Multilingual Language Models

by Tang et al. [2024], Zhao et al. [2024]

Maya Arseven & Junhong Cai

Institute of Computational Linguistics
Heidelberg University

January 16, 2025

Outline

1. Introduction
2. Tang et al.: Finding Language-Specific Neurons
3. Zhao et al.: Tracing the Roots of Facts
4. Conclusion

Motivation

- Language models can operate **multilingually**
  - sometimes even **without** explicit language alignment [Kulshreshtha et al., 2020, Cao et al., 2020]

Motivation

- Language models can operate **multilingually**
  - sometimes even **without** explicit language alignment [Kulshreshtha et al., 2020, Cao et al., 2020]

→ How do they actually **process** diverse languages?

Motivation

- Language models can operate **multilingually**
  - sometimes even **without** explicit language alignment [Kulshreshtha et al., 2020, Cao et al., 2020]

→ How do they actually **process** diverse languages?

- Do they *code-switch*?
- How do they represent **knowledge** in different languages?

Background

Neurons:

- Respond to syntactic triggers [Wang et al., 2023] and encode positional information [Voita et al., 2024]
- Store factual information [Dai et al., 2022]
- Knowledge can be edited by manipulations [Meng et al., 2022]

Background

Neurons:

- Respond to syntactic triggers [Wang et al., 2023] and encode positional information [Voita et al., 2024]
- Store factual information [Dai et al., 2022]
- Knowledge can be edited by manipulations [Meng et al., 2022]

Multilinguality:

- Both language-specific and -agnostic parameter spaces [Foroutan et al., 2022]
- Linguistic similarity correlates with cross-language transfer [Philippy et al., 2023]

Papers

Language-Specific Neurons: The Key to Multilingual Capabilities in Large Language Models by [Tang et al., 2024]

Tracing the Roots of **Facts** in Multilingual Language Models: Independent, Shared, and Transferred Knowledge by [Zhao et al., 2024]

Outline

1. Introduction
2. Tang et al.: Finding Language-Specific Neurons
3. Zhao et al.: Tracing the Roots of Facts
4. Conclusion

Main Idea

- **Language-agnostic** and **language-specific** regions in LMs

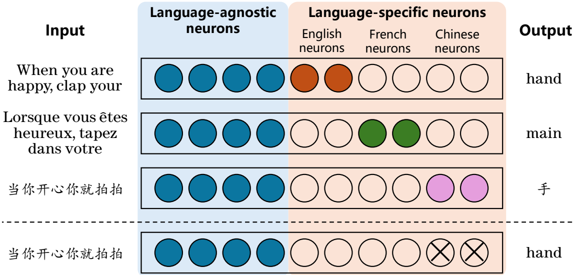

Figure 1: Activated neurons in LMs in different languages.

Main Idea

- **Language-agnostic** and **language-specific** regions in LMs

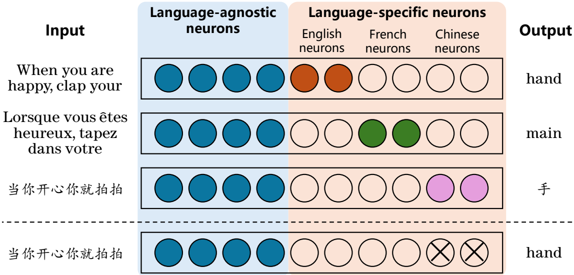

Figure 1: Activated neurons in LMs in different languages.

→ Detect **language-specific** neurons by computing **neurons activation likelihood** to different languages

Activation Probability

- Given the language k, how probable is it that the j_{th} neuron in the i_{th} layer to activate?

$$p_{i,j}^{k} = \mathbb{E}\left(\mathbb{I}\left(\text{act\_fn}(\tilde{h}^{i} W_1^{i})_j > 0\right) \mid \text{language } k\right),$$

1. **Distribution** for each neuron and each language indicating for which language a neuron fires
   - active if activation value exceeds 0
2. **L1 normalization** to convert this into a *probability* distribution

LAPE: Language Activation Probability Entropy

- The entropy of $p_{i,j}^{k}$ quantifies the neuron's activation reaction to language k

$$\text{LAPE}_{i,j} = -\sum_{k=1}^{l} p'^{k}_{i,j} \log(p'^{k}_{i,j}).$$

LAPE: Language Activation Probability Entropy

- The entropy of $p^k_{i,j}$ quantifies the neuron's activation reaction to language k

$$\mathrm{LAPE}_{i,j} = -\sum_{k=1}^{l} p'^{k}_{i,j} \log(p'^{k}_{i,j}).$$

Neurons with **low** LAPE → language **specific** neurons:

