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

Xin Zhao
The University of Tokyo
xzhao@tkl.iis.u-tokyo.ac.jp

Naoki Yoshinaga **Daisuke Oba**
Institute of Industrial Science,
The University of Tokyo
{ynaga,oba}@iis.u-tokyo.ac.jp

Abstract

Acquiring factual knowledge for language models (LMs) in low-resource languages poses a serious challenge, thus resorting to cross-lingual transfer in multilingual LMs (ML-LMs). In this study, we ask how ML-LMs acquire and represent factual knowledge. Using the multilingual factual knowledge probing dataset, mLAMA, we first conducted a neuron investigation of ML-LMs (specifically, multilingual BERT). We then traced the roots of facts back to the knowledge source (Wikipedia) to identify the ways in which ML-LMs acquire specific facts. We finally identified three patterns of acquiring and representing facts in ML-LMs: language-independent, cross-lingual shared and transferred, and devised methods for differentiating them. Our findings highlight the challenge of maintaining consistent factual knowledge across languages, underscoring the need for better fact representation learning in ML-LMs.[1]

1 Introduction

To mitigate the inherent data sparseness of low-resource languages, multi-lingual language models (ML-LMs) such as mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020a), mT5 (Xue et al., 2021), and BLOOM (Scao et al., 2022) have been developed to transfer knowledge across languages. The effectiveness of this cross-lingual transfer in ML-LMs has been demonstrated on various language tasks (Wu and Dredze, 2019; Chi et al., 2020; Pires et al., 2019; Huang et al., 2023). However, a more challenging task is the cross-lingual transfer of specific factual knowledge, such as "Greggs is a British bakery chain." In many low-resource languages, text data about such knowledge might be minimal or non-existent. Effectively transferring knowledge is vital for applications that handle factual knowledge, such as fact checking and relation extraction (Lee et al., 2020; Verlinden et al., 2021).

[1]code: https://github.com/xzhao-tkl/fact-cl

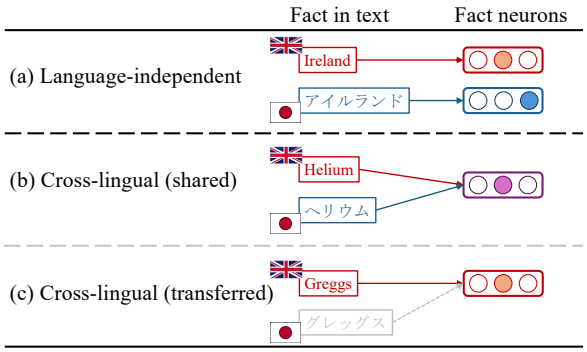

Figure 1: Three types of fact representation in ML-LMs; Facts are a) represented with distinct neurons across languages (language-independent), b) shared using the same neurons (cross-lingual (shared)), and c) transferred across languages (cross-lingual (transferred)).

Following early studies (Petroni et al., 2019; Jiang et al., 2020b) that used cloze-style queries to probe whether monolingual language models can recall factual knowledge, researchers probed ML-LMs (Jiang et al., 2020a; Kassner et al., 2021; Yin et al., 2022; Fierro and Søgaard, 2022; Keleg and Magdy, 2023). The results indicated that ML-LMs exhibit an ability to recall facts. However, the mechanism behind the acquisition and representation of facts in ML-LMs remains unclear.

In this study, we investigated whether and how low-resource languages can benefit from the cross-lingual transfer of factual knowledge (Figure 1). Concretely, we addressed three research questions:

RQ1: How does factual probing performance of ML-LMs differ across languages, and what factors affect these differences? (§4)

RQ2: Do ML-LMs represent the same fact in different languages with a shared or independent representation? (§5)

RQ3: What mechanisms during the pre-training of ML-LMs affect the formation of cross-lingual fact representations? (§6)

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 2088–2102

To answer these research questions, we started by probing two ML-LMs, mBERT and XLM-R, using the mLAMA probing dataset (Kassner et al., 2021). The results reconfirm that ML-LMs have difficulty recognizing facts in low-resource languages (Kassner et al., 2021), such as Irish and Lithuanian (§3). However, we also observed only a moderate correlation between probing performance and the amount of training data. Although the cultural bias of the mLAMA dataset may hinder probing performance in non-Latin script languages (Keleg and Magdy, 2023), the exact effect of a model's cross-lingual capabilities remains to be established.

To identify the role of cross-lingual capability in fact probing, we performed a neuron-level analysis for facts predicted correctly. By comparing active neurons across languages, we observed that identical facts in various languages are not acquired in identical ways. For specific facts, some languages exhibit similar neuron activity, while others display distinct patterns. We categorize the former as **cross-lingual fact representations**, as illustrated in Figure 1(b,c), and the latter as **language-independent representations**, as illustrated in Figure 1(a).

To further discern cross-lingual representations, we devised a method for tracing the roots of facts by verifying their presence in the knowledge source (Wikipedia for mBERT). We assume that the facts that are predicted correctly although absent in the training data are captured through cross-lingual transfer, referred to as **cross-lingual transferred** (Figure 1(c)) to differentiate from **cross-lingual shared** (Figure 1(b)). A deeper investigation into the results, however, revealed that only a fraction of those facts would be acquired through cross-lingual transfer. This underscores the limitations of current ML-LMs in cross-lingual fact representation.

The contributions of this paper are

- Evaluation of training data volume and mask token count as factors to cause discrepancies in probing results across languages and discovery of localized factual knowledge clusters,
- Establishment of methods for distinguishing among fact representations in ML-LMs, by identifying shared active neurons and tracing the roots of facts back to the training data, and
- Revelation that factual knowledge in ML-LMs has three types of representations: language-independent, cross-lingual (shared), and cross-lingual (transferred) (Figure 1).

2 Related Work

This section reviews existing studies on understanding the mechanism of cross-lingual transfer and factual knowledge probing. We first discuss key studies that investigated how knowledge is transferred across languages in ML-LMs. Next, we highlight research on how factual knowledge is perceived in pre-trained language models (PLMs).

2.1 Understanding cross-lingual transfer

Numerous studies have investigated the basic mechanisms of cross-lingual transfer in ML-LMs. Studies of the process of cross-lingual transfer have shown that, while shared tokens facilitate cross-lingual knowledge transfer, their effect is circumscribed (K et al., 2020; Conneau et al., 2020b). Subsequent studies showed that using parallel data enhances a model's cross-lingual ability (Moosa et al., 2023; Reid and Artetxe, 2023).

