Impact of Shared Tokens on Multilingual Text Retrieval's Cross-Lingual Transfer Efficiency

Anonymous ACL submission

Abstract

While great progress has been made on multilingual text retrieval in the recent years, there has been few works in exploring the influential factors behind the transfer effect on multilingual text retrieval. In this work, we study the impact of shared tokens between the source and target languages to the transfering ability on multilingual text retrieval. Via extensive experiments on MIRACL, we find that the impact of shared tokens greatly depends on whether the language is whitespace-delimited, whereas not strongly associated with the number of shared tokens between the training and target language. Our analysis further suggests the impact of shared tokens may be independent from other potential influential factors in the cross-lingual transfer of multilingual retrieval.

	Self	ar	bn	en	es	fa
bn	100.0%	92.0%	–	99.2%	97.6%	88.7%
fa	100.0%	89.4%	58.8%	98.1%	96.1%	–
hi	100.0%	90.5%	72.2%	99.1%	95.9%	87.6%

	fi	fr	hi	id	ja	ko
bn	87.5%	97.2%	81.3%	94.9%	94.7%	89.3%
fa	84.7%	95.7%	65.4%	89.8%	90.2%	82.4%
hi	86.7%	95.5%	–	92.3%	92.4%	85.9%

	ru	sw	te	th	zh	Avg.
bn	94.4%	68.6%	69.5%	85.6%	94.5%	89.0%
fa	93.0%	60.2%	52.6%	74.4%	90.0%	81.4%
hi	92.6%	68.0%	68.9%	82.9%	92.0%	86.8%

Table 1: Token overlap percentage between languages A and B in MIRACL inspired by Zhang et al. (2023a). The value at entry {row A, column B} is the percentage of number of shared tokens between corpus of language A and B over number of all tokens in corpus A. Higher values are highlighted in blue, whereas lower values are highlighted in orange.

1 Introduction

Given a user query, the task of *text retrieval* is to retrieve a list of relevant documents from a collection of millions of documents, where the documents are sorted based on their relevance to the information needs embedded in the query. Retrieval serves as a task of many downstream NLP tasks, for example, question answer, recommender system, and recent works show retrieval systems could be used to argument the language model generation.

Text retrieval techniques have greatly evolved since the pretrained language models (Devlin et al., 2019; Raffel et al., 2020) on both English and multilingual benchmarks. Recent works show that English retrieval models can be adapted to multilingual scenario without by simply switching the backbone from the English pretrained language models to the multilingual ones already yields decent multilingual retrieval effectiveness and exhibits transfer effects across languages. (Zhang et al., 2021, 2023b,a; Bonifacio et al., 2021; Lawrie et al., 2022b,a).

Meanwhile, many works have been exploring the transfer effect of multilingual pretrained language models on NLP tasks, investigating the relation between cross-lingual transfer effect regarding the shared subwords, linguistic features, and many other factors. (Pires et al., 2019; Wu and Dredze, 2019; K et al., 2020; Artetxe et al., 2020; Briakou et al., 2023). However, there are few works exploring influential factors behind the transfer effect in multilingual text retrieval.

Motivated by the observations that there are great number of overlapping tokens among retrieval corpus in different languages (Zhang et al., 2023a) (Table 1), this work aims to investigate the impact of shared tokens in the context of multilingual retrieval. We study the impact of these shared tokens on the cross-lingual transfer learning in text retrieval through extensive experiments on MIRACL. We find that the impact of shared tokens greatly depends on whether the language is whitespace-delimitered, whereas not strongly associated with

the number of shared tokens between the training and target language. Our analysis further suggests the impact of shared tokens may be independent from other potential influential factors in the cross-lingual transfer of multilingual retrieval. Code and models used in this work will be released for future research in the community.

2 Methods

2.1 Multilingual Dense Passage Retriever

This work uses the dense passage retrieval model (DPR) (Karpukhin et al., 2020), one of the earliest dense retrieval models (Karpukhin et al., 2020; Lin et al., 2021; Xiong et al., 2021). We chose the model due to its simplicity in the model architecture and the training process. DPR encodes the query q and passage p independently into vector representations $E_Q(q)$ and $E_P(p)$ by taking the representation of the [CLS] token in the final layer, and then measures the similarity between them via inner product: $s_{q,p} = \text{sim}(q,p) = E_Q(q)^T E_P(p)$. In the training time, the encoders are optimized according to the NCE loss:

$$L(q_i, p_i^+, p_{i,1}^-, ..., p_{i,n}^-) = - \log\left(\frac{\exp(s_{q_i,p_i^+})}{\exp(s_{q_i,p_i^+}) + \sum_{j=1}^{n} \exp(s_{q_i,p_{i,j}^-})}\right)$$

where p_i^+ is the representation of a relevant passage, and p_j^- is the representation of a non-relevant passage. The similarity scores $s_{q,p}$ are used to generate a ranking for query q using a nearest neighbor vector search library (in our experiments, Faiss (Douze et al., 2024; Johnson et al., 2019)).