- $\begin{bmatrix}0 & 0 & 0 & 0 & 0.8 & 0 & 0 & 0.2\end{bmatrix} \rightarrow \mathrm{LAPE} \approx 0.5004$
  - only active to few languages → uncertainty is low

LAPE: Language Activation Probability Entropy

- The entropy of $p^k_{i,j}$ quantifies the neuron's activation reaction to language k

$$\text{LAPE}_{i,j} = -\sum_{k=1}^{l} p'^k_{i,j} \log(p'^k_{i,j}).$$

Neurons with **low** LAPE $\rightarrow$ language **specific** neurons:

- $\begin{bmatrix} 0 & 0 & 0 & 0 & 0.8 & 0 & 0 & 0.2 \end{bmatrix} \rightarrow \text{LAPE} \approx 0.5004$
  - only active to few languages $\rightarrow$ uncertainty is low

Neurons with **high** LAPE $\rightarrow$ language **agnostic** neurons:

- $\begin{bmatrix} 0.2 & 0.1 & 0 & 0 & 0.3 & 0.1 & 0 & 0.3 \end{bmatrix} \rightarrow \text{LAPE} \approx 1.5048$
  - active to many languages $\rightarrow$ uncertainty is high

Language Models

Model	Number of Neurons
llama2-7b	352k
llama2-13b	553k
llama2-70b	2.29M
bloom-7.1b	492k

Table 1: Number of Neurons in Different Models [Touvron et al., 2023, Scao et al., 2023].

- LLaMA-2: bigger and better but primarily trained on English
- BLOOM: trained on a balanced dataset with different languages

Tasks & Dataset

Identify language-specific neurons in two scenarios:

- Language modeling
  - perplexity scores on multilingual[1] Wikipedia corpora

$$\text{Perplexity} = 2^{H(X)}$$

[1]Considered languages: English, simplified Chinese, French, Spanish, Vietnamese, Indonesian and Japanese (not for BLOOM)

Tasks & Dataset

Identify language-specific neurons in two scenarios:

- Language modeling
  - perplexity scores on multilingual[1] Wikipedia corpora

$$\text{Perplexity} = 2^{H(X)}$$

- Open-ended generation
  - translated Vicuna [Chiang et al., 2023] questions using gpt-4
  - resulting texts are assessed by gpt4 on a 1-10 scale

[1]Considered languages: English, simplified Chinese, French, Spanish, Vietnamese, Indonesian and Japanese (not for BLOOM)

Methods

1. LAPE
2. LAP: Language Activation Probability
   - Language-specific if a neurons activation exceeds 95%
3. LAVE: Language Activation Value Entropy
   - LAPE but **mean** activation probability **across** languages
4. PV: Parameter Variation
   - Model parameters are compared before and after monolingual instruction tuning [Zhang et al., 2024]
   - Low rate of change in few languages → language-specific
5. Random Selection (RS)

Experiment Setup

1. Input tokens to the LM
2. Compute their LAPE score
3. Select the neurons that:
   - fall within the lowest percentile (bottom 1%) of LAPE scores
   - exceed the activation probability threshold (95%): `llama2-70b` → 0.515
4. Calculate their perplexity
   - lower the better

0. Main Experiment

- Deactivating language-specific regions by setting their activation values to 0
- If diagonal: neurons do impact the multilingual capabilities

LAPE: consistent diagonal entries across models

LAVE & **LAP**: cross-lingual interference

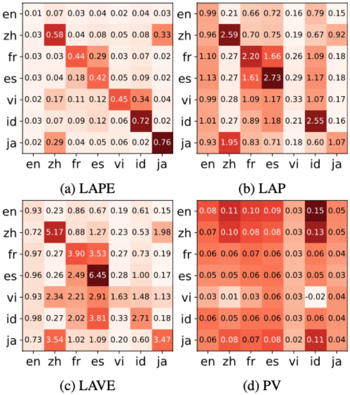

Figure 2: Four methods on the perplexity of LLaMA-2 (7B).

0. Main Experiment

	zh	fr	es	vi	id	ja
Normal Activation	4.30	4.19	3.51	3.70	4.16	2.86
Random Deactivation	4.18	4.22	3.35	3.53	4.42	2.99
zh	**2.46**	3.56	2.96	3.64	3.56	2.31
fr	3.69	**2.50**	2.29	3.01	3.59	2.76
es	3.51	2.57	**2.01**	3.14	3.34	2.56
vi	3.93	3.19	2.49	**2.74**	3.59	2.74
id	3.67	3.10	2.67	3.21	**2.84**	2.80
ja	3.21	3.69	3.07	3.49	3.37	**1.84**

Table 2: Performance of LLaMA-2 (70B) on the multilingual Vicuna as evaluated by GPT-4.

