Unraveling Cross-Lingual Dynamics in Language Models: Independent, Shared and Transferred Factual Knowledge

Anonymous ACL submission

Abstract

Acquiring factual knowledge for low-resource languages within multilingual language models (ML-LMs) presents a significant challenge due to the low coverage of real-world entities in the training data. It underscores the need for transferring knowledge from resource-rich languages to resource-poor languages, namely cross-lingual transfer. However, the effectiveness and extent of cross-lingual transfer in ML-LMs for factual knowledge remain largely unexplored. To address this research gap, we use evaluation results from the multilingual factual knowledge probing dataset - mLAMA, to conduct a neuron-level inspection of how ML-LMs (here, multilingual BERT (mBERT)) represent facts in different languages. Additionally, we analyze the knowledge source (Wikipedia) to identify the various ways in which the ML-LMs learn specific facts. As a result, we identify three types of knowledge learning and representation patterns in the ML-LMs: language-independent, cross-lingual shared, and transferred, and introduce methods to differentiate them. The findings highlight the challenge of maintaining consistent factual knowledge across various languages and emphasize the need for further research to drive improvement.

1 Introduction

To address the inherent data sparseness in low-resource languages, multi-lingual language models (ML-LMs) such as mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020a) and mT5 (Xue et al., 2021) have been developed to transfer knowledge across languages, namely cross-lingual transfer. The effectiveness of cross-lingual transfer has been confirmed in ML-LMs on various foundational linguistic tasks, such as part-of-speech tagging, dependency parsing, and named-entity recognition (Wu and Dredze, 2019; Chi et al., 2020; Pires et al., 2019). Nonetheless, a more challenging task is the cross-lingual transfer of specific real-world

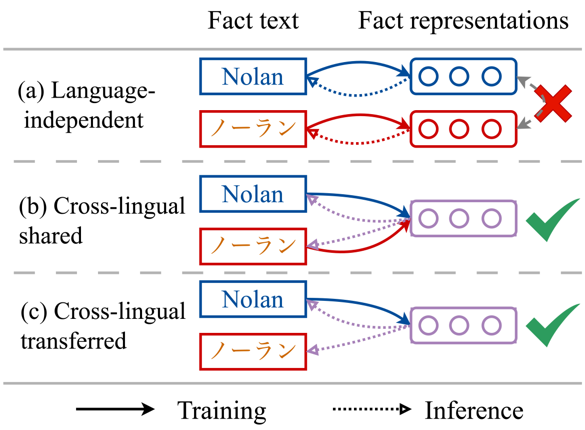

Figure 1: Three types of fact representation in ML-LMs for “Interstellar is directed by [MASK]” (“インターステラーの監督は[MASK]です “ in Japanese).

entity knowledge, for instance, understanding that “Interstellar is directed by Christopher Nolan.” In many low-resource languages, data about such entities might be minimal or non-existent. Effectively transferring knowledge is extremely important for applications that require accurate factual information, such as fact verification (Lee et al., 2020) and relation extraction (Verlinden et al., 2021).

Following early studies (Petroni et al., 2019; Jiang et al., 2020b) that investigate the capability of storing factual knowledge in pre-trained language models, several researchers have probed ML-LMs with the cloze-style queries to check whether they can recall real-world facts (Jiang et al., 2020a; Kassner et al., 2021; Yin et al., 2022; Fierro and Søgaard, 2022; Keleg and Magdy, 2023). The results of their research indicate that pre-trained language models demonstrate a competitive capacity for retrieving factual information. However, the mechanism behind the acquisition and representation of facts in ML-LMs remains unclear.

In this study, we investigate whether and how low-resource languages can benefit from the cross-lingual transfer of factual knowledge (Figure 1).

By probing deeper into this, we hope to provide background knowledge for enhancing ML-LMs in factual representation. To systematically address this, we segment our inquiry into three key research questions, each leading to a dedicated experiment:

Q1: How do factual probing performances differ across languages in ML-LMs, and what factors influence these differences? (§4)

Q2: Do languages represent specific facts within a shared semantic space, or do they maintain distinct knowledge representation? (§5)

Q3: What mechanisms during the pre-training of ML-LMs influence the formation of cross-lingual factual representations? (§6)

To address our queries, we start by probing ML-LMs (specifically, mBERT and XLM-R) using the factual knowledge probing dataset, mLAMA (Kassner et al., 2021). The results indicate that ML-LMs face difficulties in recognizing facts in low-resource languages, such as Thai, which aligns with previous research findings (Kassner et al., 2021) (§3). However, we observe only a weak correlation between probing performance and the amount of training data. Although the cultural bias of mLAMA may hinder probing performance in non-Latin script languages (Keleg and Magdy, 2023), the exact influence of models' cross-lingual capabilities remains to be established.

To discern the role of cross-lingual capability in fact probing, we perform a neuron-level analysis for facts predicted correctly. By comparing active neurons across languages, we observed that identical facts in various languages are not processed identically. For specific facts, some languages may exhibit similar neuron activity, while others display distinct patterns. We categorize the former as cross-lingual shared representations and the latter as language-independent representations.

