(en) show good transferability, but the gap with in-domain results is still considerable (>13% absolute). VUA Verbs is the best source for TroFi, likely because of the POS match between them. Overall, apart from VUA datasets, the gap between in- and out-of-domain performance is large.

The random PLM accuracies range from about 54%-64% and 50%-56% for in- and out-of-domain

		Train Dataset			
		LCC(en)	TroFi	VUA POS	VUA Verbs
Test Dataset	LCC(en)	**84.26** (54.93)	62.04 (50.05)	70.35 (50.69)	70.37 (50.14)
	TroFi	59.49 (50.58)	**68.73** (64.96)	55.38 (49.45)	59.67 (53.68)
	VUA POS	62.23 (51.47)	55.29 (50.47)	**76.86** (56.01)	71.6 (53.47)
	VUA Verbs	60.20 (50.88)	54.55 (51.73)	72.6 (56.01)	**75.21** (60.03)

Table 5: Cross-dataset edge probing accuracy results on BERT is 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., VUA Verbs → VUA Verbs), and underline the best out-of-domain (e.g., VUA POS → VUA Verbs) numbers.

cases. We hypothesize that this drop in the out-of-domain is related to the annotation biases, which a randomly initialized classifier can leverage better when testing and training sets are from the same distribution. When the sets have distinct distributions, the biases do not transfer well.

4.3.3 Comparing cross-dataset and cross-lingual

LCC(en)	LCC(es)	LCC(fa)	LCC(ru)
82.31	78.02	77.3	78.04
TroFi	VUA POS	VUA Verbs	
60.54	68.61	67.15	

Table 6: Comparing cross-dataset and cross-lingual scenarios using the same model (XLM-R), training size, testing set, i.e., LCC(en), and different training sources.

As additional transferability analysis, we compare cross-lingual and cross-dataset results, by using XLM-R and evaluating different training sources on LCC(en) test set. We make the size of each train set to be the same (3,838). The results are shown in Table 6, where the first and second rows belong to cross-lingual and cross-dataset, respectively. To base our results, we include the in-domain result of training on LCC(en), i.e., 82.31%.

Clearly, there is a substantial gap between cross-lingual and cross-dataset accuracies. Several justifications could be mentioned for this gap. As we discussed in §4.3.1, in the LCC language datasets, source and target domains of metaphors are similarly distributed (See Figures A.2 and A.3 in Appendix). Further, the annotation guideline is consistent in the LCC language datasets. On the other hand and in the cross-dataset settings, we have datasets that differ in many aspects, including annotation procedure and definitions, covered part-of-speeches (e.g., Trofi and VUA Verbs vs. LCC and VUA POS) and sentence lengths (LCC: 25.9, VUA: 19.4, Trofi: 28.3).

5 Discussion and Conclusion

Metaphors are important in human cognition, and if we seek to build cognitively inspired or plausible language understanding systems, we need to work more on their best integration in the future. Therefore, any work in this regard is impactful.

Our probing experiments showed that PLMs do in fact represent the information necessary to do the task of metaphor detection. We assume this information is related to metaphorical knowledge learned during pre-training. Further, the layer-wise analysis confirmed our hypothesis that middle layers are more informative.

Even though our probing experiments did show that metaphorical knowledge is present in PLMs, it was still unclear if this knowledge is generalizable beyond the training data. So, to probe the probe and evaluate generalization, we ran cross-lingual and cross-dataset experiments. Our results showed that the transferability across languages works quite well for the four languages in LCC annotation. However, when the definitions and annotations were inconsistent across different datasets, the cross-dataset results were not satisfactory.

Overall, we conclude that metaphorical knowledge does exist in PLM representations and in middle layers mainly, and it is transferable if the annotation is consistent across training and testing data. We will explore more the cross-lingual transfer of metaphors and the impact of cross-cultural similarities in the future. Also, the application of metaphorical knowledge for text generation is something important that we will address.

References

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. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 375–388, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

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.

A Appendices

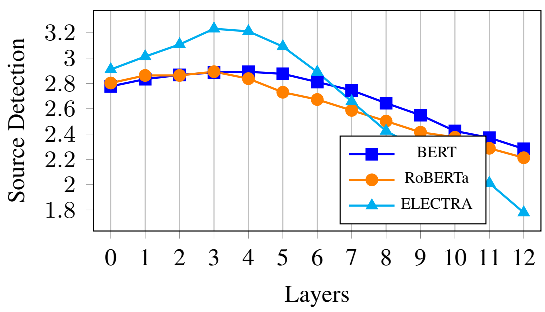

Figure A.1: MDL probing compression across layers for source and target detection. It is just for LCC(en) dataset, for it is the only dataset having the source and target concept of several samples.

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 A.1: Examples of sentences, spans, and annotations for LCC dataset in Farsi, Spanish, and Russian.

	en	es	fa	ru
en	0.0000			
es	**0.1622**	0.0000		
fa	0.1851	0.1688	0.0000	
ru	0.1833	0.2239	<u>0.2244</u>	0.0000

Table A.2: Jensen–Shannon divergence between source concept frequency distribution of different languages. The datasets are the same ones used in cross-lingual experiments where train set sizes are set to 12,238. Bold denotes the closest distributions and underline denotes the furthest distributions.

	en	es	fa	ru
en	0.0000			
es	0.4116	0.0000		
fa	<u>0.5004</u>	0.2148	0.0000	
ru	0.4291	**0.1209**	0.2141	0.0000

Table A.3: Jensen–Shannon divergence between target concept frequency distribution of different languages. The datasets are the same ones used in cross-lingual experiments where train set sizes are set to 12,238. Bold denotes the closest distributions and underline denotes the furthest distributions.