- Deactivated k-specific neurons $\rightarrow$ no more quality content generated in language k

0. Main Experiment

Question 你是一位登上珠穆朗玛峰顶峰的登山者。描述一下你... (*Translation*: You are a mountain climber reaching the summit ..)
Normal output 我是一个登上珠穆朗玛峰顶峰的登山者。当我站在山顶... (*Translation*: I am a climber who has reached the ...)
Deactivated output 我是一個登上珠穆朗瑪峰頂峰的登山者。I am a mountaineer who has climbed to the top of Mount Everest. 當我站在珠my朗ma峰頂峰，我感到非常興奮和欣慰。. . .

Table 3: Example of LLaMA-2-70B responses to a question in Chinese. The output is generated when Chinese neurons are deactivated.

1. Distribution and Identification Ratio

en	zh	fr	es	vi	id	ja
836	5,153	6,082	6,154	4,980	6,106	5,216

Table 4: The number of neurons in each language in LLaMA-2-70B.

- Total of 23k language-specific neurons
- The distribution is relatively even
  - except English, which requires fewer neurons to support the dominant language

1. Distribution and Identification Ratio

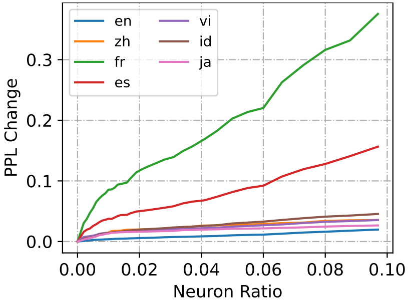

Figure 3: Change in perplexity across languages upon incremental of language-specific neurons when deactivating French neurons.

- Examining the top 1-10% of the activated neurons
- Processing French becomes harder
  - Spanish also worsens → both Romance languages

2. Structural Distribution Analysis

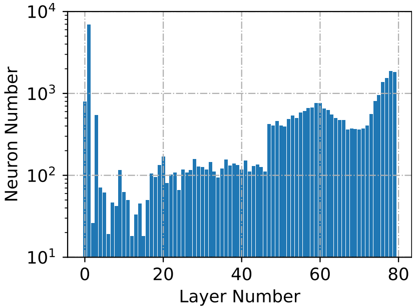

Figure 4: Distribution of language-specific neurons across different layers in LLaMA-2 (70B).

- Language processing is concentrated at **bottom** and **top** layers
  - Second layer: 7k
  - Final four layers: 1k each

2. Structural Distribution Analysis

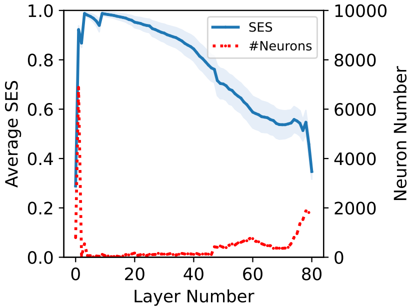

Figure 5: The mean SES between all language pairs and total language neuron numbers across layers.

- Opposite trend on the sentence embedding similarity (SES)
  - Bottom layers: mapping different languages into a **shared** representation
    - could be English
  - Top layers: **vocabulary** mapping to the respective language

3. Language Dominance Analysis

Are low-resource languaged **dominated** by high ones?

1. Compute mean sentence embeddings score (SES) for all sentence pairs between k and c
   - a larger SES indicates c has a larger dominance
2. Obtain v_k as the target language vector
3. Conduct the space mapping
4. Transfer SES into the same space around v_k
5. Observe their alignment

3. Language Dominance Analysis

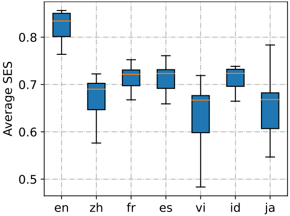

Figure 6: LLaMA-2 (70B)

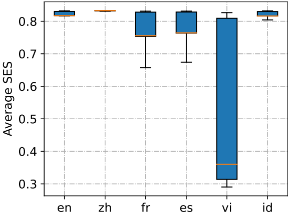

Figure 7: BLOOM (170B)

→ Low-resource languages are centered around high-resource languages

- English in very dominant in LLaMA, whereas BLOOM shows more distributed results