Concurrent studies focused on the realization of cross-lingual transfer in the parameter space within ML-LMs (Muller et al., 2021; Chang et al., 2022; Foroutan et al., 2022). They reported that ML-LMs have both language-specific and language-agnostic parameter spaces when representing identical linguistic knowledge across languages. However, they focused solely on basic linguistic tasks like dependency parsing and named-entity recognition. Cross-lingual representation of factual knowledge remains underexplored. Moreover, while these previous studies primarily provided a systematic explanation of cross-lingual transfer mechanisms, they neglected the detailed variations in how ML-LMs acquire and represent specific knowledge.

2.2 Factual knowledge probing

Understanding factual representation in PLMs has attracted much attention recently. Using fill-in-the-blank cloze question datasets, researchers explored the ability of PLMs to handle factual knowledge in the English language (Petroni et al., 2019; Heinzerling and Inui, 2021; Wang et al., 2022). To clarify the mechanism by which Transformer (Vaswani et al., 2017)-based PLMs represent facts, a few studies have conducted neuron-level investigation (Oba et al., 2021; Geva et al., 2021; Dai et al., 2022). These studies revealed that specific fact representation are linked to a specific set of neurons rather than the whole parameter space. This has led to subsequent research focused on enhancing models through neuron adjustments (De Cao et al.,

2021; Mitchell et al., 2022; Zhang et al., 2022).

Several studies have investigated the ability of PLMs to represent facts in languages other than English (Jiang et al., 2020a; Kassner et al., 2021; Fierro and Søgaard, 2022) in a multilingual setting. Their results suggest that the ability to perceive factual knowledge is not exclusive to English. Other languages demonstrated comparable proficiency. However, weaker predictability of factual knowledge has been observed for languages with limited resources. One study (Fierro and Søgaard, 2022) investigated the differences in predictability between languages and attributed them to cultural biases. However, the role of cross-lingual transfer in factual representation across languages has not been extensively explored.

3 Multilingual Factual Probing

we carried out experiments to probe the factual knowledge of ML-LMs across multiple languages. Our objective was to clarify how facts are perceived in different languages and to identify the difference in fact recognition among languages. Furthermore, we investigated how ML-LMs learn and represent these facts, seeking to understand the interplay between languages in the context of fact recognition.

3.1 Experiment setup

Datasets For the factual probing experiments, we used the mLAMA dataset[2] (Kassner et al., 2021). This dataset is a multilingual extension of LAMA (Petroni et al., 2019) and draws from sources such as TREx (Elsahar et al., 2018) and GoogleRE,[3] both of which extract information from Wikipedia. The mLAMA dataset contains 37,498 instances spanning 43 relations, represented as a fill-in-the-blank cloze, *e.g.*, "[X] plays [Y] music." where subject entity X, a relation, and object entity Y form a triplet (subject, relation, object).

Models We here focus on probing multilingual factual knowledge using encoder-based ML-LMs, multilingual BERT (mBERT)[4] (Devlin et al., 2019) and XLM-R[5] (Conneau et al., 2020a). Encoder-based models are chosen over generative models like mT5 (Xue et al., 2021) and BLOOM (Scao et al., 2022) since they are smaller than the generative models but exhibit excellent performance on language understanding tasks. Specifically for our factual knowledge probing task, which employs fill-in-the-blank queries, the encoder-based models perform well at referencing and integrating information across entire sentences, ensuring a detailed contextual understanding.

3.2 Evaluation

To determine if ML-LMs can capture specific knowledge, we substitute X with the subject tokens and replace Y with mask tokens in each cloze template to form a cloze query (*e.g.*, "The Beatles play [MASK] music."). Then, we instructed the ML-LMs to predict the mask tokens. If, in this instance, it predicted the mask token to be "rock," we considered that the knowledge was captured by the ML-LMs.

Since the object cannot necessarily be tokenized as a single token, we need to determine the number of mask tokens needed for each probed fact. Previously proposed methods used an automated technique for determining the mask counts that maximized the probability of a correct number of mask tokens (Jiang et al., 2020a; Kassner et al., 2021). In contrast, our study aimed at investigating fact representations rather than simply evaluating the probing performance of ML-LMs. We therefore adopt more lenient methods of probing facts as follows.

Protocol To correctly estimate predictable facts, we evaluated two matching methods: full-match and partial-match. In the full-match approach, we assigned the exact number of mask tokens corresponding to the object. However, this method sometimes produced correct answers containing non-essential tokens such as whitespaces. We considered these cases not as errors but as potentially valid answers. Consequently, we also examined the partial-match method. For each query template such as "[X] plays [Y] music," we listed all objects Y and their token counts associated with the template. We then probed the two ML-LMs (mBERT and XLM-R) with multiple queries, ranging from one (*e.g.*, "The Beatles plays [MASK]

[2]While DLama-v1 (Keleg and Magdy, 2023), a variant of mLAMA designed to address cultural biases, is available, we opted for mLAMA, because our focus was on cross-lingual features rather than solely assessing model competencies in factual understanding. mLAMA is suitable for this objective as it offers a consistent query set across all languages, ensuring clarity and precision in our investigation.

[3]https://github.com/google-research-datasets/relation-extraction-corpus

[4]https://huggingface.co/bert-base-multilingual-cased

[5]https://huggingface.co/FacebookAI/xlm-roberta-large

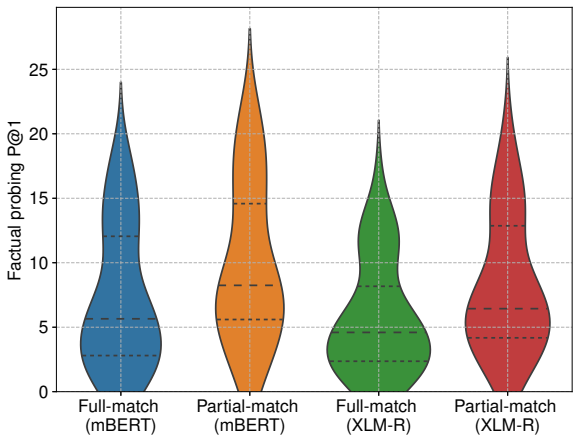

Figure 2: Probing P@1 on mLAMA for full- and partial-match methods with mBERT and XLM-R.

music") up to the longest mask token sequence for the corresponding template (*e.g.*, "The Beatles plays [MASK] [MASK] [MASK] [MASK] music"). A fact was considered correctly predicted if any version of the prompt included the correct object tokens, regardless of additional preceding or succeeding tokens.

Results Figure 2 displays the results in terms of the first rank precision (P@1). Across all experiments, we noted a consistently low P@1 score, especially for the majority of low-resource languages. Refer to Table 4 in Appendix A for details.