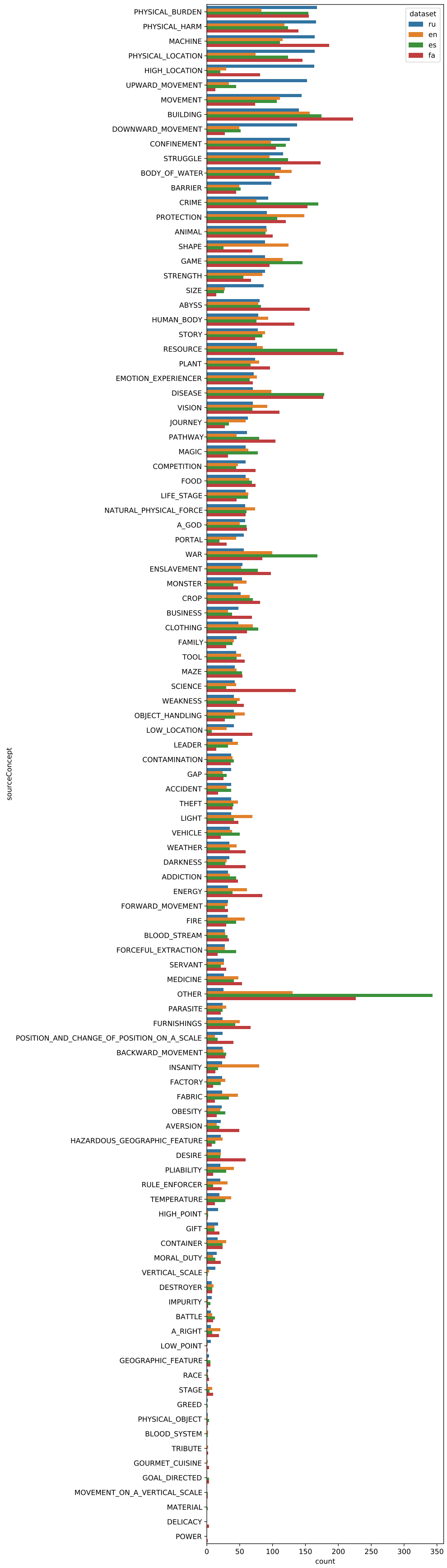

Figure A.2: Source concept frequency in training set of cross-lingual datasets.

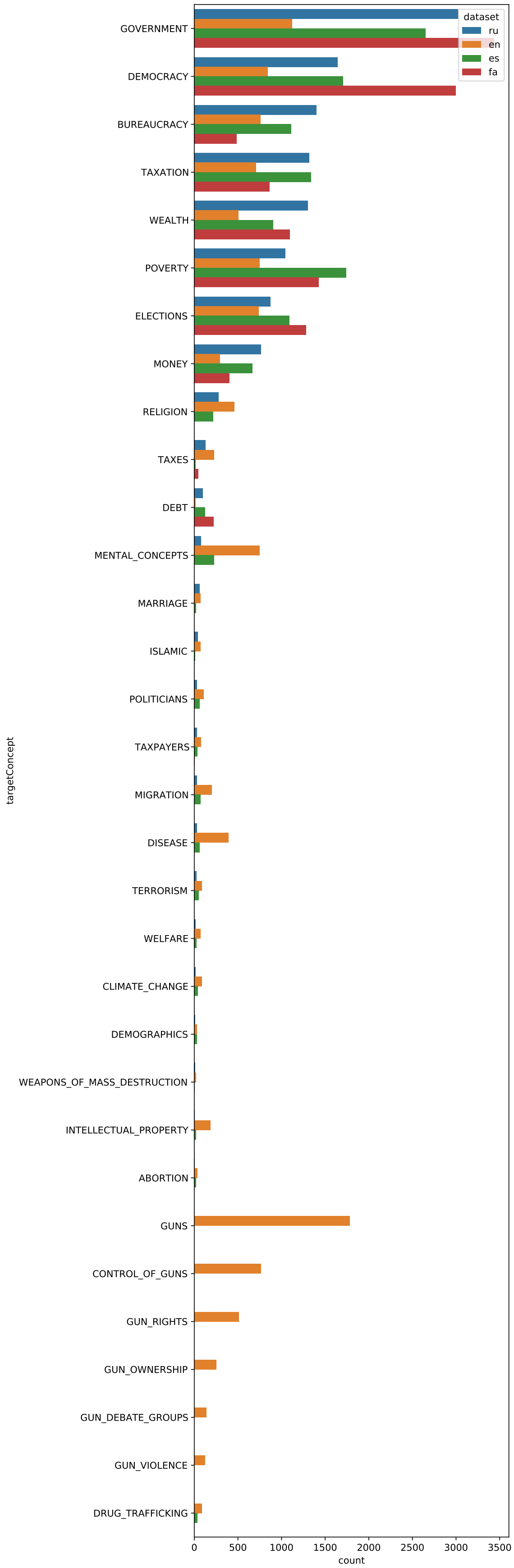

Figure A.3: Target concept frequency in training set of cross-lingual datasets.