To further identify the origins of cross-lingual shared representations, we propose a novel method by checking the presence of specific facts in the knowledge source (Wikipedia for mBERT). We assume the facts that are predicted correctly while absent in the training has high probability to be learnt by cross-lingual transfer in ML-LMs, which we refer to as cross-lingual transferred representation. However, the results reveal that, although such facts do emerge, few are acquired via cross-lingual transfer. This underscores the current limitations of ML-LMs in cross-lingual fact representation.

Contributions In this paper, We deeply investigate how ML-LMs capture and represent factual knowledge across diverse languages. Our findings highlight that ML-LMs differentiate factual knowledge through several methods, namely: language-independent, cross-lingual shared, and cross-lingual transferred representations (Figure 1). Moreover, we introduce analytical techniques to discern among these representations, elaborated in the subsequent sections.

2 Related Work

This section reviews two main areas of research in language models: cross-lingual transfer mechanisms and factual knowledge probing. We first discuss key studies that investigate how knowledge transfers across languages in ML-LMs. Next, we highlight research on how factual knowledge is perceived in language models.

2.1 Understanding cross-lingual transfer in ML-LMs

Numerous research have investigated the foundational mechanisms of cross-lingual transfer in ML-LMs. As for the acquisition of cross-lingual transfer, serveral studies ascertain that while shared tokens facilitate knowledge transfer, their impact remains circumscribed (K et al., 2020; Conneau et al., 2020b). Further research highlights the help of using parallel data in enhancing model's cross-lingual competency (Moosa et al., 2023; Reid and Artetxe, 2023).

Concurrently, another body of work direct its inquiry towards 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). It was discerned that ML-LMs incorporate a blend of language-specific and language-agnostic parameter spaces when representing identical knowledge across diverse languages. While previous research primarily offers a generalized overview of cross-lingual transfer mechanisms, they often neglect the nuanced variations in how ML-LMs represent and learn different knowledge.

2.2 Factual knowledge probing

Understanding factual representation in language models has gained great attention recently. Using fill-in-the-blank cloze question datasets, various studies (Petroni et al., 2019; Heinzerling and Inui, 2021; Wang et al., 2022) have explored the proficiency of language models in handling factual knowledge within the English language. Regarding the mechanism by which Transformer-based language models represent facts, several works (Geva et al., 2021; Dai et al., 2022) have conduct neuron-level investigation. These studies reveal that specific factual insights are linked to a select 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).

In a multilingual setting, several studies have investigated the capacity for factual representation in non-English languages (Jiang et al., 2020a; Kassner et al., 2021; Fierro and Søgaard, 2022). These studies suggest that the ability to perceive factual knowledge is not exclusive to English. Other languages can demonstrate comparable proficiency. However, a decline in the predictability of factual knowledge has been observed for languages with limited resources. Further research (Fierro and Søgaard, 2022) delves into the discripencies between languages, attributing these differences to cultural biases. However, the role of cross-lingual transfer in factual representation across languages has not been extensively explored.

3 Multilingual Factual Probing

In this section, we carry out experiments to probe the factual knowledge of ML-LMs across multiple languages. Our objective is to clarify how facts are perceived in different languages and to discern the difference in factual recognition among languages. Additionally, we delve into 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 adopt the mLAMA dataset (Kassner et al., 2021).[1] This dataset is a multilingual extension of LAMA (Petroni et al., 2019) and draws from sources of TREx (Elsahar et al., 2018) and GoogleRE,[2] 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 was created by Y." where entity X, relation and Y form a triplet of (object, relation, subject).

Models In our research, we focus on probing multilingual factual knowledge using prominent encoder-based ML-LMs, notably mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020a). Encoder-based models are chosen over alternatives like mT5 (Xue et al., 2021) due to their proficiency in language understanding tasks. Specifically for our fact-probing task, which employs fill-in-the-blank queries, these models perform well at referencing and integrating information across entire sentences, ensuring a detailed contextual understanding.

3.2 Evaluation

There is an issue in evaluating facts containing multi-token entities, since we need to determine the number of mask tokens for each probed fact. Earlier methods (Jiang et al., 2020a; Kassner et al., 2021) proposed automated techniques for determining mask counts by maximizing the probability of a correct number of mask tokens. Our approach, however, leans toward capturing facts representation rather than purely evaluating the probing performance of ML-LMs.

To maximize the predictable factual prompts, we evaluate two matching methods: full-match and partial-match. In the full-match approach, we assign the exact number of mask tokens corresponding to the object. However, we noticed that this method sometimes misses prompts and yields correct answers containing non-essential tokens such as whitespaces. We may consider these cases not as errors but as potentially valid answers.

Consequently, we introduced the partial-match method. For a query like "[X] was directed by [Y]," we list all objects and their token counts associated with this relation. We then probe ML-LMs with multiple queries, ranging from one mask token (*e.g.*, "[X] was directed by [MASK]") up to the longest mask token sequence for that relation (*e.g.*, "[X] was directed by [MASK] ... [MASK]"). A fact is considered correctly predicted if any version of the prompt includes the right object tokens,

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

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

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: Facts captured by partial-match. The tokens in "()" are extra tokens compared to golden dataset.

regardless of additional preceding or succeeding tokens.

Evaluation result Figure 2 lists the results comparing two ML-LMs and their respective probing methods in terms of the precision at the first rank (P@1) metric. Across all experiments, we noted a consistently low P@1 accuracy, especially for low-resource languages. Detailed results for individual languages can be found in Appendix A.