Interestingly, the partial-match method demonstrated noticeably better factual probing by considering partially matched predictions. A deeper analysis revealed four unique prediction patterns, specifically discernible for the partial-match method, as examplified in Table 1. These patterns illustrate the limitation of the factual probing based on the fill-in-the-blank dataset: the answers are restricted to a single standard format and thus do not reflect the diversity in entity expressions in text. These observations indicate a direction for future improvements in probing techniques.

For clarity in our subsequent analysis, we will primarily focus on mBERT, a 12-layer Transformer-based ML-LM pre-trained on Wikipedia text across 103 languages. This decision is motivated by the comparable results between mBERT and XLM-R. Although the partial-match method offers a richer representation for exploration, it sometimes includes irrelevant tokens that can introduce noise (*e.g.*, whitespace in Table 1). Therefore, the following discussions are predominantly based on results obtained using the full-match approach.

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.

Table 1: Four patterns discerned in facts predicted by partial-match method. The tokens in "()" are extra compared with those in the ground-truth dataset.

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

Table 2: Correlation between the training data volume and probing P@1 on mLAMA with mBERT.

4 What Factors Influence Discrepancy in Factual Probing across Languages?

The vertical bars in Figure 3 showing the results of factual probing in various languagesreveal substantial differences among languages. In this section, we will evaluate the potential factors for these differences and examine how they relate to proficiency in the cross-lingual transfer ability of ML-LMs.

4.1 Training data volume for learning facts

The first factor relates to the amount of distinct factual knowledge seen in training the ML-LMs. Since it is difficult to estimate the amount of factual knowledge in the training data, we explored several metrics on the training data volume instead. Specifically, we calculated the Pearson correlation coefficient between probing accuracy (P@1) and five metrics on the training data of mBERT, Wikipedia:[6] the number of Wikipedia articles and raw and compressed data sizes for abstracts and full articles.

Table 2 lists the correlation between P@1 and the metrics on the training data volume. All metrics show a moderate correlation with P@1. We depict the data size of abstracts, which correlated the most among the five metrics, in Figure 3. The moderate correlation indicates a limited impact of the training data volume on learning factual knowledge.

Among high-resource languages, we observed the probing P@1 of 16.94% for Italian (it) and 1.34% for Japanese (ja), as shown in Table 3. Prior research has highlighted potential cultural biases

[6]We crawled Wikipedia dumps as of the time just before the released date of mBERT. See Appendix B.1 for details.

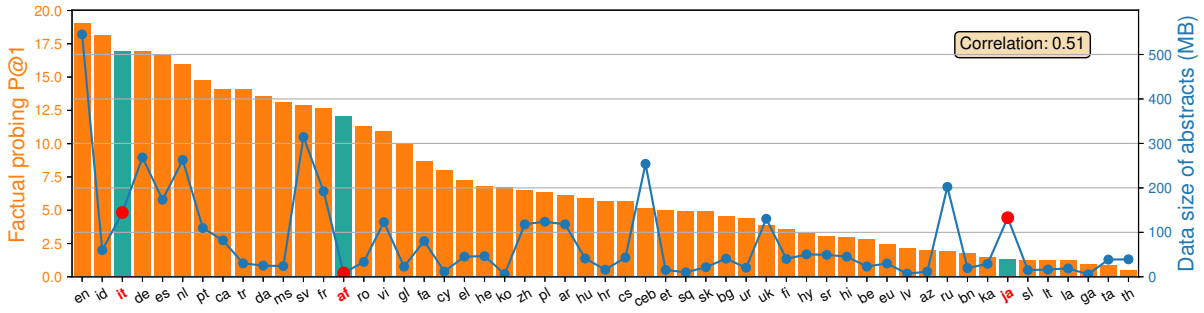

Figure 3: Wikipedia data size of abstracts vs. Factual probing P@1 on mLAMA in mBERT in 53 languages.

	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

Table 3: P@1 and one-token object counts for mBERT and XLM-R in Italian (it), Japanese (ja) Afrikaans (af).

in mLAMA, particularly affecting non-Latin script languages (Keleg and Magdy, 2023). However, these biases alone do not explain the substantial difference between the training data volume and probing performance. Meanwhile, some low-resource languages, such as Afrikaans (af), perform relatively well despite having limited Wikipedia data. The ability of Afrikaans (af) to represent such a breadth of knowledge, even in the face of potential cultural biases, is indeed remarkable.

4.2 Mask token count making inference hard

There was notable -0.81 (mBERT) and -0.74 (XLM-R) correlations between P@1 and the number of subwords in the target entities. As shown in Table 3, while both ML-LMs had similar P@1 scores in predicting one-token entities, the XLM-R tokenizer captured more one-token entities in Japanese (ja), resulting in more accurate predictions. The XLM-R tokenizer often produced shorter tokens for non-Latin scripts, enhancing its performance for non-Latin languages. However, this does not explain the differences in prediction accuracy across languages, as Afrikaans (af) outperformed Japanese (ja) for one-token P@1.

4.3 Presence of localized knowledge cluster

The higher accuracy for low-resource languages might have resulted from the model being proficient at cross-lingual factual knowledge sharing. To investigate this possibility, we assessed shared facts between languages using Jaccard similarity:

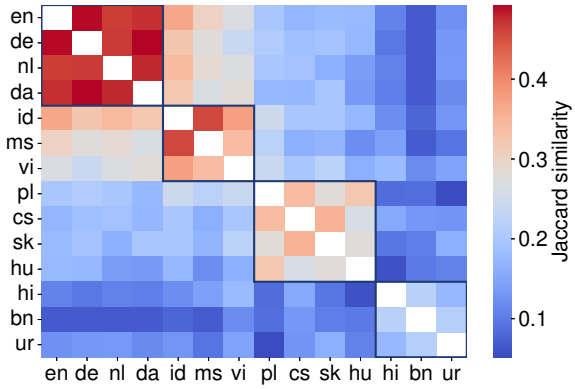

Figure 4: Jaccard similarity matrix of shared factual knowledge across languages with mBERT.

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

where A and B are sets of facts predictable by two languages.

Figure 4 reveals that languages in geographical proximity showed greater overlap in shared facts. Geographically proximate languages, like Indonesian (id), Malay (ms), and Vietnamese (vi), had higher similarities, indicating substantial shared content. This suggests that cross-lingual knowledge transfer does not occur universally across languages. Instead, it appears to be localized, driven more by shared culture and vocabulary. We will explore this phenomenon in the subsequent sections.

Summary We examined training data volume and mask token count as factors that will influence the discrepancies in factual knowledge comprehension across languages. Our findings revealed localized knowledge sharing patterns among languages, hinting at the potential for cross-lingual transfer.