Following the previous works (Asai et al., 2021; Zhang et al., 2021, 2023a), we initialize the model with multilingual BERT (mBERT) to support retrieval in multiple non-English languages. We refer to the model as “mDPR”.

2.2 Shared Token Removal

In order to examine the relationship between overlapping tokens and cross-lingual transfer learning, as well as relationship with linguistic features, we perform shared token removal. By comparing results from evaluating performance with token overlap to results from evaluating with no token overlap, we wish to make conclusions about the role token overlap plays in transfer learning between languages.

Given a language L, we use an off-the-shelf language detection library `langdetect`.[1] We pre-process the MIRACL training data for the language L by first tokenizing using the mBERT tokenizer, then removing all tokens that `langdetect` finds to be in a script other than the script of L. For example, if L is Russian, our process will remove all *non-Cyrillic* tokens from the training passages. The remaining tokens, which are in the same script with L, form the new data used for fine-tuning mDPR.

We perform the same pre-processing step in the stage of encoding. We consider an encoding language L which refers to the language of the dataset we will be encoding and the language we will evaluate on. In this step, we remove all tokens in the passages and the queries of the encoding dataset that are in a script other than the script of L.

Note that due to the limitation of the language detection accuracy, the remaining tokens are only in the same script with the given language L, but not necessarily belong to the language. However, since we focus on training and language pairs that are in different scripts, this procedure effectively reduces the shared tokens between training and inference data. We thus keep this design for the analysis.

3 Experimental Setup

Data. All experiments on this work are based on MIRACL (Zhang et al., 2023b), a large-scale multilingual retrieval dataset that provides training and evaluation data for 16 diverse languages,[2] released under Apache-2.0 License. MIRACL provides 2k training queries per language on average, where each query are equipped with human-labelled relevant and irrelevant passages (hard negatives) draw from a hybrid retrieval system.

Training and Inference. All experiments in this work are based on Tevatron (Gao et al., 2023), a flexible framework for the training and inference of common retrieval methods. That is, for each language L, we initialize with mBERT provided by HuggingFace (Wolf et al., 2020),[3] then fine-tune on the training data in L from MIRACL with the non-self script tokens removed or not. We end up with two models for each language L supported by MIRACL. The first is a baseline model fine-tuned

[1] https://pypi.org/project/langdetect

[2] MIRACL does not provide training data for German and Yoruba.

[3] https://huggingface.co/google-bert/bert-base-multilingual-uncased

on the official MIRACL training data.[4] The other is fine-tuned on the pre-processed version of the training data, where all tokens that are not in the same script as L are removed.

Similar to the training step, we adopted two versions of evaluation collections for each language L: one is the official MIRACL collection,[5], the other is the collection that removes all tokens not in the same script with L.

We use the same hyper-parameters following previous works (Zhang et al., 2023a): all experiments are fine-tuned for 40 epochs with a batch size of 64 and a learning rate of 1e-5. We limit the query lengths to 32 and the passages lengths to 256 on both training and inference steps.

4 Impact of Shared Tokens in Multilingual Retrieval Transferring

Table 2 compare the zero-shot results before and after removing the shared tokens in the training and evaluation data. We evaluate the zero-shot scores on three target languages (`fa`, `hi`, `bn`) from MIRACL with the models fine-tuned on 14 source languages. Under each source–target language pair, we report three scores: (1) *Baseline*, where we use the official training and evaluation data; (2) *No Shared Token*, where the tokens in non-self script are removed as mentioned in Section 2.2; (3) *Relative Perc.*, the relative percentage of *No Shared Token* scores relative to *Baseline*. We share our observation regarding the impact of removing shard tokens as follows:

Whitespace. Under the percentage row in Table 2, we highlight the entries with values larger than 90% in blue, and the ones smaller than 80% in orange, where we observe clear clustering in the zeros-shot scores with regard to whether the language separate words using whitespace: All languages with whitespace are colored in blue and only slightly impacted by the shared tokens: Most of them achieved over 95% zero-shot scores compared to the baseline. However, all three non-whitespace languages (Japanese (`ja`), Thai (`th`), and Chinese (`zh`)) are greatly impacted by the shared token removal: Most of the relative performance are under 80%. Looking at source languages `zh` specifically, removing shared tokens results in only 60-66% effectiveness compared to the baseline. This suggests that the current sub-word tokenizers are sub-optimal for the non-whitespace languages, and more delicately designed tokenizers are necessary.