Interestingly, the partial-match method demonstrated a noticeably better performance in factual probing by considering partially matched predictions. A deeper analysis revealed four unique prediction patterns, specifically discernible using the partial-match method, as highlighted in Table 1. These patterns emphasize the constraints of the mask-token-based probing method, which restricts answers to a single standard format, neglecting the diversity in expressing entities in text. This insight indicates a direction for future improvements in probing techniques.

For clarity in our subsequent analysis, we will primarily focus on mBERT, a 12-layer Transformer language model 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. Therefore, the following discussions will be predominantly based on results obtained using the full-match approach.

4 Languages Discrepancy in Factual Probing

Figure 3 shows the results of factual probing by languge demonstrating a significant difference among languages. In this section, we will delve into the potential reasons for such discrepancy and its relationship with the cross-lingual transfer proficiency of ML-LMs (hereafter, mostly mBERT).

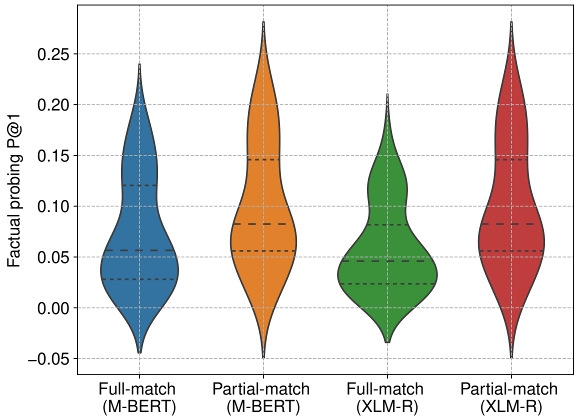

Figure 2: Probing P@1 for full-match and partial-match method in mBERT and XLM-R.

Data vs. Accuracy disparity In English, we observe a prediction accuracy of 19.06%, while in languages like Japanese, it dropped to 1.34%. While prior research has highlighted potential cultural biases in mLAMA, particularly impacting non-Latin script languages (Keleg and Magdy, 2023). However, this bias alone does not explain the pronounced discrepancy between the volume of training data and probing performance. For example, some languages, such as Afrikaans, perform exceptionally well despite having limited Wikipedia data, as shown in Figure 3. The ability of Afrikaans to represent such a breadth of knowledge, even in the face of potential cultural biases, is indeed remarkable.

To assess the influence of training data volume on factual knowledge acquisition, we computed the Pearson correlation coefficient between the quantity of Wikipedia articles[3] utilized for mBERT training and the P@1 score. The correlation yielded a value of 0.43 (Figure 3), indicating a limited effect of training data on learning fact knowledge. We consider several reasons for the inconsistency.

Difficulty in predicting multi-token object entites There is a notable -0.81 correlation between mBERT's P@1 scores and the number of subwords in the target entities. While mBERT and XLM-R have similar P@1 scores in predicting one-token entities, XLM-R's tokenizer captures more one-token

[3]We used the Wikipedia dump closest to the mBERT release date (2018/11) as our data source, ranging from 2018/10/01 to 2018/11/20. The data was dumped from https://archive.org/details/wikipediadumps. We only recorded article with namespace 0 - https://en.wikipedia.org/wiki/Wikipedia:What_is_an_article(Main/Article) as the meaningful article page.

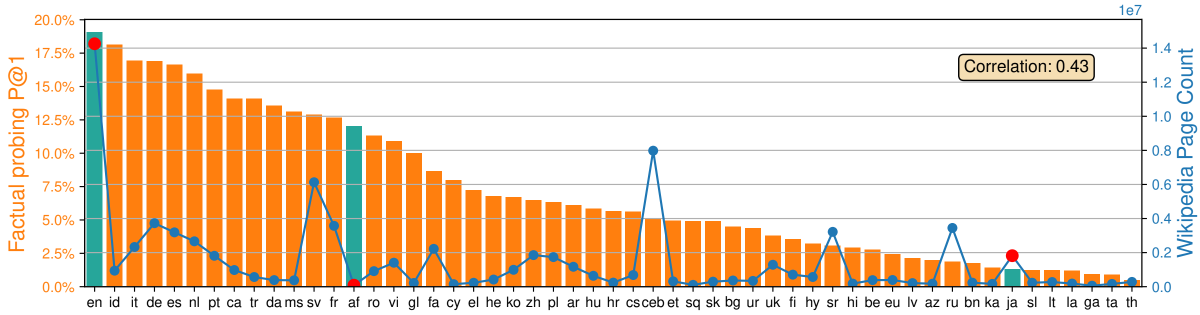

Figure 3: Wikipedia page count vs. Factual probing P@1 for mBERT on 53 languages.

	en	ja	af
mBERT P@1	19.07%	1.34%	12.05%
One-token P@1	15.1%	15.3%	17%
One-token entities	2464	126	498
XLM-R P@1	17.07%	4.78%	8.17%
One-token P@1	15.5%	14.7%	16.6%
One-token entities	1390	244	333

Table 2: P@1 scores and one-token object counts for mBERT and XLM-R for English, Japanese, and Afrikaans.