5 Do ML-LMs Have Fact Representations Shared across Languages?

In this section, we discuss how ML-LMs represent facts within their parameter spaces by exploring two scenarios. In one scenario, a copy of the same fact is independently maintained in different languages, as illustrated in Figure 1(a); ML-LMs based on this scenario are referred to as "language-independent." In the other scenario, fact representations in different languages are unified in an embedding space, as illustrated in Figure 1(b,c); ML-LMs based on this scenario are referred to as "cross-lingual." The language-independent scenario will hinder cross-lingual transfer in fine-tuning ML-LMs on downstream tasks, where the training data is available in a few languages.

5.1 Factual neuron probing

In Transformer-based PLMs, the feed-forward network (FFN) plays a pivotal role in the knowledge extraction and representation process (Durrani et al., 2020; Dai et al., 2022). Formally, an FFN is defined as:

$$\text{FFN}(x) = \text{f}(x\mathbb{K}^T + b_1)\mathbb{V} + b_2 \qquad (2)$$

where $\mathbb{K}$, $\mathbb{V}$, b_1, and b_2 are trainable parameters.

Experiment setup We analyzed the representation of cross-lingual facts in ML-LMs by identifying their active neurons across languages. We used a method called PROBELESS (Antverg and Belinkov, 2022) - an efficient and explicit technique that measures neuron activity by contrasting value differences among facts. Specifically, PROBELESS identifies neurons as active when their values deviate greatly from the average for specific knowledge representations.

More specifically, we analyzed neuron activity for each correctly-predicted fact, represented as (subject, relation, object). For probing, we considered other predictable facts that share the same relation but vary in subject-object pairs. We collected the neurons of the mask tokens and identified their active neurons as signatures of the fact representations. For multi-token masks, we used average pooling across all tokens. As our goal was to investigate fact representations across languages, we collected the active neurons for the same fact in various languages for further analysis. Because the reliability of fact probing is lower when the availability of predicted facts is limited, we focused on the top 30 languages by P@1 score.

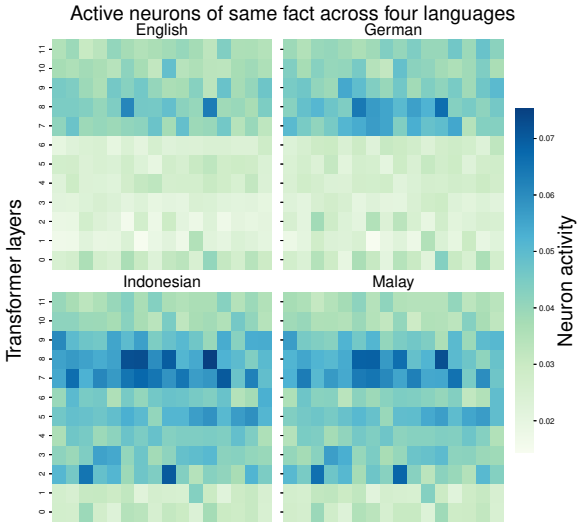

Figure 5: Neuron activity with mBERT in four languages, English, German, Indonesian, and Malay, in response to the query "William Pitt the Younger used to work in [MASK]." Color intensity indicates neuron activity; neurons in each Transformer layer are grouped into 16 bins. Distinct activation patterns in the English-German and Indonesian-Malay pairs indicate cross-lingual knowledge neurons, while differences between the pairs indicate language-independent representations.

5.2 Results and discussion

Do cross-lingual fact representations exist? In our neuron probing, we identified and used active neurons to distinguish cross-lingual fact representations from language-independent neurons. Similar patterns in active neurons across languages suggest that there is a common cross-lingual semantic space for fact representation. Our findings indicate that while some languages exhibit similar neuron activity patterns for a given fact, others exhibit distinct distributions, as depicted in Figure 5. This indicates the presence of both language-independent and cross-lingual fact representations in ML-LMs, even for the same fact.

Quantification of cross-lingual sharing To precisely measure the extent of cross-lingual sharing of facts between two languages, we calculated Jaccard similarity based on the top 50 active factual neurons. We then measured the general language similarity among all languages by computing the average similarity for all shared facts.

Figure 6 shows the results of computing general language similarity in terms of shared facts. Surprisingly, our findings revealed no consistent geographical boundaries among languages, suggesting

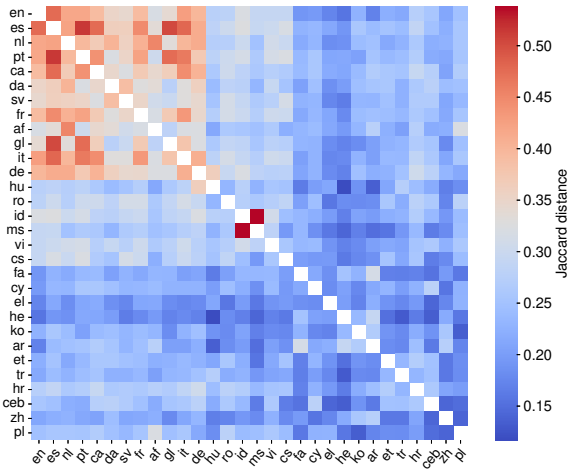

Figure 6: Language similarity based on top 50 shared active neurons by probing on mLAMA with mBERT.

that the use of either the language-independent scenario or the cross-lingual sharing scenario largely depends on the specific factual knowledge itself, so such analysis should be tailored to specific factual knowledge. For instance, despite English (en) and Chinese (zh) exhibiting a relatively low neuron correlation (0.21, compared with the 0.24 average), they still displayed similar active neuron patterns for certain facts, often rooted in shared tokens, like "Google" in Chinese "developed-by" relations.

The results of neuron probing revealed that active fact neurons in low-resource languages have more activity and are distributed more in the shallow layers of Transformers compared with high-resource languages. This finding contradicts the findings of previous research (Oba et al., 2021; Dai et al., 2022), and suggests that only a few neurons in the higher Transformer layers are responsible for representing facts. This difference indicates a potential reason for the lower expression ability of low-resource languages, for which the hierarchical structure of knowledge is not acquired as well as in resource-rich languages.

Summary Our exploratory analysis using neuron probing of fact representation in ML-LMs and examination of whether languages share common representations or maintain unique knowledge spaces for specific facts revealed the presence of both language-independent and cross-lingual neural activity patterns across languages. The results of Jaccard similarity analysis of active factual neurons revealed inconsistent geographical boundaries in knowledge sharing, indicating the complexity of cross-lingual knowledge representation.

6 How Are Cross-lingual Representations of Facts Formed in ML-LMs?

Having confirmed the presence of cross-lingual representations, we subsequently explored their formation within ML-LMs and assessed whether they are learned individually from distinct language corpora and subsequently aligned into a common semantic space (Figure 1(b)) or whether they are acquired through cross-lingual transfer (Figure 1(c)).