Number of Share Tokens. We then examine whether the impact of shared tokens depends on the number of shared tokens between the training data in the source languages and the evaluation data in the target language. However, as shown in Figure 1, we do not observe any clear correlation between them, whether considering the unique number of tokens or the total number of shared tokens.

We additionally inspect sampled passages that contains shared tokens, and find that the majority of these tokens originate from "(entities) translation pair" and "references" (Raffel et al., 2020). We hypothesize that these bilingual pairs only provide

		Self	fa	hi	bn	ru
fa	Baseline	0.474	–	0.318	0.349	0.390
	No Shared Token	–	–	0.318	0.322	0.372
	Relative Perc.	–	–	100.1%	92.5%	95.3%
hi	Baseline	0.385	0.343	–	0.343	0.335
	No Shared Token	–	0.331	–	0.340	0.315
	Relative Perc.	–	96.5%	–	99.3%	94.1%
bn	Baseline	0.597	0.441	0.359	–	0.502
	No Shared Token	–	0.396	0.380	–	0.458
	Relative Perc.	–	89.8%	105.8%	–	91.2%
		fr	**es**	**th**	**zh**	**sw**
fa	Baseline	0.2825	0.380	0.410	0.312	0.303
	No Shared Token	0.287	0.376	0.373	0.189	0.307
	Relative Perc.	101.7%	98.8%	91.0%	60.6%	101.3%
hi	Baseline	0.218	0.297	0.4723	0.244	0.273
	No Shared Token	0.221	0.293	0.296	0.162	0.253
	Relative Perc.	101.6%	98.8%	62.6%	66.4%	92.8%
bn	Baseline	0.263	0.369	0.523	0.3367	0.362
	No Shared Token	0.274	0.365	0.493	0.203	0.364
	Relative Perc.	104.0%	99.0%	94.2%	60.4%	100.5%
		fi	**id**	**ja**	**te**	**ar***
fa	Baseline	0.383	0.337	0.393	0.371	0.433
	No Shared Token	0.383	0.334	0.286	0.369	–
	Relative Perc.	99.9%	99.1%	72.8%	99.4%	–
hi	Baseline	0.308	0.299	0.3393	0.327	0.363
	No Shared Token	0.309	0.307	0.271	0.308	0.342
	Relative Perc.	100.6%	102.7%	79.8%	94.4%	94.3%
bn	Baseline	0.453	0.432	0.482	0.4971	0.498
	No Shared Token	0.466	0.430	0.381	0.521	0.495
	Relative Perc.	102.8%	99.6%	79.0%	104.8%	99.3%

Table 2: nDCG@10 scores on the development set of Persian (`fa`), Hindi (`hi`), Bengali (`bn`) from MIRACL. Column headers are the training languages, and rows are grouped by the evaluation languages. Under each group, the first two rows are scores based on the official MIRACL training data (*Baseline*) and the data without non-self script tokens (*No Shared Token*), and the third row (Relative Perc.) is the ratio of *No Shared Token* scores over the *Baseline* scores, where the blue entries have values greater than 90%, whereas the orange ones are less than 80%. *"No Shared Token" score is not computed for pair `fa`–`ar` since they are in the same script.

[4] https://huggingface.co/miracl/miracl

[5] https://huggingface.co/miracl/miracl-corpus

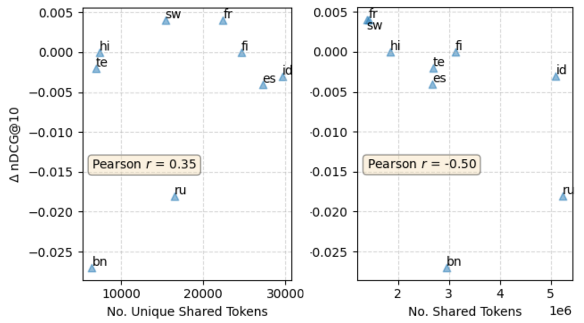

Figure 1: nDCG@10 difference (y-axis) on `fa` with respect to the number of shared tokens (x-axis) between the training and the evaluation data. (*Left*: shared unique tokens; *Right*: all shared tokens). Similar results are observed on `hi` and `bn` (See Figure 5).