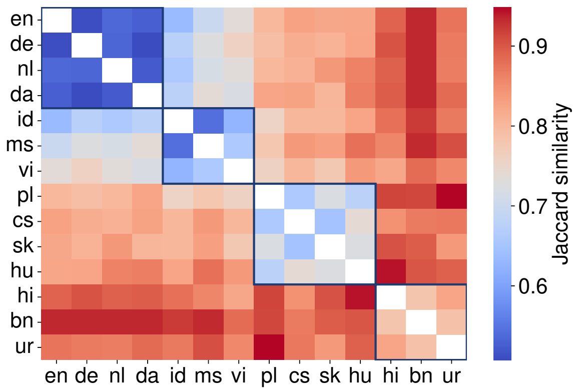

Figure 4: Jaccard similarity matrix of shared factual knowledge across languages in mBERT. Geographically closer languages, like Indonesian, Malay, and Vietnamese from the Southeast Asian group, display higher similarities, signifying substantial shared content.

entities in Japanese, resulting in more accurate predictions. Additionally, the tokenizer of XLM-R often produces shorter tokens for non-Latin scripts, enhancing its performance for non-latin languages. However, this does not completely explain the difference in prediction accuracy shown in Table 2, as Afrikaans greatly outperforms Japanese despite having much less training data.

Localized knowledge cluster The higher accuracy in low-resource languages might result from the model's capacity for cross-lingual factual knowledge sharing. To delve deeper into this, we assessed shared facts between languages using the Jaccard similarity, defined as

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

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

As shown in Figure 4, languages in geographical proximity show greater overlap. This suggests that the model's cross-lingual transfer capacity for factual knowledge might not be universally applicable across all languages. Instead, it appears to be localized, driven more by shared culture and vocabulary. We will explore this phenomenon in subsequent sections.

In this section, we examined various factors that might influence the discrepancies in factual knowledge comprehension across languages. These factors encompass training data volume, tokenizer specifics, and the intrinsic nature of the probing dataset. Our findings reveal localized knowledge sharing patterns among languages, hinting at the potential for cross-lingual transfer capabilities.

5 Language-independent vs. Cross-lingual Shared Knowledge

In this section, we delve into the fact representations and examine how ML-LMs capture such representations. We explore two scenarios: one where ML-LMs maintain several copies of the same fact in different languages, referred as "language-independent" (Figure 1.a) and not optimal for cross-lingual applications. This method is not ideal for cross-lingual applications. Conversely, the "cross-lingual shared/transferred" representation consolidates facts from different languages into a unified embedding space (Figure 1.b & 1.c).

5.1 Factual neuron probing

In Transformer-based language models, the feed-forward network (FFN) plays a pivotal role in the knowledge extraction and representation process. Formally, an FFN is defined as:

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

where $\mathbb{K}$, $\mathbb{V}$, b_1, and b_2 are trainable parameters. Through this function, the FFN selects and recombines knowledge. The specific units of its intermediate layers $f(x\mathbb{K}^T + b_1)$, termed **neurons**, have been shown to possess the capability to express specific knowledge (Geva et al., 2021, 2022). Furthermore, it has been observed that certain factual knowledge is typically encoded in a sparse set of neurons (Dai et al., 2022). These neurons are activated when corresponding knowledge is represented within models.

Experiment setup In our study, we analyze the representation of cross-lingual facts in ML-LMs by identifying their active neurons across languages. We employed the probeless method (Antverg and Belinkov, 2022) - an efficient and explicit technique that measures neurons' activity by contrasting value differences among facts. Specifically, probeless identifies neurons as active when their values deviate significantly from the average for specific knowledge representations.

In detail, our research analyzes neuron activity for each correctly-predicted fact, represented as (subject, relation, object). For probing, we consider other predictable facts that share the same relation but vary in subject-object pairs. We collect the neurons of the mask tokens and identify their active neurons as representatives of the facts. For multi-token masks, we use average pooling across all tokens. As our goal is to investigate facts representation across languages, we collect the active neurons for the same fact in various languages for further analysis. Importantly, the reliability of fact probing decreases when limited predicated facts are available. As such, we focused on the top 30 languages by P@1 score.

5.2 Results & discussion

Cross-lingual and language-independent representation both exist In our neuron probing, we identify active neurons to discern between language-independent and cross-lingual shared/transferred fact representations. Similar patterns in active neurons across languages suggest

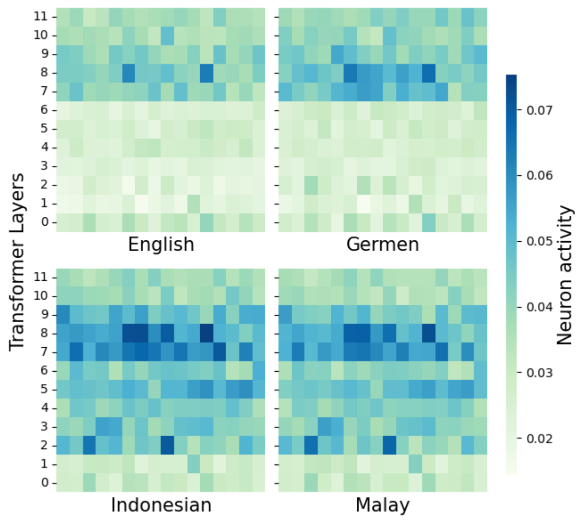

Figure 5: Neuron activity in mBERT across four languages (English, German, Indonesian, Malay) in response to the fact "William Pitt the Younger used to work in London." Color intensity indicates neuron activity, with neurons in each transformer layer grouped into 16 blocks. Distinct activation patterns within the English-German and Indonesian-Malay pairs suggest cross-lingual shared/transferred, while differences between pairs imply language-independent representations.