6.1 Tracing the roots of facts back to data

To identify the reason behind the formation of a cross-lingual representation, it is crucial to verify if the fact originates from the training data. We used a simple yet effective method to check the presence of a fact in text: for a fact triplet (subject, relation, object), we examined the occurrences of the subject and object in mBERT's training data, Wikipedia.[6] If they could be found, the fact was considered present in the data. Although this approach may not provide precise quantitative results, it helps in exploring cross-lingual transfer possibilities.

To determine whether a fact was traced back to the data, we used subject-object co-occurrence as an approximation method. We rigorously adhered to the preprocessing and sentence-splitting guidelines for mBERT (Devlin et al., 2019). Using the WikiExtractor,[7] we extracted only text passages, deliberately omitting lists, tables, and headers. Each extracted document was segmented into multiple lines, with each line containing no more than 512[8] tokens. Using string matching between the object-subject pair and Wikipedia text, we assessd the co-occurrence of the object and subject for a given fact. If there was co-occurrence, we considered the fact to be present; otherwise, it was considered to be absent.

6.2 Analysis of absent yet predictable facts

We assessed the absence rate of all and correctly predicted facts, respectively. As shown by the results for 53 languages in Figure 7, languages with more training data exhibited better factual knowledge coverage, as anticipated. Nonetheless, several facts, such as those in Afrikaans (af) and Albanian (sq), were accurately predicted even without verifiable existence in the training corpus, indicating a high possibility of effective cross-lingual transfer.

[7] https://github.com/attardi/wikiextractor

[8] The maximum number of tokens that can be input to mBERT in training.

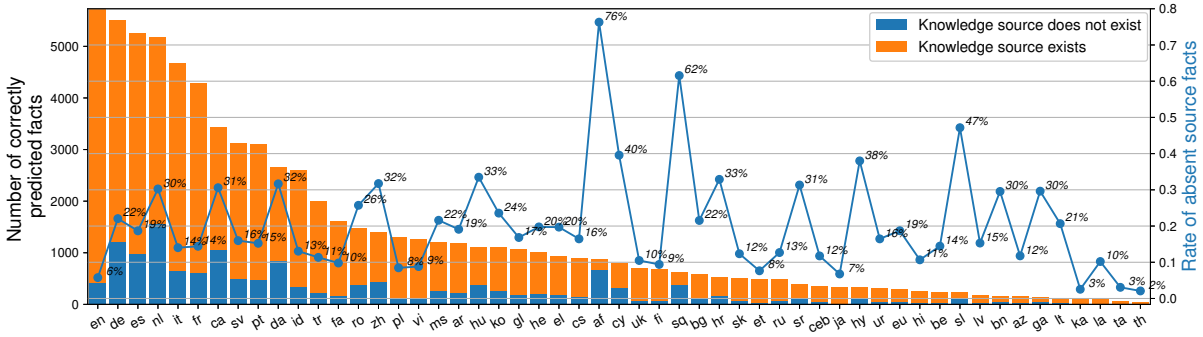

Figure 7: Number of correctly-predicted facts with mBERT in terms of existence of knowledge source.

What kinds of facts are absent yet predictable? Analysis revealed that many of the facts that were absent in the knowledge source but correctly predicted were relatively easy to predict. We categorized these easy-to-predict facts into two types: shared entity tokens and naming cues. Along with other facts, we grouped them into a total of three categories by using a rule-based method (see to Appendix B.2 for the criteria of fact classification).

Shared entity tokens: Some probing queries ask object entities whose tokens are shared with the subject entities; for example, 'Sega Sports R&D is owned by Sega.' We regard correctly predicted facts to be of this type when the subject and object entities share subwords.

Naming cues: Some probing queries are related to entity-universal association across person names, countries, and languages (see Table 6 in Appendix for details), which allows the ML-LMs to guess the object entity from subwords of the subject entity; for example, 'The native language of Go Hyeon-jeong is Korean.' We regard facts correctly-predicted on the basis of such a relation to be of this type.

Others: The remaining facts are difficult to infer from the entities only, indicating the high possibility of cross-lingual transfer. *e.g.*, 'Crime & Punishment originally aired on NBC.'

Figure 8 shows the proportions of facts correctly predicted without knowledge sources by mBERT for the three types. The predictability of easy-to-predict facts suggests that the ML-LMs can rely on simple deductions rather than encoding specific facts to make predictions, highlighting the need to enhance probing datasets to enable more effective evaluation of model proficiency in fact representation. Without the easy-to-predict facts, the absence

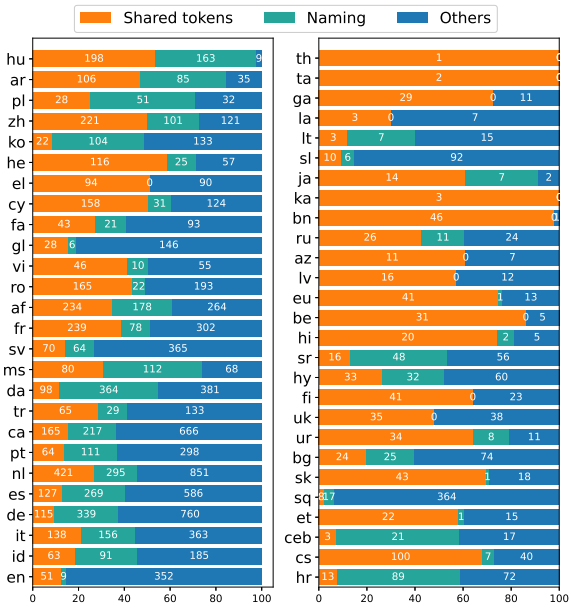

Figure 8: The count of three types of absent and predictable facts with mBERT.

rate drops but is still not zero (blue bar in Figure 8) for some of the languages, such as Albanian (sq), Slovenian (sl), and Galician (gl), indicating that ML-LMs indeed possess cross-lingual transfer capabilities for factual knowledge, even though the knowledge sources for some languages are limited. See Table 7 through 9 in Appendix for examples of correctly-predicted facts of the three types.

Summary We trace the roots of facts back to the mBERT's pre-training data, specifically the Wikipedia text. We categorized correctly predicted but absent facts into three types, two of which can be resolved through simple inference. Our statistics show that while cross-lingual transfer of factual knowledge in ML-LMs does occur, it is limited, highlighting the challenges in achieving effective cross-lingual factual knowledge transfer.

