Investigating Information Inconsistency in Multilingual Open-domain Question Answering

Shramay Palta, Haozhe An, Yifan Yang, Shuaiyi Huang, Maharshi Gor
Department of Computer Science
University of Maryland, College Park
{spalta, haozhe, yang7832, huangshy, mgor}@umd.edu

Abstract

Retrieval based open-domain QA systems use retrieved documents and answer-span selection over retrieved documents to find best-answer candidates. We hypothesize that multilingual Question Answering (QA) systems are prone to information inconsistency when it comes to documents written in different languages, because these documents tend to provide a model with varying information about the same topic. To understand the effects of the biased availability of information and cultural influence, we analyze the behavior of multilingual open-domain question answering models with a focus on retrieval bias. We analyze if different retriever models present different passages given the same question in different languages on TyDi QA and XOR-TyDi QA, two multilingual QA datasets. We speculate that the content differences in documents across languages might reflect cultural divergences and/or social biases.

1 Introduction

The answer to a question is not always unique under different cultural or language backgrounds. When one asks a search engine\(^1\) in English, "How old do you need to be to be able to drink?", the first relevant passage retrieved states "nationally, the legal drinking age is 21." However, if the same question is asked in Chinese, the search engine retrieves a Chinese passage that indicates, "Usually one has to be at least 16 years old to drink, but there is no minimum legal age." Given the same question in different languages, it is possible that a QA system gives inconsistent answers as a result of training on a corpus that reflects various cultural, historical, and social landscapes. We hypothesize that information bias in documents causes models to deliver varying performance. Information bias occurs when key information is either measured, collected, or interpreted differently in the training documents. We empirically investigate the validity of our hypothesis by comparing and contrasting the performance of QA models trained on multilingual dataset.

In addition, in-depth investigations of cultural and/or social biases, specifically in the context of open-domain question answering, are still under explored. We speculate that the content differences in documents across languages are effective indicators of some cultural divergences and/or social biases.

In this work, the focus is on understanding retrieval inconsistencies in multilingual open-domain question answering. Specifically, given a question in any language (e.g. Korean), we feed it into a multilingual passage retriever to retrieve multilingual passages as its context from a huge Wikipedia corpus covering 13 languages including English, Arabic, Finnish, Japanese, Korean, Russian, Bengali, Telugu, Indonesian, Thai, Hebrew, Swedish, and Spanish. We then propose a strategy to get paired questions (questions having the same content but in different languages) and their in-language context. In this way, the difference in retrieved context given questions in different languages (e.g. Korean context for a Korean question, English context for a paired English question) would be a good indicator of information inconsistency in a multilingual setting.

In order to have a quantitative analysis of retrieval differences, we feed the question and its retrieved in-language passage into a shared multilingual answer generator and directly compare the final answer results. We investigate two different settings (oracle and non-oracle setting) for a comprehensive analysis.

Our contributions are summarized as follows:

- We investigate whether QA datasets exhibit information discrepancy or not.

\(^1\)https://www2.bing.com/
We propose a strategy to obtain paired questions for cross-lingual analysis.

We evaluate to what extent each component in the QA system contributes to inconsistent answers on the XOR-TyDi QA benchmark across different languages, with a focus on retrieval inconsistencies.

## 2 Methods

### 2.1 Overview

An overview of our pipeline is shown in Fig. 1. Our framework consists of a multilingual Dense Passage Retriever (mDPR) and a multilingual answer generator. Given a question in any language, we feed it into the question encoder to get an encoding of the question. We retrieve relevant multilingual passages from a huge Wikipedia corpus covering multiple languages by comparing the similarity between question encodings and passage encodings. We select the proper in-language context as its final retrieved context. We then feed the question and its retrieved context into the multilingual answer generator to generate the final answer. We will first describe the details of the mDPR in Sec. 2.2 followed by the multilingual answer generator in Sec. 2.3. Finally, we elaborate on the two settings for analysis as described in Sec. 2.4.

### 2.2 Multilingual Dense Passage Retriever

We follow Asai et al. (2020) for training a multilingual dense passage retriever. We use a multilingual dense passage retriever fine-tuned on XOR-TyDi QA benchmark. The details are as below:

**Training:** mDPR extends Dense Passage Retriever (DPR; Karpukhin et al. (2020)) to a multilingual setting. mDPR uses an iterative training approach to fine-tune a pre-trained multilingual language model (e.g., mBERT; Devlin et al. (2018)) to encode passages and questions separately. Once training is done, the representations for all passages are computed offline and stored locally. We refer the reader to (Asai et al., 2020) for more details.