4. Steering Output Language

Steering the output language can mitigate the **off-target** language issue [Gu et al., 2019, Sennrich et al., 2024]
↪ prompted in Chinese but answered in English

- Language-specific neurons are **not** rightfully activated
- **Manual** activation by assigning an activation value

4. Steering Output Language

Metrics	Settings	zh	fr	es	vi	id	ja
Language accuracy	normal	0.87	0.73	0.81	0.60	0.40	0.79
	steered	0.99	0.90	0.93	0.97	0.99	1.00
Content quality	normal	4.30	4.19	3.51	3.70	4.16	2.86
	steered	4.57	4.35	4.02	3.57	4.28	2.91

Table 5: The language accuracy of the normal and steered output by activating language-specific neurons.

- More responses in the target language and improvement in the generation quality

4. Steering Output Language

Question
¿Cómo puedo mejorar mis habilidades de gestión del tiempo? (*Translation*: How can I improve my time management skills?)
Steered output in Chinese
Time管理是一个非常重要的技能，因为它可以提高你的生产力并使你更有效地完成任务。下面是一些提示... (*Translation*: Time management is a very important skill because it increases your productivity and allows...)

Table 6: An example of asking question in Spanish and answering it in Chinese.

- Altering the output language through neuron-level manipulations
- By **deactivating** the question language and **activating** answer language neurons

Main Findings

1. Language-specific neurons are at a **minuscule** proportion
   - 23k in 2.29M $\approx$ 1%
   - when deactivated, understanding and generation in that language significantly decreases

Main Findings

1. Language-specific neurons are at a **minuscule** proportion
   - 23k in 2.29M $\approx$ 1%
   - when deactivated, understanding and generation in that language significantly decreases
2. They are mostly located in the **bottom** and **top** layers of model
   - bottom: process the input to a higher representation
   - top: project this back to the target language

Main Findings

1. Language-specific neurons are at a **minuscule** proportion
   - 23k in 2.29M $\approx$ 1%
   - when deactivated, understanding and generation in that language significantly decreases
2. They are mostly located in the **bottom** and **top** layers of model
   - bottom: process the input to a higher representation
   - top: project this back to the target language
3. Generation can be **steered** by selectively activating and deactivating these neurons
   - solution to off-target language issue

Outline

Research Questions

Tracing the Roots of Facts in Multilingual Language Models: Independent, Shared, and Transferred Knowledge

1. How does **factual probing performance** of ML-LMs differ across languages, and what factors affect these differences?
2. Do ML-LMs represent the same fact in different languages with a **shared** or **independent** representation?
3. What mechanisms during the **pre-training** of ML-LMs affect the formation of cross-lingual fact representations?

Previous Works on Multilingual Factual Probing

- Use the fill-in-blank cloze question dataset to query PLMs to explore their ability of handling factual knowledge.[Petroni et al., 2019]

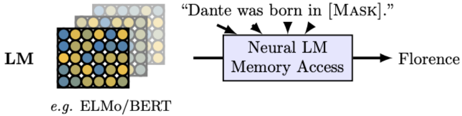

Figure 8: Query LMs for factual knowledge

- Specific fact representation are linked to specific set of neurons rather than the whole space. $\rightarrow$ Enhance models through neurons adjustment.[De Cao et al., 2021]

Previous Works and Targets

- Investigate PLMS' representation of facts in languages other than English. → Languages with limited resources have weaker predictability.[Devlin et al., 2019]
- **Cultural biases** of the datasets might affect the predictability of PLMs[Fierro and Søgaard, 2022].

Targets:

- Clarify how facts are perceived and identify the difference in fact recognition among languages.
- Investigate how ML-LMs learn and represent facts.

Experiment Setup of Probing Factual Knowledge

Datasets:

- mLAMA: multilingual extension of LAMA, contains 37,498 instances spanning 43 relations. (represented as fill in blank cloze)[Kassner et al., 2021].

Example: "[X] is the capital of [Y]."

Models:

Encoder-based ML-LMs

- multilingual Bert(mBERT)[Devlin et al., 2019]
- XLM-R[Conneau et al., 2020]

Why not generative models?

Protocol of Probing ML-LMs

- **Full Match:**
  Assign exact number of mask tokens of object Y.
- **Partial Match:**
  List all object Y and their token counts associated with the template.
  A fact was considered correctly predicted if any version of the prompt included the correct object tokens, regardless of additional preceding or succeeding tokens.

  "The Beatles plays [MASK] music."