there is cross-lingual common semantic space for fact representation. Our findings indicate that while some languages exhibit similar neuron activity patterns for a given fact, others may display distinct distributions, as depicted in Figure 5. This reveals the presence of both language-independent and cross-lingual shared/transferred representations within ML-LMs, even for the same fact.

Quantifying cross-lingual sharing Furthermore, to precisely measure the extend of cross-lingual sharing of facts between two languages, we propose a method using the Jaccard similarity based on the top 50 active factual neurons. We then measure the general language similarity between all languages by computing the average similarity for all shared facts, as shown in Figure 6.

Surprisingly, our findings reveal no consistent geographical boundaries among languages, suggesting that either cross-lingual sharing and language-independent are highly depends the fact itself, and such analysis should be tailored to specific factual knowledge. For instance, despite English and Chinese exhibiting a relatively low neuron correlation (0.21, compared to the 0.24 average), they still display similar patterns in active neurons for certain

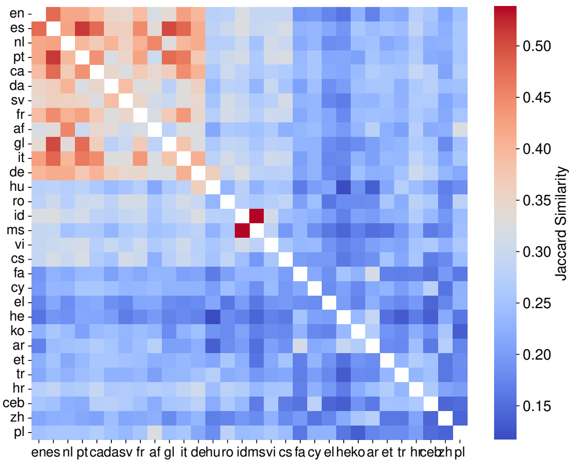

Figure 6: Pairwise similarity between languages measured by shared top 50 active neurons.

facts, often rooted in shared tokens, like "Google" in Chinese "developed-by" relations. See Appendix B for additional analysis.

In this section, we investigated if languages share common fact representations or maintain unique knowledge spaces. Through neuron probing, we found both cross-lingual shared/transferred and language-independent neural activity patterns across languages. Using the Jaccard similarity with active factual neurons, we observed inconsistent geographical boundaries in knowledge sharing, indicating the complexity of cross-lingual knowledge representation.

6 Are cross-lingual representations learned from cross-lingual transfer?

Acknowledging the presence of cross-lingual representation, we subsequently explore its formation mechanism within ML-LMs, assessing whether they are learned individually from distinct language corpora and subsequently aligned into a common semantic space, or if they are acquired through cross-lingual transfer (Figure 1.c).

6.1 Tracing fact origins

To determine the reason behind the formation of the cross-lingual representation, it is crucial to verify if the fact originates from the training corpus. We propose a simple yet effective method to check the presence of a fact in the corpus: for a fact with a triplet (subject, relation, object), we examine the occurrence of the subject and object in the ML-LM training corpus. If they can be found, the fact is considered present. Although this approach may not provide precise quantitative results, it helps in exploring cross-lingual transfer possibilities.

Details Specifically, we are using the Wikipedia data that were dumped just before the release of the mBERT model for verification, as mBERT utilized Wikipedia as its training corpus. We gathered public Wikipedia dumps archived[4] for 53 languages, spanning the period between October 1, 2018, and November 20, 2018. We chose this timeframe to align closely with mBERT's release date, ensuring the data source closely resembled the actual training data of mBERT.

To determine whether a fact is sourced from our training data, we employ subject-object co-occurrence as an approximation method. We rigorously adhere to the preprocessing and sentence-splitting guidelines set out by mBERT, as detailed in (Devlin et al., 2019). We extract only text passages, deliberately omitting lists, tables, and headers using WikiExtractor.[5] Each extracted document is segmented into multiple lines, with each line containing no more than 512[6] tokens. By conducting string matching between the object/subject and Wikipedia, we then assess the co-occurrence of the object and subject for a given fact. If they co-occur, we consider the fact to be present; if not, it's deemed absent.

6.2 Analysis of absent facts

We assessed both the overall absence rate of facts and the absence rate within correctly predicted facts. Figure 7 presents fact verification results for 53 languages, revealing that languages with more training data typically exhibit superior factual knowledge coverage, as anticipated. Nonetheless, several facts, such as those in Afrikaans, are accurately predicted even without verifiable existence in the training corpus, implying a high possibility of cross-lingual transfer effectiveness.

Correctly-predicted facts without knowledge source Upon analysis, we identified that many of the facts that were absent in the knowledge source but correctly predicted are relatively easy to predict. We categorized these into two types. Including other facts, we grouped them into a total of three categories with rule-based methods (See Appendix C).