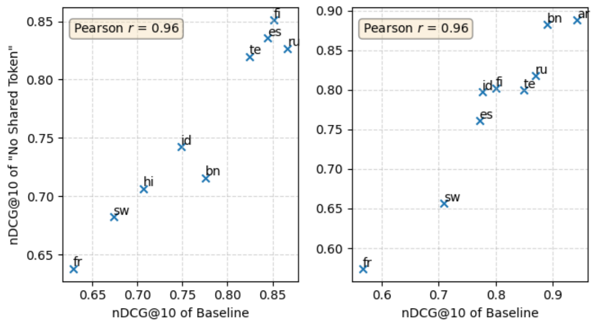

Figure 2: The correlation between zero-shot scores on `fa` (*Left*) and `hi` (*Right*) before and after removing shared tokens, where we observe a strong positive correlation with Pearson $r = 0.96$. Similar results are observed on `bn` (See Figure 4).

minor connections for the bilingual pair, rather than critical aid in the cross-lingual transfer. We share some examples in Appendix B.

Impact of shared token may be orthogonal to other linguistic features. Figure 2 shows the zero-shot scores of *Baseline* (x-axis) and *No Shared Tokens* (y-axis) when using all non-whitespace languages as the training data, where we observe a clear positive correlation (Pearson correlation coefficient $r = 0.96$). This suggests that while removing shared tokens hurt the transfer effect to some extent, this impact is independent of the other (unknown) influential factors in cross-lingual transfer learning, since the relative performance of the languages remains similar.

We additionally investigate the effectiveness difference against the linguistic properties provided by Zhang et al. (2023b) in Appendix C, where we did not observe any clear correlation on all linguistic feature, which also hint the same hypothesis. We leave it for future work to understand the other influential factors behind the cross-lingual transfer.

5 Related Works

There has been many works that studied the impact of shared tokens on the NLP tasks. Pires et al. (2019) showed that while the transfer is possible between languages with no token overlap, more shared tokens lead to better cross-lingual transfer results on the NER task. Wu and Dredze (2019) concurrently examined the impact of the shared token on 5 different NLP tasks (Document Classification, XNLI, NER, POS, and Dependency Parsing), and observed that the amount of shared tokens positively correlates to the cross-lingual transfer ability on all but XNLI tasks. On the opposite, K et al. (2020) found that the shared tokens does not play a significant role in the cross-lingual transfer, and Artetxe et al. (2020) proved that shared tokens are a not necessary requirement in multilingual pretraining. However, all of the above works are conducted on NLP tasks, and it remained unclear how would the share token influence the cross-lingual transfer ability on text retrieval.

6 Conclusion

In this work, we examined the impact of shared token on the transfer effect of cross-lingual transfer effect in the multilingual text retrieval task. We specifically focused on the mDPR model architecture and conduced extensive experiments on MIRACL, a large-scale multilingual retrieval dataset. We find that the impact of shared tokens greatly depends on whether the language is whitespace-delimitered, whereas not strongly associated with the number of shared tokens between the training and target language. Our analysis further suggests the impact of shared tokens may be independent from other potential influential factors in the cross-lingual transfer of multilingual retrieval.

Limitations

This work is limited on that we only considered the dense retrieval models in the multilingual retrieval setup, but not other architectures, e.g., sparse models (Formal et al., 2021), rerankers (Nogueira and Cho, 2019; Nogueira et al., 2020), which we leave for future studies.

References

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Akari Asai, Xinyan Yu, Jungo Kasai, and Hanna Hajishirzi. 2021. One question answering model for many languages with cross-lingual dense passage retrieval. In *Advances in Neural Information Processing Systems*, volume 34, pages 7547–7560.

Luiz Henrique Bonifacio, Vitor Jeronymo, Hugo Queiroz Abonizio, Israel Campiotti, Marzieh Fadaee, , Roberto Lotufo, and Rodrigo Nogueira. 2021. mMARCO: A multilingual version of ms marco passage ranking dataset. *ArXiv*, abs/2108.13897.