  The Beatles plays [MASK][MASK][MASK]MASK] music"

  How to decide the longest mask token sequence?

Probing Results

- Why we have the distribution here? Isn't P@1 a number?

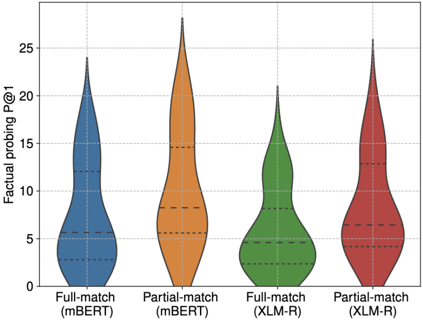

Figure 9: Probing P@1 on mLAMA for full and partial match methods with mBERT and XLM-R.

Probing Results

ISO	Language	mBERT Full	mBERT Partial	XLM-R Full	XLM-R Partial	ISO	Language	mBERT Full	mBERT Partial	XLM-R Full	XLM-R Partial
en	English	19.07	**22.57**	17.08	21.17	cs	Czech	5.63	**8.62**	1.21	4.34
id	Indonesian	18.15	**22.43**	13.99	19.23	ceb	Cebuano	5.11	**5.84**	0.76	0.88
it	Italian	16.94	**19.78**	10.80	13.53	et	Estonian	4.97	**8.24**	3.82	6.01
de	German	16.91	**20.33**	12.06	14.78	sq	Albanian	4.93	**5.62**	3.31	4.13
es	Spanish	16.65	**20.28**	10.51	12.87	sk	Slovak	4.90	**7.08**	2.84	4.84
nl	Dutch	15.98	**18.30**	10.47	13.04	bg	Bulgarian	4.51	6.58	5.07	**7.44**
pt	Portuguese	14.76	**17.96**	14.05	17.12	ur	Urdu	4.41	**8.02**	4.40	6.31
ca	Catalan	14.11	**17.05**	5.23	8.60	uk	Ukrainian	3.84	**6.56**	0.64	4.18
tr	Turkish	14.08	**17.65**	13.79	17.47	fi	Finnish	3.58	7.11	4.43	**8.54**
da	Danish	13.56	**16.61**	12.01	15.63	hy	Armenian	3.25	**5.01**	3.90	4.66
ms	Malay	13.14	**16.99**	11.20	14.76	sr	Serbian	3.07	**5.95**	2.45	5.59
sv	Swedish	12.89	**15.32**	11.63	13.63	hi	Hindi	2.95	5.63	3.78	**6.61**
fr	French	12.68	**20.18**	7.79	13.81	be	Belarusian	2.80	**4.49**	0.78	1.54
af	Afrikaans	12.05	**14.47**	8.17	10.09	eu	Basque	2.45	**5.42**	1.19	2.46
ro	Romanian	11.33	14.23	13.38	**17.46**	lv	Latvian	2.15	**3.79**	1.66	2.94
vi	Vietnamese	10.93	14.58	11.78	**15.67**	az	Azerbaijani	1.99	5.60	3.21	**6.38**
gl	Galician	10.00	**13.03**	6.04	8.00	ru	Russian	1.90	**5.98**	0.79	4.07
fa	Persian	8.67	**12.47**	7.30	9.36	bn	Bangla	1.76	3.12	2.67	**4.10**
cy	Welsh	7.98	**9.16**	5.08	6.05	ka	Georgian	1.45	1.79	1.89	**2.31**
el	Greek	7.24	**8.17**	5.68	7.41	ja	Japanese	1.34	4.85	4.78	**5.26**
he	Hebrew	6.78	**9.09**	4.60	6.44	sl	Slovenian	1.26	**3.80**	1.77	3.70
ko	Korean	6.73	**9.24**	7.18	6.44	lt	Lithuanian	1.25	1.94	2.31	**3.42**
zh	Chinese	6.51	**11.95**	4.05	5.91	la	Latin	1.21	2.24	1.83	**2.53**
pl	Polish	6.33	**8.45**	5.09	8.30	ga	Irish	0.96	**1.31**	0.56	0.75
ar	Arabic	6.11	**8.25**	6.16	7.63	ta	Tamil	0.90	**1.93**	0.93	1.24
hu	Hungarian	5.86	10.08	5.42	**11.17**	th	Thai	0.49	0.65	2.75	**4.26**
hr	Croatian	5.65	**9.51**	2.36	5.27	Macro average		8.85	**11.84**	6.88	9.52

Figure 10: P@1 for 53 languages on mLAMA using full- and partial-match methods with mBERT and XLM-R.