**Inference:** We get a passage encoding from multilingual BERT given a passage with fixed-length sequence of tokens from multilingual documents. At inference, mDPR independently obtains a d-dimensional (d = 768) encoding of the questions from multilingual BERT. It retrieves k passages with the k highest relevance scores to the question, where the relevance score between a passage and a question q is estimated by the inner product of their encoding vectors.

---

2 https://github.com/mia-workshop/MIA-Shared-Task-2022
With the trained DPR, we retrieve relevant passages across all available languages including English, Arabic, Finnish, Japanese, Korean, Russian, Bengali, Telugu, Indonesian, Thai, Hebrew, Swedish, and Spanish. For each question in its original language, we get its paired question by translating it into other languages and applying the mDPR module for retrieval.

2.3 Multilingual Answer Generator
We follow Clark et al. (2020a) for training the multilingual answer generator. Our multilingual answer generator is MT5 (Xue et al., 2021) pretrained from m4C corpus covering 101 languages. We then finetune the MT5 answer generator on the TyDi-QA training split (Clark et al., 2020a) covering 11 languages.

2.4 Inconsistency and Error Analysis

Oracle Setting: For the oracle setting, we use the ground truth answer as a filter to select the context and the questions to be fed into the answer extractor for analysis. Specifically, we keep the top 20 retrieval results. We keep questions whose retrieval results contain context in at least two languages, and have the ground truth answers contained within the multilingual languages.

Non-Oracle Setting: For the non-oracle setting, we do not use the ground truth answer as a filter to select the context and the questions. Instead, we only keep the top-one score passage. We believe that using the oracle gives bias towards retrieving consistent answers across different languages, and by removing the filter, we can acquire samples that are more likely to reveal information inconsistency.

3 Experiments

In this section, we first elaborate on our implementation details in Sec. 3.1, and follow that with the quantitative and qualitative results in Sec. 3.2.

3.1 Implementation Details

3.1.1 Datasets
TyDi QA (Clark et al., 2020a) is a question answering dataset containing 204K question-answer pairs in 11 typologically diverse languages. It contains instances of language phenomena that would not be found in English-only corpora. In order to reduce the priming effects and provide a realistic and a robust information seeking task, the questions in TyDi QA are written by people who don’t know the answer, and want to know what it is, (unlike SQuAD (Rajpurkar et al., 2016) and its descendants) and the data is collected directly in each language without implementing any translation module (unlike MLQA (Lewis et al., 2019) and XQuAD (Artetxe et al., 2019)). We use TyDi QA for fine-tuning our multilingual answer extractor.

XOR-TyDi QA is a dataset based on TyDi QA. To support research in multilingual open-domain QA, Asai et al. (2020) constructed a dataset, XOR-TyDi QA, of 40K annotated questions across 7 diverse non-English languages. The questions in this dataset are inherited from TyDi QA (Clark et al., 2020) and consists of questions that TyDi QA could not find the same-language answers for. The rationale behind choosing XOR-TyDi QA is that it is the first large-scale cross-lingual open-retrieval QA dataset that consists of information-seeking questions from native speakers and multilingual reference documents. We use the XOR-TyDi QA training split for fine-tuning our multilingual Dense Passage Retriever. The XOR-TyDi QA dev split is used for generating paired questions across languages and for analysing the different types of inconsistencies observed.

3.1.2 Evaluation Metric
We follow the same setup as Lewis et al. (2019). For Chinese, we use the jieba package to perform tokenization; for all other languages, we use white-space tokenization. We remove punctuation characters by checking the corresponding character type in Unicode. We adopt the mean token F1 score to estimate the performance metrics in line with previous extractive QA work to evaluate multilingual question answering accuracy. We also examine the discrepancies in the models by comparing the percentage of answers that are different. Finally, we manually annotate the types of discrepancies/errors observed.

3.1.3 Experimental Configuration
For fine-tuning the multilingual Dense Passage Retriever, we first fine-tune a pretrained multilingual BERT on the Natural Questions (NQ) (Kwiatkowski et al., 2019) dataset. Then we fine-tune our mDPR on the XOR-TyDi QA training split. For fine-tuning the multilingual Answer Generator, we use a pretrained MT5 and then fine-tune it on the TyDi QA training split. We use the AdamW
optimizer with a learning rate of $3e^{-5}$ for fine-tuning the answer generator, with a batch size of 12 and two training epochs.

3.2 Multilingual QA Results

We describe our analysis for multilingual question answering results as below. We identify five types of inconsistency errors which we categorize as translation issues, retriever issues, answer extractor issues, inconsistent facts, and different levels of granularity.