[4]https://archive.org/details/wikipediadumps

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

[6]The maximum number of tokens allowed to input to mBERT in training.

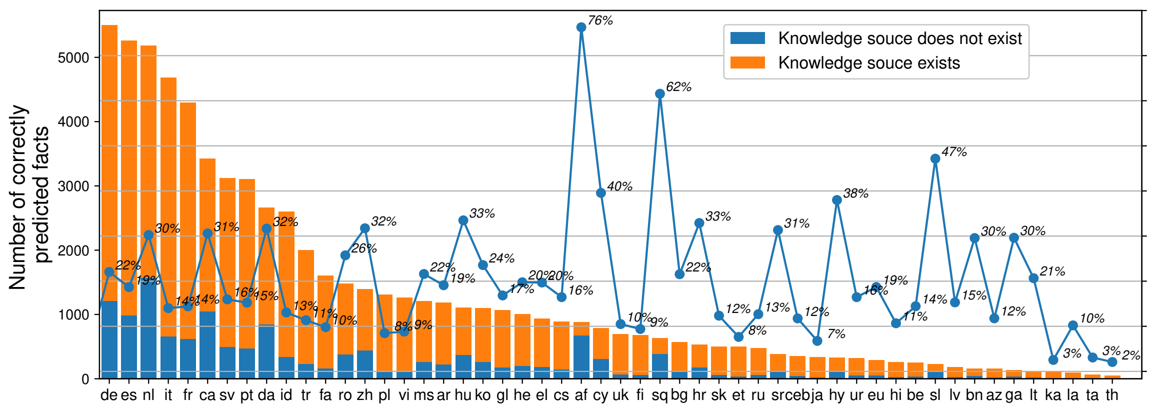

Figure 7: The number of correctly-predicted facts in terms of the existence of possible knowledge source in mBERT.

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

Naming cues: Some probing queries relate to the entity-universal association across person names, countries, and languages, which allows the ML-LMs to guess the object entities from subwords of the subject entities; for example, 'The native language of Go Hyeon-jeong is Korean.' We regard facts related to those relations as this type (see Table 4 in the Appendix for details).

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

Figure 8 shows the counts of correctly-predicted facts by mBERT in each type. The predictability of easy-to-predict facts suggests that the language model can rely on inherent deductions rather than encoding specific facts to make predictions, highlighting the need to enhance factual knowledge probing datasets to more effectively evaluate model proficiency in fact representation. Besides the easy-to-predict facts, the absent rate drops but still not zero (blue bar in Figure 8) for some of the languages, such as Galician, indicating that ML-LMs indeed possess cross-lingual transfer capabilities for factual knowledge, while the applicable languages are limited. More comprehensive results

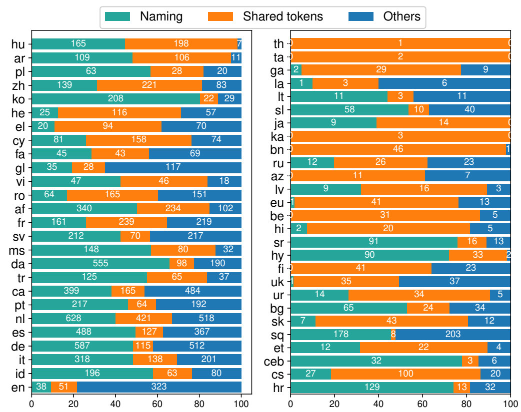

Figure 8: The count of three types of absent & predictable facts

of fact origin checking and examples about the correctly-predicted facts without the knowledge source are given in Appendix C.

7 Conclusions

Our research provides insights and establishes the groundwork for further studies in understanding cross-lingual factual knowledge representation. We identified three distinct patterns for acquiring and representing factual knowledge across languages in ML-LMs: language-independent, cross-lingual shared, and cross-lingual transferred mechanisms. We also introduce methods to quantify these patterns. Our analysis on factual probing reveals the challenges involved in achieving effective cross-lingual transfer of factual knowledge from high-resource to low-resource languages in ML-LMs. In the future, we encourage enhancing the cross-lingual transfer capacity for factual knowledge in ML-LMs and the development of a more precise factual probing dataset.

8 Limitations

We primarily examined two encoder-based Transformer models for language understanding tasks, mBERT and XLM-R. Therefore, our findings may not directly apply to the recent, large-scale decoder-based LMs such as LLama2 (Touvron et al., 2023) and GPT-3 (Brown et al., 2020). Future research should explore these latest models to gain more precise insights.

Moreover, the dataset we utilized has certain limitations. Native speakers identified corrections needed for certain language prompts. Additionally, the dataset focuses on a limited set of relation types, implying that some nuances in fact prediction may lie beyond the scope of our current research.

9 Ethical Consideration

This research is designed the reveal the inner working of factual knowledge learning within language models. We strictly adhered to ethical guidelines, ensuring data privacy and integrity. All datasets utilized were publicly accessible and did not involve sensitive information. The findings and interpretations presented are unbiased and intended for academic purposes. The authors acknowledge and respect the diverse linguistic contexts and toolkits represented in the study.

References

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

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

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.

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, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, 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.

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 *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 *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.