7 Conclusions

Our research establishes the groundwork for further studies in understanding cross-lingual factual knowledge representation. Through comprehensive factual probing experiments and analysis for 53 languages using the mBERT multilingual language model, we evaluated key factors in the differences between their proficiencies in cross-lingual transfer of factual knowledge, such as the training data volume and mask token count, and identified knowledge sharing patterns among geographically proximate language clusters.

We leverage the existing neuron probing and the proposed knowledge tracing methods to identify three types of patterns for acquiring and representing factual knowledge across languages in ML-LMs: language-independent, cross-lingual shared, and cross-lingual transferred. Analysis revealed the challenges involved in achieving effective cross-lingual transfer of factual knowledge from high-resource to low-resource languages in ML-LMs.

Future work aims to enhance the cross-lingual fact representation learning in ML-LMs and develop a more precise factual probing dataset.

8 Limitations

We primarily examined two encoder-based models for language understanding tasks, mBERT and XLM-R. Therefore, our findings may not directly apply to the recent, large decoder-based multilingual language models such as BLOOM (Scao et al., 2022). Future research is needed to explore these larger generative models in order to gain more insights into the mechanism of cross-lingual knowledge transfer in ML-LMs.

Moreover, the dataset we used has certain limitations. A review of the relation template in mLAMA by one first author, who is a native Chinese speaker, identified necessary corrections for certain Chinese language prompts. Meanwhile, the dataset focuses on a limited set of relation types, indicating that fact prediction in other relations may lie beyond the scope of our current research.

9 Ethical Statement

All datasets used in our experiment are publicly accessible and do not contain sensitive information.[9] The findings and interpretations presented are unbiased and intended for academic purposes.

[9] https://en.wikipedia.org/wiki/Wikipedia:What_Wikipedia_is_not

Acknowledgements

This work was partially supported by the special fund of Institute of Industrial Science, The University of Tokyo, by JSPS KAKENHI Grant Number JP21H03494, JP21H03445, and by JST, CREST Grant Number JPMJCR19A, Japan.

References

Omer Antverg and Yonatan Belinkov. 2022. On the pitfalls of analyzing individual neurons in language models. In *International Conference on Learning Representations*.

Tyler Chang, Zhuowen Tu, and Benjamin Bergen. 2022. The geometry of multilingual language model representations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 119–136, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ethan A. Chi, John Hewitt, and Christopher D. Manning. 2020. Finding universal grammatical relations in multilingual BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5577, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020a. 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, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online. Association for Computational Linguistics.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502, Dublin, Ireland. Association for Computational Linguistics.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6491–6506, Online and Punta Cana, Dominican Republic. 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.

Nadir Durrani, Hassan Sajjad, Fahim Dalvi, and Yonatan Belinkov. 2020. Analyzing individual neurons in pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4865–4880, Online. Association for Computational Linguistics.

Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest, and Elena Simperl. 2018. T-REx: A large scale alignment of natural language with knowledge base triples. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Constanza Fierro and Anders Søgaard. 2022. Factual consistency of multilingual pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3046–3052, Dublin, Ireland. Association for Computational Linguistics.

Negar Foroutan, Mohammadreza Banaei, Rémi Lebret, Antoine Bosselut, and Karl Aberer. 2022. Discovering language-neutral sub-networks in multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7560–7575, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Benjamin Heinzerling and Kentaro Inui. 2021. Language models as knowledge bases: On entity representations, storage capacity, and paraphrased queries. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1772–1791, Online. Association for Computational Linguistics.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12365–12394, Singapore. Association for Computational Linguistics.

Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. 2020a. X-FACTR: Multilingual factual knowledge retrieval from pretrained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5943–5959, Online. Association for Computational Linguistics.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020b. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual bert: An empirical study. In *The Eighth International Conference on Learning Representations*.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. Multilingual LAMA: Investigating knowledge in multilingual pretrained language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3250–3258, Online. Association for Computational Linguistics.

Amr Keleg and Walid Magdy. 2023. DLAMA: A framework for curating culturally diverse facts for probing the knowledge of pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6245–6266, Toronto, Canada. Association for Computational Linguistics.

Nayeon Lee, Belinda Z. Li, Sinong Wang, Wen-tau Yih, Hao Ma, and Madian Khabsa. 2020. Language models as fact checkers? In *Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER)*, pages 36–41, Online. Association for Computational Linguistics.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. 2022. Fast model editing at scale. In *The Tenth International Conference on Learning Representations*.

Ibraheem Muhammad Moosa, Mahmud Elahi Akhter, and Ashfia Binte Habib. 2023. Does transliteration help multilingual language modeling? In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 670–685, Dubrovnik, Croatia. Association for Computational Linguistics.

Benjamin Muller, Yanai Elazar, Benoît Sagot, and Djamé Seddah. 2021. First align, then predict: Understanding the cross-lingual ability of multilingual BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2214–2231, Online. Association for Computational Linguistics.

Daisuke Oba, Naoki Yoshinaga, and Masashi Toyoda. 2021. Exploratory model analysis using data-driven neuron representations. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 518–528, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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)*, pages 2463–2473, Hong Kong, China. 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.

Machel Reid and Mikel Artetxe. 2023. On the role of parallel data in cross-lingual transfer learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5999–6006, Toronto, Canada. Association for Computational Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Thomas Wolf, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Severine Verlinden, Klim Zaporojets, Johannes Deleu, Thomas Demeester, and Chris Develder. 2021. Injecting knowledge base information into end-to-end joint entity and relation extraction and coreference resolution. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1952–1957, Online. Association for Computational Linguistics.

Xiaozhi Wang, Kaiyue Wen, Zhengyan Zhang, Lei Hou, Zhiyuan Liu, and Juanzi Li. 2022. Finding skill neurons in pre-trained transformer-based language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11132–11152, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. 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)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Da Yin, Hritik Bansal, Masoud Monajatipoor, Liunian Harold Li, and Kai-Wei Chang. 2022. GeoMLAMA: Geo-diverse commonsense probing on multilingual pre-trained language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2039–2055, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhengyan Zhang, Yankai Lin, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2022. MoEfication: Transformer feed-forward layers are mixtures of experts. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 877–890, Dublin, Ireland. Association for Computational Linguistics.

A Probing P@1 on mLAMA with mBERT and XLM-R

Table 4 lists the probing P@1 for the 53 languages on mLAMA using full-match and partial match methods with mBERT and XLM-R, respectively, to complement the overall results shown in § 3.2. In most of the languages, mBERT with the partial-match method achieved the best P@1. This is probably because the number of facts in Wikipedia used in mBERT will be larger than that in CC-100 used in XLM-R, although CC-100 is larger than Wikipedia (Conneau et al., 2020a). Meanwhile, XLM-R outperforms mBERT on languages with non-Latin scripts such as Hindi, Bangla, Georgian, Japanese, and Thai and several Eastern European languages such as Romanian, Hungarian, Bulgarian, Finnish, Azerbaijani, and Georgian. These results will be probably due to the larger mask token count in the non-Latin scripts (§ 4.2) and local knowledge clusters (§ 4.3) that do not include resource-rich languages.

