Metaphors in Pre-Trained Language Models: Probing and Generalization Across Datasets and Languages

Ehsan Aghazadeh*, Mohsen Fayyaz*, Yadollah Yaghoobzadeh

School of Electrical and Computer Engineering,
University of Tehran, Tehran, Iran

Introduction

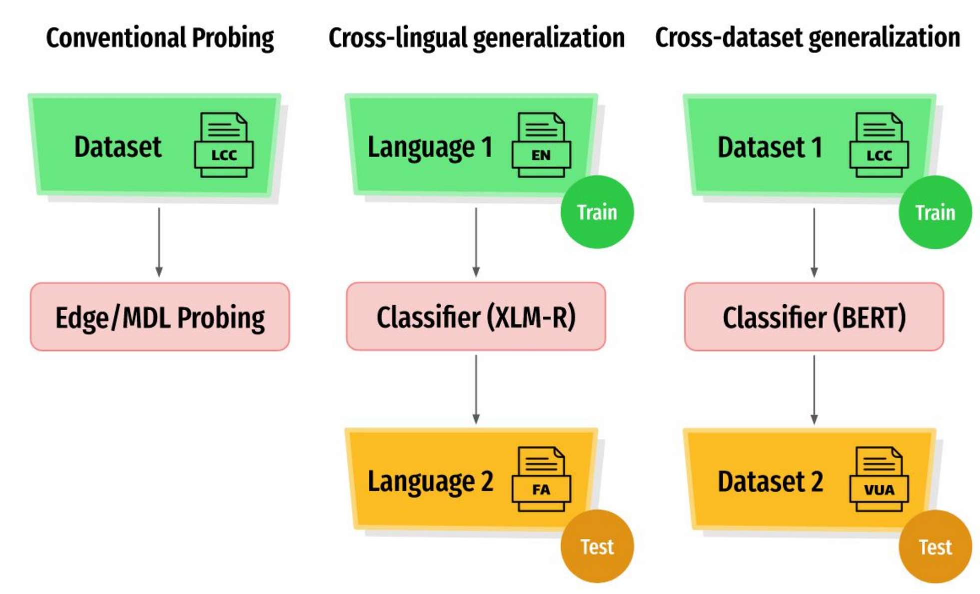
Summary

- Metaphors are essential in human communication and constructing human-like computational systems.
- We analyze and answer this question:

"do our pre-trained language models represent metaphors?"

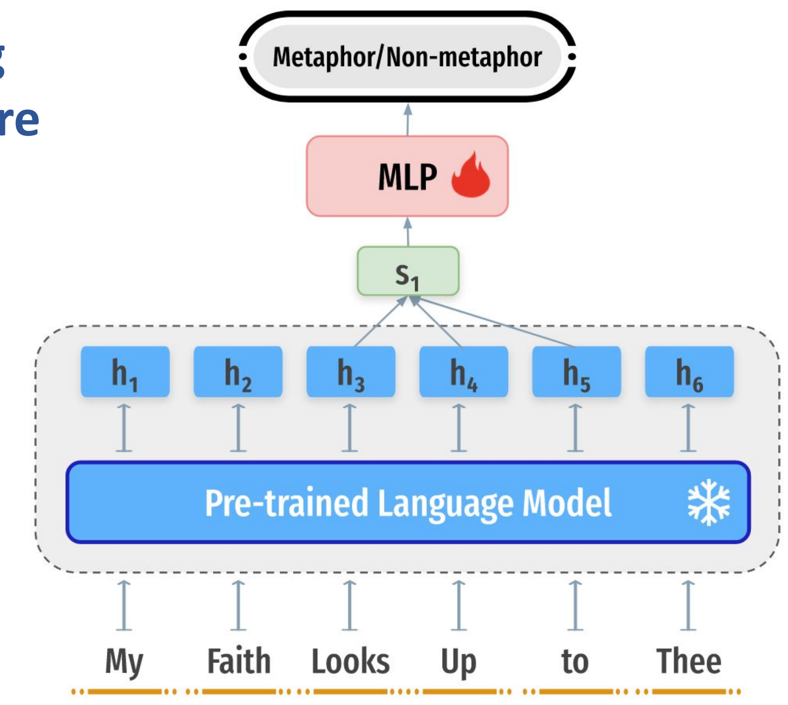
- We find that:
  - PLMs do encode metaphorical knowledge
  - Metaphorical knowledge is encoded better in the middle layers
  - Metaphorical knowledge is transferable between languages and datasets

Our probing and generalization scenarios



- To see if PLMs encode generalizable metaphorical knowledge, we evaluate them in settings where testing and training data are in different distributions.
- We present studies in multiple metaphor detection datasets and in four languages (i.e., English, Spanish, Russian, and Farsi).