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.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin

Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

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 Factual probing P@1 score

We report the factual probing P@1 scores by both full-match and partial match methods, on mBERT and XLM-R in Table 3.

B Extra result & analysis based on neuron analysis

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

C Additional Data about Fact-origin Checking

C.1 Rules of define fact types

We classify the three types of absent & predictable facts by rules simple.

Shared tokens across entities: We assess whether the object is a substring of the subject or if they share common subwords. Although the latter method might include some irrelevant details, our analysis indicates that the majority of these instances reflect evident shared information. Examples in different languages can be found in Table 5.

Naming cues: We manually selected several relations that contains the information between human name, location and countries, as illustrated in Table 4. Examples in different languages can be found in Table 6.

Others: The left facts are regards as others. Examples in different languages can be found in Table 7.

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

Table 3: Overall P@1 score (Part 2)

Ids	Relation	Example
P103	The native language of [X] is [Y] .	The native language of Jean-Baptiste Say is French .
P37	The official language of [X] is [Y] .	The official language of Aigle is French .
P937	[X] used to work in [Y] .	George Osborne used to work in London .
P17	[X] is located in [Y] .	Noyon is located in France .
P407	[X] was written in [Y] .	El Espectador was written in Spanish .
P20	[X] died in [Y] .	Pius III died in Rome .
P140	[X] is affiliated with the [Y] religion .	Abdullah Ahmad Badawi is affiliated with the Islam religion .
P19	[X] was born in [Y] .	Boniface III was born in Rome .
P364	The original language of [X] is [Y] .	The original language of The Second Sex is French .
P190	[X] and [Y] are twin cities .	New Delhi and Chicago are twin cities .
P1412	[X] used to communicate in [Y] .	Pere Gimferrer used to communicate in Spanish .
P27	[X] is [Y] citizen .	Giovanni Lista is Italy citizen .

Table 4: Relations that contain mostly name, country and location entities.

Language	Absent & Predictable fact
Afrikaans	Vlag van Jamaika is ’n wettige term in Jamaika.
Azerbaijani	Split hava limanı Split adını daşıyır.
Belarusian	Сталцай камуна Гётэбарг з’яляецца Гётэбарг.
Bulgarian	Декларация за създаване на държавата Израел е легален термин в Израел.
Catalan	Govern de Macau és un terme legal en Macau.
Cebuano	Ang Nokia X gihimo ni Nokia.
Czech	Guvernér Kalifornie je právní termín v Kalifornie.
Welsh	Mae seicoleg cymdeithasol yn rhan o seicoleg.
Danish	Danmarks Justitsminister er en juridisk betegnelse i Danmark.
German	Die Hauptstadt von Gouvernorat Bagdad ist Bagdad.
Greek	Υπουργ Δικαιοσνη τη Δανα εναι να νομικ ρο στο Δανα.
English	Sega Sports R&D is owned by Sega .
Spanish	Honda Express es producido por Honda.
Estonian	Seim (Poola) on Poola -is juriidiline termin.
Basque	orbita ekliptiko orbita azpi-klasea da.
Finnish	1955 Dodge tuottaa Dodge.
French	Massacre de Cologne se trouve dans Cologne.
Irish	Tá Contae Utah suite i Utah.
Galician	Sheffield United F.C. recibe o nome de Sheffield.
Croatian	Sjedište Valencia C.F. B je u Valencia.
Hungarian	Honda Fit -et Honda állítja elő.
Indonesian	Menteri Kehakiman Denmark adalah istilah hukum dalam Denmark.
Italian	Nagoya Railroad Co., Ltd è stata fondata a Nagoya.
Japanese	アンフィオン級水艦は水艦のサブクラスです 。
Korean	모빌군의 수도는 모빌입니다.
Latin	Ethica adhibita est pars ethica.
Lithuanian	Stokholmas savivaldybė sostinė yra Stokholmas.
Latvian	Voterfordas grāfiste galvaspilsēta ir Voterforda.
Malay	Sony Alpha 99 dihasilkan oleh Sony.
Dutch	Aluminiumsulfaat bestaat uit aluminium.
Polish	Cadillac Series 60 jest wytwarzany przez Cadillac.
Portuguese	cooperativa autogestionária é uma subclasse de cooperativa.
Romanian	Festivalul Internațional de Film de la Calgary este localizat în Calgary.
Russian	Сенат Теннесси является юридическим термином в Теннесси.
Slovak	BMW N52 sa vyrába v BMW.
Slovenian	Narodno gledališče München se nahaja v München.
Albanian	BBC Music është pjesë e BBC.
Serbian	Аеродром Минск е назван по Минск.
Swedish	Huvudstaden till Guvernementet Bagdad är Bagdad.
Turkish	Waterford County ’un başkenti Waterford’ dir.
Ukrainian	Законодавча асамблея штату Орегон - юридичний термн в Орегон.

Table 5: Examples of easy-to-predict facts with shared tokens in entities on more languages.