B Experimental Details

B.1 Wikipedia dumps

The Wikipedia data used in this study for assessing the effect of training data volume on factual knowl-

ISO	Language	mBERT		XLM-R		ISO	Language	mBERT		XLM-R	
		Full	Partial	Full	Partial			Full	Partial	Full	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

Table 4: P@1 for 53 languages on mLAMA using full- and partial-match methods with mBERT and XLM-R.

Date	ISO-639 Language Codes
20181001	ru, el, uk, la
20181101	ms, ca, ko, he, fi, ga, ka, th, zh, eu, da, pt, fr, sr, et, sv, hy, cy, sq, hi, hr, bg, ta, sl, bn, id, be, ceb, fa, pl, az, ar, gl, lt, cs, sk, lv, tr, af, vi, ur, ro
20181120	en, nl, ja, it, es, hu, de

Table 5: Dates of downloaded Wikipedia dumps for the 53 languages supported by mLAMA.

edge acquisition (§ 4.1) and tracing the roots of facts (§ 6.1) were taken from the Wikipedia dumps in the Internet Archive.[10] We extracted articles in the Main and Article namepspace.[11]

We collected public Wikipedia dumps for 53 languages, spanning the period between October 1 and November 20, 2018. We chose this timeframe to align with mBERT's release date, ensuring that data source resembled the training data of mBERT. The download URLs for each language follow this format: `https://archive.org/download/{language_code}wiki-{timestamp}`; the dumps were downloaded on the basis of data availability during the target period. The datas of the downloaded dumps for each language are listed in Table 5.

B.2 Rules for classifying types of predictable facts

We classify the three types of predictable facts by the following rules.

Shared entity tokens: We normalized entities by lowercasing strings and unifying Chinese traditional/simplified characters, and then assessed if the object is a substring of or shares subwords with the subject. Examples of this type can be found in Table 7.

Naming cues: We manually selected several relations that contain information among person name, location, and countries entities, as illustrated in Table 6. Examples of this type can be found in Table 8.

Others: The facts other than those classified into shared tokens across entities and naming cues are regarded as others. Examples of this type can be found in Table 9.

[10] The Internet Archive (https://archive.org/) is a non-profit library of millions of free books, movies, software, music, websites, and more.

[11] https://en.m.wikipedia.org/wiki/Wikipedia:What_is_an_article%3F#Namespace

IDs	Relation	Examples
P103	The native language of [X] is [Y].	The native language of Jean-Baptiste Say is French. The native language of Nie Weiping is Chinese.
P17	[X] is located in [Y].	Noyon is located in France. Gavrilovo-Posadsky District is located in Russia.
P140	[X] is affiliated with the [Y] religion.	Abdullah Ahmad Badawi is affiliated with the Islam religion. Noriyasu Hirata is Japan citizen.
P1412	[X] used to communicate in [Y].	Pere Gimferrer used to communicate in Spanish. Susan McClary used to communicate in English.
P27	[X] is [Y] citizen.	Priyanka Vadra is India citizen. Giovanni Lista is Italy citizen.

Table 6: Relations containing mostly name, country, and location entities.

ISO	Language	Examples of absent yet predictable facts
af	Afrikaans	Vlag van Jamaika is 'n wettige term in Jamaika.
az	Azerbaijani	Split hava limanı Split adını daşıyır.
be	Belarusian	Сталіцай камуна Гётэбарг з'яўляецца Гётэбарг.
bg	Bulgarian	Декларация за създаване на държавата Израел е легален термин в Израел.
ca	Catalan	Govern de Macau és un terme legal en Macau.
ceb	Cebuano	Ang Nokia X gihimo ni Nokia.
cs	Czech	Guvernér Kalifornie je právní termín v Kalifornie.
cy	Welsh	Mae seicoleg cymdeithasol yn rhan o seicoleg.
da	Danish	Danmarks Justitsminister er en juridisk betegnelse i Danmark.
de	German	Die Hauptstadt von Gouvernorat Bagdad ist Bagdad.
el	Greek	Υπουργός Δικαιοσύνης της Δανίας είναι ένας νομικός όρος στο Δανία.
en	English	Sega Sports R&D is owned by Sega .
es	Spanish	Honda Express es producido por Honda.
et	Estonian	Seim (Poola) on Poola -is juriidiline termin.
eu	Basque	orbita ekliptiko orbita azpi-klasea da.
fi	Finnish	1955 Dodge tuottaa Dodge.
fr	French	Massacre de Cologne se trouve dans Cologne.
ga	Irish	Tá Contae Utah suite i Utah.
gl	Galician	Sheffield United F.C. recibe o nome de Sheffield.
hr	Croatian	Sjedište Valencia C.F. B je u Valencia.
hu	Hungarian	Honda Fit -et Honda állítja elő.
id	Indonesian	Menteri Kehakiman Denmark adalah istilah hukum dalam Denmark.
it	Italian	Nagoya Railroad Co., Ltd è stata fondata a Nagoya.
ja	Japanese	アンフィオン級潜水艦は潜水艦のサブクラスです。
ko	Korean	모빌군의 수도는 모빌입니다.
la	Latin	Ethica adhibita est pars ethica.
lt	Lithuanian	Stokholmas savivaldybė sostinė yra Stokholmas.
lv	Latvian	Voterfordas grāfiste galvaspilsēta ir Voterforda.
ms	Malay	Sony Alpha 99 dihasilkan oleh Sony.
nl	Dutch	Aluminiumsulfaat bestaat uit aluminium.
pl	Polish	Cadillac Series 60 jest wytwarzany przez Cadillac.
pt	Portuguese	cooperativa autogestionária é uma subclasse de cooperativa.
ro	Romanian	Festivalul Internațional de Film de la Calgary este localizat în Calgary.
ru	Russian	Сенат Теннесси является юридическим термином в Теннесси.
sk	Slovak	BMW N52 sa vyrába v BMW.
sl	Slovenian	Narodno gledališče München se nahaja v München.
sq	Albanian	BBC Music është pjesë e BBC.
sr	Serbian	Аеродром Минск је назван по Минск.
sv	Swedish	Huvudstaden till Guvernementet Bagdad är Bagdad.
th	Thai	เมืองหลวงของ เมืองหลวงคอร์ก คือ คอร์ก
tr	Turkish	Waterford County 'un başkenti Waterford' dir.
uk	Ukrainian	Законодавча асамблея штату Орегон - юридичний термін в Орегон.
vi	Vietnamese	Vốn của Hạt Waterford là Waterford.
zh	Chinese	意大利雜菜湯是汤的子类。

Table 7: Examples of easy-to-predict facts by using shared entity tokens in mLAMA.