Confusions of their Arguments

- Non-essential tokens such as whitespaces are produced with both full and partial match.
- Partial match offers better representation, but they choose full-match approach in the following discussions.

Type	Example
Whitespace	Petr Kroutil was born in Prague ( ).
Preposition	Galactic halo is part of (the) galaxy.
Related noun	Surinder Khanna was born in Delhi (,) (India).
Adjective	Pokhara Airport is a (popular) airport.

Figure 11: Four patterns discerned in facts predicted by partial-match method.

Discrepancy in Factual Probing across Languages

- Training data volume
- Mask token count
- Presence of localized knowledge cluster

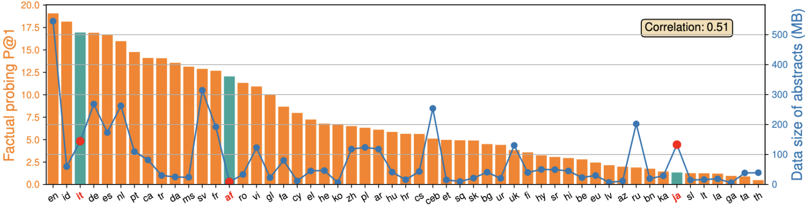

Figure 12: Wikipedia data size of abstracts vs. Factual probing P@1 on mLAMA in mBERT in 53 languages.

Training Data Volume

- **Pearson correlation coefficient** between P@1 and five metrics on the training data of mBERT:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- Moderate correlation indicates a limited impact of the training data volume on learning factual knowledge.

Statistics	Pearson's zr with P@1
The number of page count	0.43
The data size of articles	0.44
The data size of articles (bzipped)	0.45
The data size of abstracts	**0.51**
The data size of abstracts (bzipped)	0.48

Mask Token Count

- Is One-token P@1 the sub experiment of mBERT P@1?
- Potential cultural biases in mLAMA alone can't explain the substantial difference between mBERT P@1 performance of Italian and Japanese.
- XLM-R tokenizer captured more one-token entities in Japanese. Better performance on non-Latin languages.

	it	**ja**	**af**
mBERT P@1	16.94	1.34	12.05
One-token P@1	15.27	15.34	17.00
One-token entities	1675	126	498
XLM-R P@1	10.80	4.78	8.17
One-token P@1	13.67	14.73	16.58
One-token entities	923	244	333

Cross-Lingual Knowledge Sharing

- Jaccard Similarity:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}, \quad (1)$$

- Cross-lingual knowledge transfer does not occur universally across languages. $\rightarrow$ Localized knowledge sharing pattern.

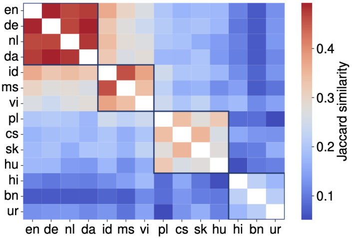

Figure 13: Jaccard similarity matrix of shared factual knowledge across languages with mBERT. How many facts two languages share.

Fact Representations

- Do ML-LMs Have Fact Representations Shared Across Language? → Two scenarios:

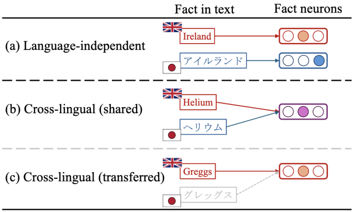

Figure 14: Three types of fact representation in ML-LMs

- Copy of same fact is independently maintained across language.
- Fact representations in different languages are unified in an embedding space.

Factual Neuron Probing

- Analyzed the representation of cross-lingual facts in ML-LMs by identifying their active neurons across languages. $\rightarrow$ **PROBLESS**[Antverg and Belinkov, 2022].
- Detect the deviation of neurons values from the average, so both positive and negative deviation is considered active.
- Predictable facts that share the same relation but vary in subject-object pairs.
- Collect neurons of [MASK] token identify active neurons as signatures of the fact representations. **Average Pooling** for multi-tokens masks.
- Collect active neurons for the same fact in various languages.Focused on the top 30 languages by P@1 score.

Results of Factual Neuron Probing

- The presence of both independent and cross language fact representations in ML-LMs.

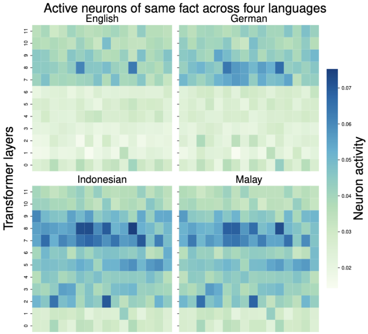

Figure 15: Neuron activity with mBERT in four languages in response to the query "William Pitt the Younger used to work in [MASK]."