Probing Structure



[1], [2]

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 is not a **[disease]**$_1$. → 3.0
	But the Supreme Court says it's not a way to **[hurt]**$_1$ the Second Amendment → 2.0
	Is he angry that gun rights **[progress]**$_1$ has been done without him? → 1.0
	I mean the 2nd amendment **[suggests]**$_1$ a level playing field for all of us. → 0.0

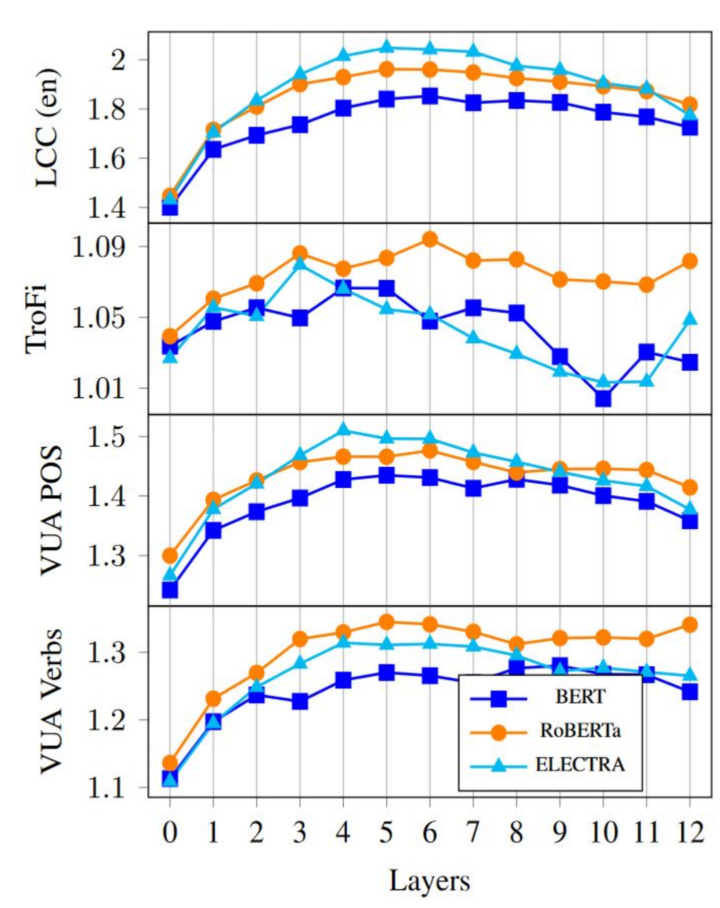
Probing Results

MDL Probing Compression (Best Among Layers) / Edge Probing Accuracy

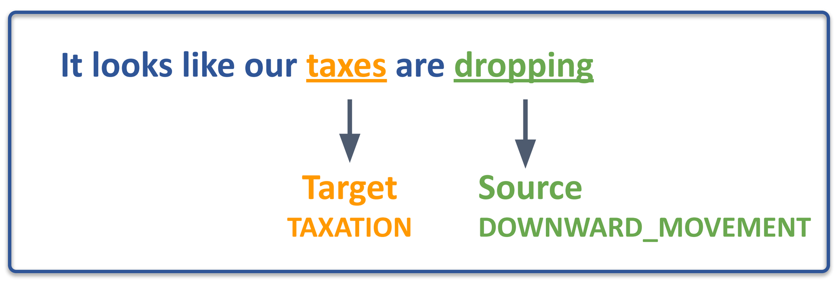
Dataset	Baseline Acc.	Baseline Comp.	BERT Acc.	BERT Comp.	RoBERTa Acc.	RoBERTa Comp.	ELECTRA Acc.	ELECTRA Comp.
LCC (en)	74.86	1.05_2	88.25	1.85_6	88.06	1.96_5	**89.30**	**2.05_5**
TroFi	67.34	1.01_4	**68.58**	1.07_4	68.46	**1.09_6**	68.07	1.08_3
VUA POS	65.92	1.03_0	80.32	1.43_5	81.72	1.48_6	**83.03**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

- RoBERTa and ELECTRA are shown to encode metaphorical knowledge better than BERT.
- This is consistent with their better performance on various tasks

MDL Probing Compression across layers



- Metaphorical information is more concentrated in the middle layers.
- To detect metaphors, we mainly need to predict if the *source* and *target* domains contrast. That is done in the earlier and middle layers.



Generalization Experiments

Cross-lingual Generalization for XLM-R (and its random version)

Test Lang	Train Lang: en	es	fa	ru
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)

- 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 metaphoricity transfer.
- Moreover, the innate similarities of metaphors in distinct languages can contribute to higher transferability, despite the lexicalization differences.

Cross-dataset Generalization for BERT (and its random version)

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

- PLM is much better than random in all out-of-distribution cases, suggesting the presence of generalizable metaphorical information.
- The random PLM accuracies range from about 54%-64% and 50%-56% for in- and out-of-distribution cases. We hypothesize that this drop in the out-of-distribution is related to the annotation biases, which a randomly initialized classifier can leverage better when testing and training sets are from the same distribution.
- There is a substantial gap between cross-lingual and cross-dataset accuracies. This can be attributed to that the annotation guideline is consistent in the LCC language datasets, while for the cross-dataset settings, we have datasets that differ in many aspects.

Conclusions

- We confirm that contextual representations in PLMs do encode metaphorical knowledge.
- We show that metaphorical knowledge is encoded better in the middle layers of PLMs.
- Our extensive experiments suggest that metaphorical knowledge is transferable between languages and datasets, especially when the annotation is consistent across training and testing sets.

References

- [1] Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy.
- [2] 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.
- [3] 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.