Language	Absent & Predictable fact
Afrikaans	Die moedertaal van Jean-Baptiste Say is Frans.
Bulgarian	Официалният език на Бермудски острови е английски език.
Catalan	La llengua nativa de Alain Mabanckou és francès.
Cebuano	Ang Giovanni Lista usa ka lungsuranon sa Italya.
Czech	Embrik Strand pracoval v Berlín.
Welsh	Mae Guillaumes wedi'i leoli yn Ffrainc.
Danish	Mødesproget til Pierre Blanchar er fransk.
German	Die Originalsprache von Young Foolish Happy ist Englisch.
Greek	Ζωρζ Ντυαμλ γεννθηκε στο Παρσι.
English	The original language of Campeones de la vida is Spanish .
Spanish	Bruno Racine solía comunicarse en francés.
Estonian	New Jersey osariik ametlik keel on inglise keel.
Basque	Umar II.a Islam erlijioarekin erlazionatuta dago.
French	Silent Alarm a été écrit en anglais.
Irish	Rugadh Salvador Puig Antich i Barcelona.
Galician	Romain Rolland usado para traballar en París.
Croatian	Izvorni jezik Die Zeit je njemački jezik.
Hungarian	John Hutton az angol nyelven történő kommunikációhoz használt.
Indonesian	Adrian Knox adalah warga negara Australia.
Italian	La lingua originale di The Lunchbox è inglese.
Japanese	ウィリアムハウイットの母語は英語です。
Korean	알랭 마방쿠의 모국어는 프랑스어입니다.
Latin	Paulus Manutius mortuus apud Roma.
Lithuanian	Oficiali Patna kalba yra hindi.
Latvian	Džhārkhanda oficiālā valoda ir hindi.
Malay	Bahasa ibunda Jean-Baptiste Say ialah Bahasa Perancis.
Dutch	The Christian Century is geschreven in Engels.
Portuguese	John Pye costumava trabalhar em Londres.
Romanian	Abdolkarim Soroush este afiliat cu religia islam.
Russian	Насир уд-Дин Абу-л-Фатх Мухаммад связан с религией ислам.
Slovak	Pôvodný jazyk Die Zeit je nemčina.
Slovenian	Hideki Shirakawa se je rodil v mestu Tokio.
Albanian	Georges Rouault vdiq në Paris.
Serbian	Изворни език Жан Батист Се е француски език.
Swedish	Pierre-Jean Mariette brukade arbeta i Paris.
Turkish	The Massacre, İngilizce dilinde yazılmıştır.
Ukrainian	Ренцо Пано використовуться для роботи в Рим.

Table 6: Examples of easy-to-predict facts of naming cues on more languages.

Language	Absent & Predictable fact
Afrikaans	Die hoofstad van Verenigde Koninkryk is Londen.
Azerbaijani	Slovakiya Sosialist Respublikası -nin paytaxtı Bratislava.
Belarusian	Сталцай Татарская АССР з'яляецца Казань.
Bulgarian	Ембриология е част от медицина.
Catalan	Jean-Baptiste-Claude Chatelain va néixer a París.
Cebuano	Kuala Lumpur (estado) mao ang kapital sa Malaysia.
Czech	Beijing College Student Film Festival se nachází v Peking.
Welsh	Mae Meade Lux Lewis yn chwarae piano.
Danish	Jean-Baptiste-Claude Chatelain blev født i Paris.
German	Surinder Khanna wurde in Delhi geboren.
Greek	Πιρ Λεκμτ ντου Νου γεννθηκε στο Παρσι.
English	Aleksandar Novaković was born in Belgrade .
Spanish	Aleksandar Novaković nació en Belgrado.
Estonian	Serbia kuningriik pealinn on Belgrad.
Basque	Libano Mendiko eskualdea hiriburua Beirut da.
Finnish	Art Davis soittaa jazz -musiikkia.
French	Rhigos est un village.
Irish	Is é Toulouse príomhchathair Haute-Garonne.
Galician	Giuliano Giannichedda xoga na posición centrocampista.
Croatian	Glavni grad Narodna Socijalistička Republika Albanija je Tirana.
Hungarian	State University of New York székhelye Albany -ben található.
Indonesian	Ibukota Republik Rakyat Sosialis Albania adalah Tirana.
Italian	Vernon Carroll Porter è nato a Cleveland.
Korean	머피 브라운는 원래 CBS에 방영되었습니다.
Latin	Gulielmus Marx Est politicus per professionis.
Latvian	Itālijas futbola izlase ir loceklis no FIFA.
Malay	Power Rangers Samurai pada mulanya ditayangkan pada Nickelodeon.
Dutch	Power Rangers: Samurai werd oorspronkelijk uitgezonden op Nickelodeon.
Polish	Gregg Edelman to aktor z zawodu.
Portuguese	Jean-Baptiste-Claude Chatelain nasceu em Paris.
Romanian	Capitala lui Republica Populară Socialistă Albania este Tirana.
Russian	Штаб-квартира Jim Beam находится в Чикаго.
Slovak	Leicestershire zdiel'a hranicu s Lincolnshire.
Slovenian	Dilawar Hussain se je rodil v Lahore.
Albanian	Guy Doleman është një aktor me profesion.
Serbian	Седиште компание Чикашка берза е у Чикаго.
Swedish	Jean-Baptiste-Claude Chatelain föddes i Paris.
Turkish	Aruba Futbol Federasyonu, FIFA üyesidir.
Ukrainian	Штаб-квартира Партя «Новий Азербайджан» знаходиться в Баку.

Table 7: Examples of non-easy-to-predict facts on more languages.