Eleftheria Briakou, Colin Cherry, and George Foster. 2023. Searching for needles in a haystack: On the role of incidental bilingualism in PaLM's translation capability. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9432–9452, Toronto, Canada. 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.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library.

Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2021. SPLADE v2: Sparse lexical and expansion model for information retrieval. *arXiv:2109.10086*.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Tevatron: An efficient and flexible toolkit for neural retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 3120–3124, New York, NY, USA. Association for Computing Machinery.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

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*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Dawn J Lawrie, James Mayfield, Douglas W. Oard, and Eugene Yang. 2022a. Hc4: A new suite of test collections for ad hoc clir. *ArXiv*, abs/2201.09992.

Dawn J Lawrie, Eugene Yang, Douglas W. Oard, and James Mayfield. 2022b. Neural approaches to multilingual information retrieval. In *European Conference on Information Retrieval*.

Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. 2021. In-batch negatives for knowledge distillation with tightly-coupled teachers for dense retrieval. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 163–173, Online.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. *arXiv:1901.04085*.

Rodrigo Nogueira, Zhiying Jiang, and Jimmy Lin. 2020. Document ranking with a pretrained sequence-to-sequence model. *arXiv:2003.06713*.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online.

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.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *Proceedings of the 9th International Conference on Learning Representations (ICLR 2021)*.

Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. 2021. Mr. TyDi: A multi-lingual benchmark for dense retrieval. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 127–137, Punta Cana, Dominican Republic.

Xinyu Zhang, Kelechi Ogueji, Xueguang Ma, and Jimmy Lin. 2023a. Toward best practices for training multilingual dense retrieval models. *ACM Trans. Inf. Syst.*, 42(2).

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2023b. MIRACL: A Multilingual Retrieval Dataset Covering 18 Diverse Languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131.

A Additional Figures

Figure 4 and Figure 5 shows the visualization on all three evaluation languages (`fa`, `hi`, `bn`) corresponding to Figure 2 and Figure 1.

B Example of Passages Containing Cross-lingual Shared Tokens

Figure 3 shows examples of passages that contain shared tokens from `bn`, `fi`, and `hi`.

- ইংরেজি ভাষার দৈনিক পত্রিকাগুলোও প্রতিদিন প্রকাশিত হয়ে থাকে, উদাহরণস্বরুপ দি জাকার্তা পোস্ট এবং দি জাকার্তা গ্লোব। চাইনিজ ভাষার পত্রিকাগুলো হল ইন্দোনেশিয়া শাং বাও (印尼商报), হারিয়ান ইন্দোনেশিয়া (印尼星洲日报) এবং গুয়ো জি রি বাও (国际日报)
- لازم به ذکر است کوماتسو زادگاه این شرکت نام شهری است که فرودگاه بین‌المللی ایشیکاوا در آن قرار دارد.این شرکت در حال حاضر یکی از بزرگترین شرکت های دنیا در زمینه ماشین های سنگین است. از دیگر شرکت بزرگترین شرکت()渋谷工業:به ژاپنی)(Shibuya kogyo)های معروف این استان میتوان به شرکت شیبویا 津田駒工業 :به ژاپنی)(Tsudakoma)(در زمینی ماشینهای بسته بندی در جهان) و شرکت تسودا کما (بزرگترین شرکت در زمینی ماشینهای ریسندگی و ماشین‌های صنعتی) اشاره کرد)業
- काला-पैर बिल्ली (Black-footed cat), जिसका वैज्ञानिक नाम फ़ीलिस निग्रिपेस (Felis nigripes) है, बिल्लीनुमा (फ़ेलिडाए) स्तनधारियों के फ़ीलिस वंश की एक जीववैज्ञानिक जाति है। यह छोटे आकार की होती है और दक्षिणी अफ़्रीका क्षेत्र के बोत्स्वाना, नमीबिया और दक्षिण अफ़्रीका देशों में रहती है।

(a) Example of shared tokens originated from translation pairs.