Quantification of Cross-Language Sharing

- Given a set of shared facts, to what extent the two languages share the top 50 active neurons.
- No consistent geographical boundaries among languages.

Figure 16: Language similarity based on top 50 shared active neurons by probing on mLAMA with mBERT.

Formation of Cross-Lingual Representations of Facts

So far:

- Confirm the presence of cross-lingual representations by neuron probing and Jaccard similarity.

Next step:

- Access whether (1) Fact representations are learned individually from **distinct language corpora** and subsequently aligned into a **common semantic space**. (2) Acquire through **cross-lingual transfer**.

Roots – Data

- Verify if the fact originates from training data.
- Examine the occurrences of the subject and object.
- Using string matching between the object subject pair and Wikipedia text. Then access the co-occurrence.

Absent yet Predictable Facts

- Languages with more training data have better factual knowledge coverage.
- Languages like Afrikaans and Albanian can predict fact correctly even without existence in the training corpus.
- Indicate high possibility of effective cross-lingual tansfer.

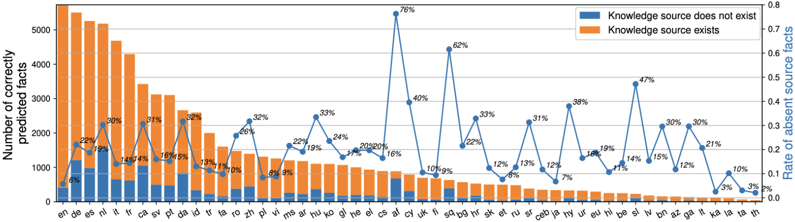

Figure 17: Number of correctly-predicted facts with mBERT in terms of existence of knowledge source.

Easy to Predict Facts

- Shared Entity Tokens:
  'Sega Sports R&D is owned by Sega.'
- Naming Cues:
  'The native language of Go Hyeon-jeong is Korean.'
- Other:
  The remaining facts are difficult to infer from the entities only, indicating the high possibility of cross-lingual transfer. Why???

Irregular!

- Statistics show that while cross-lingual transfer of factual knowledge in ML-LMs does occur, it is limited.

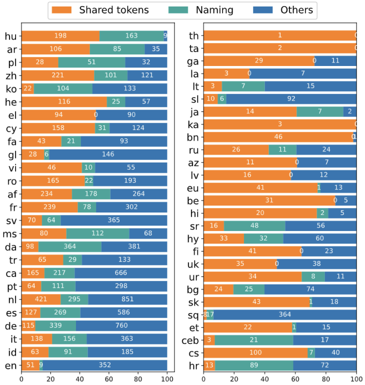

Figure 18: The count of three types of absent and predictable facts with mBERT.

Outline

Tang et al. [2024] Conclusion

1. Language-specific neurons are at a **minuscule** proportion.
2. They are mostly located in the **bottom** and **top** layers of model.
3. Generation can be **steered** by selectively activating and deactivating these neurons.

Zhao et al. [2024] Conclusion

1. Two methods including full and partial match are applied to prob factual knowledge of two ML-LMs mBERT and XLM-R. P@1 scores are relatively low especially in low resource languages.
2. Key factors like data volume, mask token count are evaluated on their influence to the discrepancy in factual probing across language.
3. Contradictions in sharing patterns among geographically proximate language clusters.
4. Three types of patterns for acquiring and representing factual knowledge across languages in MLLMs are identified through neuron probing.
5. Future work aims to enhance the cross-lingual fact representation learning in ML-LMs and develop a more precise factual probing dataset.

Questions?

Questions

Questions to Tang et al. [2024]

Q1:

Is the L1 normalization based on single neuron activations? If so, may that discriminate against language specific polysemantic neurons that activate in combination, though get disregarded since their single activation may lie below the chosen threshold?

Questions to Tang et al. [2024]

Q1:

Is the L1 normalization based on single neuron activations? If so, may that discriminate against language specific polysemantic neurons that activate in combination, though get disregarded since their single activation may lie below the chosen threshold?

They are applied to the **distribution** of neuron activations. I think they still acount for polysemanticy by not resticting each neuron to be responsible for one language. The idea is more to find the language specific **regions**, and not 1-1 mapping of language-neurons.

Questions to Tang et al. [2024]

Q2:

Why and how do the authors decide to target the “bottom 1%” of neurons? Where does the 1% come from?