ISO	Language	Examples of absent yet predictable facts
af	Afrikaans	Die moedertaal van Jean-Baptiste Say is Frans.
bg	Bulgarian	Родният език на Лионел Жоспен е френски език.
ca	Catalan	La llengua nativa de Alain Mabanckou és francès.
ceb	Cebuano	Ang Giovanni Lista usa ka lungsuranon sa Italya.
cs	Czech	Rodný jazyk Danielle Darrieuxová je francouzština.
cy	Welsh	Mae Guillaumes wedi'i leoli yn Ffrainc.
da	Danish	Mødesproget til Pierre Blanchar er fransk.
de	German	Die Muttersprache von Pierre Blanchar ist Französisch.
en	English	The native language of Hamidou Benmassoud is French .
es	Spanish	Bruno Racine solía comunicarse en francés.
et	Estonian	Dominic Seiterle on Kanada kodanik.
eu	Basque	Umar II.a Islam erlijioarekin erlazionatuta dago.
fr	French	Bayazid Bastami est affilié à la religion islam.
gl	Galician	Toulouges está situado en Francia.
hr	Croatian	Izvorni jezik Jean-Baptiste Say je francuski jezik.
hu	Hungarian	John Hutton az angol nyelven történő kommunikációhoz használt.
id	Indonesian	Adrian Knox adalah warga negara Australia.
it	Italian	La lingua madre di Victor Riqueti de Mirabeau è francese.
ja	Japanese	ウィリアム・ハウイットの母国語は英語です。
ko	Korean	알랭 마방쿠의 모국어는 프랑스어입니다.
lt	Lithuanian	Gimtoji kalba Nikolajus Dobroliubovas yra rusų kalba.
ms	Malay	Bahasa ibunda Jean-Baptiste Say ialah Bahasa Perancis.
nl	Dutch	De moedertaal van Jacques Legras is Frans.
pl	Polish	Abdolkarim Soroush jest powiązany z religią islam.
pt	Portuguese	O idioma nativo de Georges Hugnet é francês.
ro	Romanian	Abdolkarim Soroush este afiliat cu religia islam.
ru	Russian	Насир уд-Дин Абу-л-Фатх Мухаммад связан с религией ислам.
sk	Slovak	Rodný jazyk Vergílius je latinčina.
sl	Slovenian	Ernesto Tornquist je državljan Argentina.
sq	Albanian	Gjuha amtare e Andrew Jackson është anglisht.
sr	Serbian	Изворни језик Жан Батист Сеј је француски језик.
sv	Swedish	Sibirkhanatet är anslutet till islam -religionen.
tr	Turkish	Guillaumes, Fransa 'da bulunur.
vi	Vietnamese	Uzhhorod và Moskva là hai thành phố sinh đôi.
zh	Chinese	円珍,隸屬於佛教宗教。

Table 8: Examples of easy-to-predict facts by using naming cues in mLAMA.

ISO	Language	Examples of absent yet predictable facts
af	Afrikaans	Die hoofstad van Verenigde Koninkryk is Londen.
az	Azerbaijani	Slovakiya Sosialist Respublikası -nin paytaxtı Bratislava.
be	Belarusian	Сталіцай Татарская АССР з'яўляецца Казань.
bg	Bulgarian	Ембриология е част от медицина.
ca	Catalan	Jean-Baptiste-Claude Chatelain va néixer a París.
ceb	Cebuano	Kuala Lumpur (estado) mao ang kapital sa Malaysia.
cs	Czech	Beijing College Student Film Festival se nachází v Peking.
cy	Welsh	Mae Meade Lux Lewis yn chwarae piano.
da	Danish	Jean-Baptiste-Claude Chatelain blev født i Paris.
de	German	Surinder Khanna wurde in Delhi geboren.
el	Greek	Πιέρ Λεκόμτ ντου Νουί γεννήθηκε στο Παρίσι.
en	English	Aleksandar Novaković was born in Belgrade .
es	Spanish	Aleksandar Novaković nació en Belgrado.
et	Estonian	Serbia kuningriik pealinn on Belgrad.
eu	Basque	Libano Mendiko eskualdea hiriburua Beirut da.
fi	Finnish	Art Davis soittaa jazz -musiikkia.
fr	French	Rhigos est un village.
ga	Irish	Is é Toulouse príomhchathair Haute-Garonne.
gl	Galician	Giuliano Giannichedda xoga na posición centrocampista.
hr	Croatian	Glavni grad Narodna Socijalistička Republika Albanija je Tirana.
hu	Hungarian	State University of New York székhelye Albany -ben található.
id	Indonesian	Ibukota Republik Rakyat Sosialis Albania adalah Tirana.
it	Italian	Vernon Carroll Porter è nato a Cleveland.
ko	Korean	머피 브라운는 원래 CBS에 방영되었습니다.
la	Latin	Gulielmus Marx Est politicus per professionis.
lt	Lithuanian	Ernst & Young būstinė yra Londonas.
lv	Latvian	Itālijas futbola izlase ir loceklis no FIFA.
ms	Malay	Power Rangers Samurai pada mulanya ditayangkan pada Nickelodeon.
nl	Dutch	Power Rangers: Samurai werd oorspronkelijk uitgezonden op Nickelodeon.
pl	Polish	Gregg Edelman to aktor z zawodu.
pt	Portuguese	Jean-Baptiste-Claude Chatelain nasceu em Paris.
ro	Romanian	Capitala lui Republica Populară Socialistă Albania este Tirana.
ru	Russian	Штаб-квартира Jim Beam находится в Чикаго.
sk	Slovak	Leicestershire zdieľa hranicu s Lincolnshire.
sl	Slovenian	Dilawar Hussain se je rodil v Lahore.
sq	Albanian	Guy Doleman është një aktor me profesion.
sr	Serbian	Сједиште компаније Чикашка берза је у Чикаго.
sv	Swedish	Jean-Baptiste-Claude Chatelain föddes i Paris.
tr	Turkish	Aruba Futbol Federasyonu, FIFA üyesidir.
uk	Ukrainian	Штаб-квартира Партія «Новий Азербайджан» знаходиться в Баку.
vi	Vietnamese	Chiếc giày vàng Giải bóng đá Ngoại hạng Anh là một giải thưởng.
zh	Chinese	拉尔克·沃里斯是专业上的演員。

Table 9: Examples of non-easy-to-predict facts in mLAMA.