- ১৫ মার্চ (১৯৭২) সংখ্যায় The Supreme Test শিরোনামে সম্পাদকীয় লেখার দায়ে সম্পাদক আবদুস সালামকে পত্রিকার চাকরি থেকে অব্যাহতি দিয়ে ওবায়েদ উল হককে সম্পাদকের দায়িত্ব দেয়া হয়
- درخت بی برای اولین بار در مقاله 《Organization and Maintenance of Large Ordered Indices. Acta Informatica 1: 173–189 (1972)》 به وسیله‌ی و ادوارد م. مک کریت توصیف شد.
- आज Realme कपंती 22 April 2019 को अपना New Smartphone Realme 3 Pro को दोपहर 12 बजे सस्ते price पर india में launch करेगी। Realme 3 Pro Full Specifications स्पेसिफिकेशन्स की बात करि जाए तो इस फ़ोन (Phone) में आपको स्नैपड्रैगन 710 प्रोसेसर (Snapdragon 710 Processor) का पावरफुल प्रोसेसर देखने को मिलेगा। यह फ़ोन 6, 4 जीबी रैम के साथ आ सकता है।

(b) Example of shared tokens originated from reference.

Figure 3: Examples of passages that contains shared tokens.

C Impact of Shared Tokens v.s. Linguistic Features

Figure 6 shows the visualizations on the impact of shared tokens with respect to the linguistic features provided by MIRACL (Zhang et al., 2023b), including language family, synthesis, word order, and whether the language uses grammatical gender system, In all figures, x-axis are the groups in the linguistic features, and y-axis are the percentage difference after removing shared tokens.

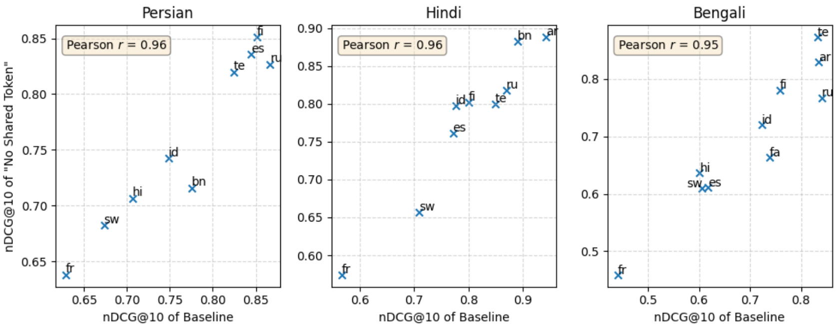

Figure 4: Figure 2 on all three languages: The correlation between zero-shot scores on `fa`, `hi` and `bn` before and after removing shared tokens.

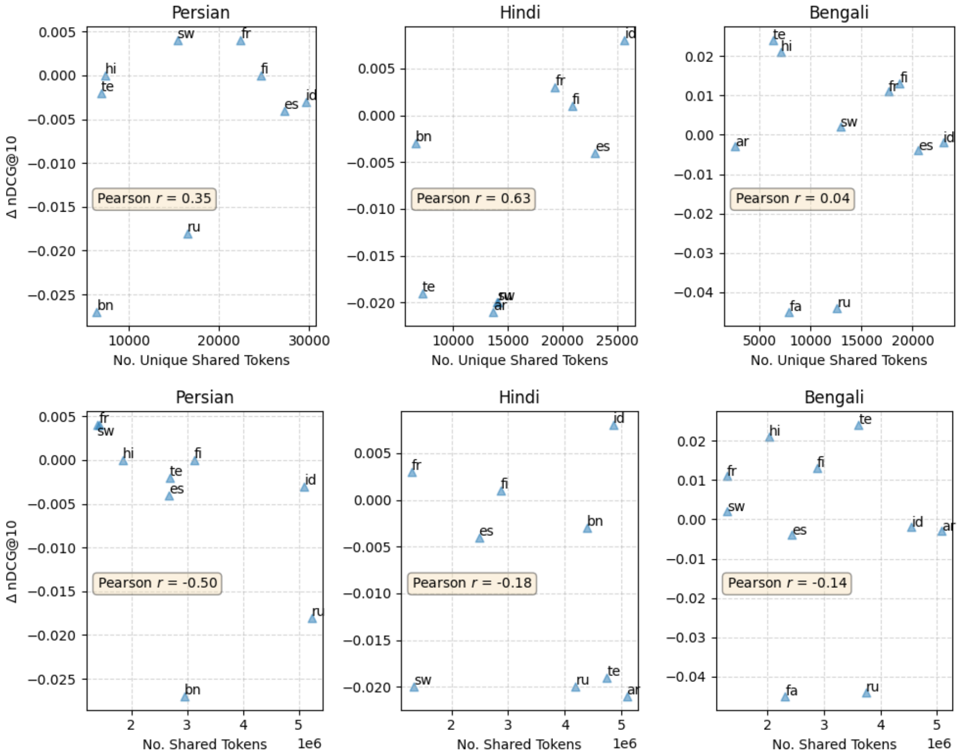

Figure 5: nDCG@10 difference (y-axis) on `fa` with respect to the number of shared tokens (x-axis) between the training and the evaluation data. (*Left*: number of shared unique tokens; *Right*: number of all shared tokens)