3.2.1 Oracle Setting

In the oracle setting, we use the ground truth answer as a filter to select the context and the questions to be fed into the answer extractor for analysis as described in Sec. 2.4.

Answer Inconsistency Results: There are 146 questions obtained for analysis from a total of 5K questions in the dev set. We observe that 66 questions yield inconsistent answers across languages. We visualize the first 20 questions for simplicity as shown in Fig. 2. We observe that given questions with same meaning but in different languages (e.g. sample index 52 shows question in English (en) and Finnish (fi)), the answers could be different (e.g. the answer asked in fi is correct while the answer asked in en is wrong). This confirms our hypothesis that QA datasets exhibit information discrepancy across languages.

Inconsistency Error Analysis Results: The distribution of the error types is shown in Fig. 2. We manually analyze the different inconsistency errors, and find that only three question samples consist of inconsistent facts in different languages.

Some error are due to issues with the answer translation. The current pipeline translates the gold answer to the target language to compute F1 and Exact Match scores. While this is reasonable, answer translations are not consistent all the time. For examples, name translations across different languages (for example, es → ja) can be different.

3.2.2 Non-Oracle Setting

In the non-oracle setting, instead of only selecting the context with the gold answer, we return the top scored context for each language in the total top 1,000 retrieved passages and use them for generating answers as described in Sec. 2.4.

Answer Inconsistency Results: We perform the same process as mentioned previously for all the 5K samples in the dev set. We observe that almost all questions give inconsistent answers in different languages. We visualize the first 200 questions for simplicity as shown in Fig. 3.
Inconsistency Error Analysis Results: The distribution of the error types is shown in Fig. 3. We manually analyze the different inconsistency errors. As we expand the search space to the top 1,000 retrieved passages, we find that most of the retrieved context is either unrelated or related but does not contain the answer to the question asked. For context that are related, we find that 4 of the question samples consist of inconsistent facts in different languages.

One of these samples asks the question “What is the British population?”. Answers retrieved in different languages present statistics in different years (some language’s Wiki have data for 2010, some has 2007) whereas the context in the English Wiki gives a detailed break-down of population in different areas. Given these different contexts, the answer extractor returns different answers.

We also observe bias in the retrieved context. For a question that asks about the governor of the state of Punjab, all the retrieved context, although in different languages, is either related to Afghanistan or Pakistan.

4 Related Work

4.1 Open-domain Question Answering

Open-Domain QA (OpenQA) has been studied closely with research in Natural Language Processing (NLP), Information Retrieval (IR), and Information Extraction (IE). Traditional OpenQA systems mostly follow a pipeline with three stages, i.e. Question Analysis, Document Retrieval and Answer Extraction (Zhu et al., 2021).

With the advancement of deep learning techniques, OpenQA systems have seen significant developments in every stage. Lei et al. (2018) and Xia et al. (2018) used a CNN-based model and a LSTM-based model in order to develop a question classifier. Khattab et al. (2020), Karpukhin et al. (2020) and Nishida et al. (2018) have worked on developing neural retrieval models for searching documents in a latent space. With the development of large-scale datasets like SQuAD (Rajpurkar et al., 2016), there has been significant progress in the area of neural Machine Reading Comprehension (MRC) methods (Wang et al., 2017b; Seo et al., 2016; Devlin et al., 2018). The traditional OpenQA systems have now developed into the “Retriever-Reader” architecture (Chen et al., 2017; Wang et al., 2017a; Das et al., 2019; Guu et al., 2020). Here the Retriever retrieves the relevant documents for a given question, hence behaving like an IR system. The Reader, which is mostly a neural MRC model, tries to infer the answer from the retrieved documents. Based on this "Retriever-Reader" architecture, there has been progress along the lines of re-ranking the retrieved documents before feeding them into the Reader (Lee et al., 2018; Lin
et al., 2018), using an iterative method to retrieve the relevant documents for a question, (Feldman and El-Yaniv, 2019), and doing end-to-end training on the OpenQA System (Lee et al., 2019).

4.2 Multilingual Question Answering

A lot of research has been going on in the domain of multilingual question answering. Recently, Asai et al. (2020) introduced a new task framework named Cross-lingual Open Retrieval Question Answering (XOR QA), that consists of three new tasks involving cross-lingual document retrieval from multilingual resources.