Questions to Tang et al. [2024]

Q2:

Why and how do the authors decide to target the "bottom 1%" of neurons? Where does the 1% come from?

To **restrict** the amount of neurons they have to analyze. Since this is a percentage, the bottom 1% is an **empirical** threshold that they have set.

Questions to Zhao et al. [2024]

Q3:

Which kind of the 3 fact representations would be "ideal", which one would we actually wish for?

Only children select one answer from multiple-choice questions, the paper's experiment in neuron probing shows that both language-independent and cross-language exist. Human cannot wish, only god wishes. As a normal human, i would prefer explicit factual knowledge transfer among languages.

Questions to Zhao et al. [2024]

Q4:

Do you agree with the authors easy-to-learn fact types and ruleset or may there be others that they disregarded?

They only summarize two types of easy-to-learn fact, but the remaining other might be splited into finer categories? But i am not sure.

More Questions?

Questions

References I

[1] Omer Antverg and Yonatan Belinkov. On the pitfalls of analyzing individual neurons in language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=8uz0EWPQIMu.

[2] Steven Cao, Nikita Kitaev, and Dan Klein. Multilingual alignment of contextual word representations. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=r1xCMyBtPS.

[3] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

[4] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747/.

[5] Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.581. URL https://aclanthology.org/2022.acl-long.581/.

[6] Nicola De Cao, Wilker Aziz, and Ivan Titov. Editing factual knowledge in language models. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6491–6506, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.522. URL https://aclanthology.org/2021.emnlp-main.522/.

References II

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *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, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

[8] Constanza Fierro and Anders Søgaard. Factual consistency of multilingual pretrained language models. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3046–3052, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.240. URL https://aclanthology.org/2022.findings-acl.240/.

[9] Negar Foroutan, Mohammadreza Banaei, Rémi Lebret, Antoine Bosselut, and Karl Aberer. Discovering language-neutral sub-networks in multilingual language models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7560–7575, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.513. URL https://aclanthology.org/2022.emnlp-main.513/.

[10] Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor O.K. Li. Improved zero-shot neural machine translation via ignoring spurious correlations. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1258–1268, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1121. URL https://aclanthology.org/P19-1121/.

References III

[11] Nora Kassner, Philipp Dufter, and Hinrich Schütze. Multilingual LAMA: Investigating knowledge in multilingual pretrained language models. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3250–3258, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.284. URL https://aclanthology.org/2021.eacl-main.284/.

[12] Saurabh Kulshreshtha, Jose Luis Redondo Garcia, and Ching-Yun Chang. Cross-lingual alignment methods for multilingual BERT: A comparative study. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 933–942, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.83. URL https://aclanthology.org/2020.findings-emnlp.83/.

[13] Kevin Meng, David Bau, Alex J Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=-h6WAS6eE4.

[14] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1250. URL https://aclanthology.org/D19-1250/.

[15] Fred Philippy, Siwen Guo, and Shohreh Haddadan. Towards a common understanding of contributing factors for cross-lingual transfer in multilingual language models: A review. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5877–5891, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.323. URL https://aclanthology.org/2023.acl-long.323/.

References IV

[16] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, and François Yvon et al. Bloom: A 176b-parameter open-access multilingual language model, 2023. URL https://arxiv.org/abs/2211.05100.

[17] Rico Sennrich, Jannis Vamvas, and Alireza Mohammadshahi. Mitigating hallucinations and off-target machine translation with source-contrastive and language-contrastive decoding. In Yvette Graham and Matthew Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 21–33, St. Julian's, Malta, March 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.eacl-short.4/.

[18] Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. Language-specific neurons: The key to multilingual capabilities in large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5701–5715, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.309. URL https://aclanthology.org/2024.acl-long.309/.

[19] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, and Guillem Cucurull et al. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

[20] Elena Voita, Javier Ferrando, and Christoforos Nalmpantis. Neurons in large language models: Dead, n-gram, positional. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1288–1301, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.75. URL https://aclanthology.org/2024.findings-acl.75/.

[21] Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

References V

[22] Zhihao Zhang, Jun Zhao, Qi Zhang, Tao Gui, and Xuanjing Huang. Unveiling linguistic regions in large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6228–6247, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.338. URL https://aclanthology.org/2024.acl-long.338/.

[23] Xin Zhao, Naoki Yoshinaga, and Daisuke Oba. Tracing the roots of facts in multilingual language models: Independent, shared, and transferred knowledge. In Yvette Graham and Matthew Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2088–2102, St. Julian's, Malta, March 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.eacl-long.127/.