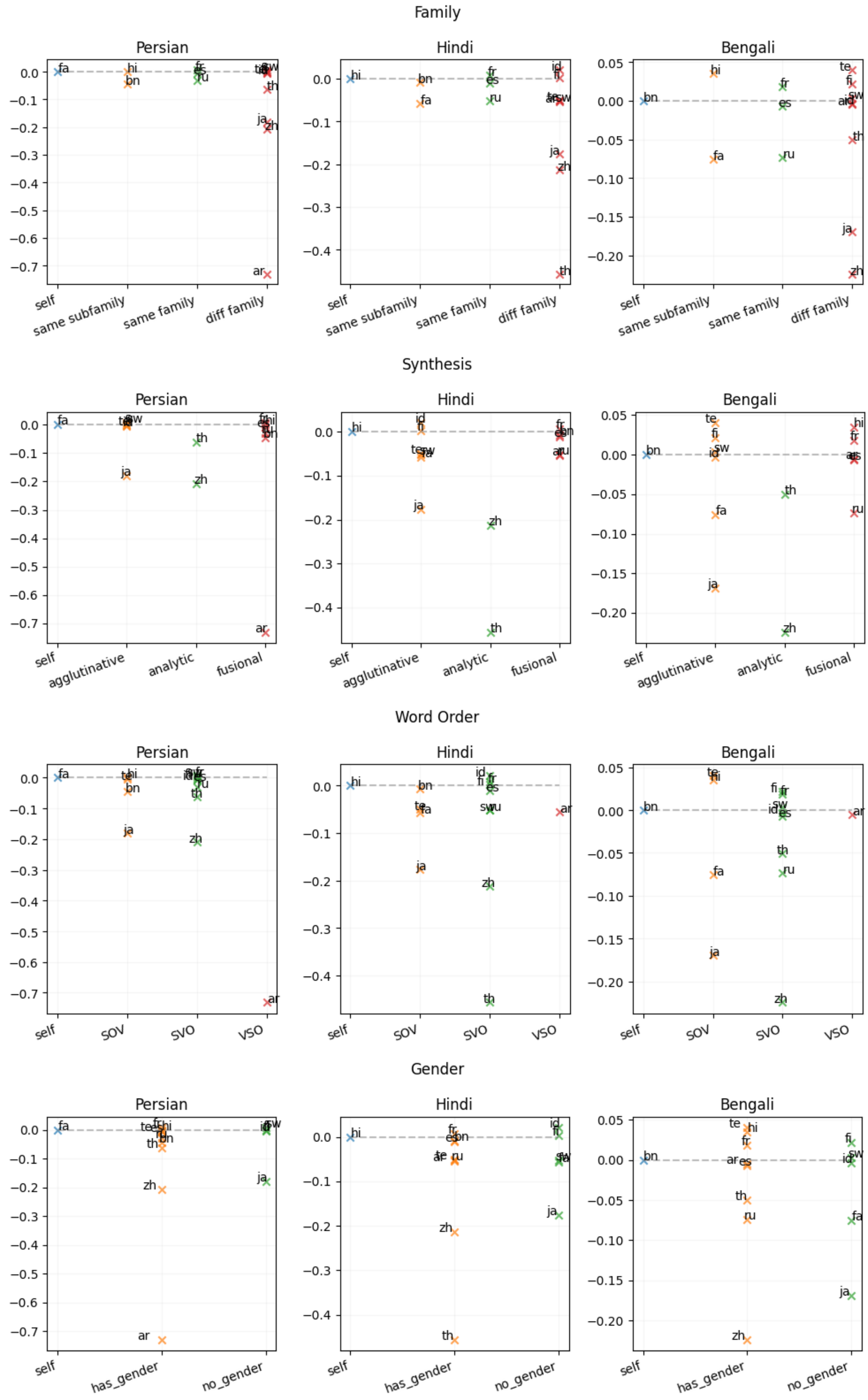

Figure 6: nDCG@10 difference (y-axis) on `fa`, `hi`, and `bn` with respect to the linguistic group.