In order to overcome the issue of data-scarcity in non-English languages, several non-English QA datasets have been created like TyDi QA (Clark et al., 2020b), XOR-TyDi QA (Asai et al., 2020) etc. Along similar lines is the dataset MLQA (Lewis et al., 2019), which consists of over 12K QA instances in English and 5K each in other languages (Arabic, German, Spanish, Hindi, Vietnamese and Simplified Chinese). Each QA instance is parallel between at least 4 languages on average. MLQA serves as a cross-lingual extension to existing extractive QA datasets, given it’s built using a novel alignment context strategy on articles from Wikipedia. Recently, Hardalov et al. (2020) developed a cross-lingual and multilingual dataset for high school examinations. Liu et al. (2019) developed XQA which has a great degree of lexical overlap between reference paragraphs and the questions.

4.3 Bias in Question Answering

Although current linguistic machine learning models achieve numerically-high performance on the question answering task, they often lack optimization for reducing implicit biases. There are many works that try to analyze or tackle biases in QA. Zhu et al. try to capture and eliminate position bias in large language models in question answering. Ahn and Oh (2021) developed a metric called Categorical Bias score and attempted to reduce ethnic bias in BERT models for multilingual question answering. Recent attempts to tackle bias try to either automatically generate additional language sources to balance the training data distribution (Chen et al., 2020; Abbasnejad et al., 2020) or to learn a biased model with the source of bias as the input, and then attempt to debias it through log-its re-weighting or logitsensembling (Clark et al., 2019; Cadène et al., 2019).

5 Conclusion

In this work, we investigate whether multilingual QA tasks expose information inconsistencies across different languages. We evaluate to what extent each component of the QA system contributes to inconsistent answers on the XOR-TyDi QA benchmark across nine languages. We observe inconsistent answers given questions with the same content, but in different languages. We analyze the potential reasons that could lead to such inconsistencies including retriever issues, answer extractor issues, different levels of granularity, translation issues, and inconsistent facts, of which retriever issues and answer extractor issues are the most prominent. Our analysis provides new insights into inconsistencies in multilingual question answering, and is one step closer towards a more realistic application of this task.

References

Ehsan Abbasnejad, Damien Teney, Amin Parvaneh, Javen Shi, and Anton van den Hengel. 2020. Counterfactual vision and language learning. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 10041–10051.

Jaimeen Ahn and Alice Oh. 2021. Mitigating language-dependent ethnic bias in bert. arXiv preprint arXiv:2109.05704.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. CoRR, abs/1910.11856.

Akari Asai, Jungo Kasai, Jonathan H Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2020. Xor qa: Cross-lingual open-retrieval question answering. arXiv preprint arXiv:2010.11856.

Rémi Cadène, Corentin Dancect, Hedi Ben-youtes, Matthieu Cord, and Devi Parikh. 2019. Rubi: Reducing unimodal biases in visual question answering. CoRR, abs/1906.10169.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. arXiv preprint arXiv:1704.00051.

Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. 2020. Counterfactual samples synthesizing for robust visual question answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10800–10809.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Don’t take the easy way out: ensemble
Table 1: An example of the question pairs from the English and Russian subsets of TyDi QA found by using question embedding cosine similarity as the metric.

| English questions | Russian questions translated into English | Cosine similarity |
|-------------------|------------------------------------------|-------------------|
| How many people died during WW1? | How many Germans were captured in the Battle of Stalingrad? | 0.71 |
| When was the Nazi party founded? | When was the KPSS founded? | 0.78 |
| When was the University of Maryland football team established? | When was the Rome Football Club established? | 0.78 |
| When was Siri introduced by Apple? | What year was Apple formed? | 0.74 |
| When was the term anthropology first used? | What does the word "anthropology" mean? | 0.72 |

Figure 4: (a) t-SNE projection of 500 English question embeddings and 500 Russian question embeddings. The well-mixed scatters indicate the existence of similar questions across English and Russian. (b) No significant correlation between sentence embedding similarity and translation BLEU score is observed.

Based methods for avoiding known dataset biases. arXiv preprint arXiv:1909.03683.

Jonathan H Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020a. Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. Transactions of the Association for Computational Linguistics, 8:454–470.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020b. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. Transactions of the Association for Computational Linguistics, 8:454–470.

Rajarshi Das, Shehzaad Dhuliawala, Manzil Zaheer, and Andrew McCallum. 2019. Multi-step retriever-reader interaction for scalable open-domain question answering. CoRR, abs/1905.05733.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Yair Feldman and Ran El-Yaniv. 2019. Multi-hop paragraph retrieval for open-domain question answering. arXiv preprint arXiv:1906.06606.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: retrieval-augmented language model pre-training. CoRR, abs/2002.08909.

Momchil Hardalov, Todor Mihaylov, Dimitrina Zlatkova, Yoan Dinkov, Ivan Koychev, and Preslav Nakov. 2020. EXAMS: A multi-subject high school examinations dataset for cross-lingual and multilingual question answering. CoRR, abs/2011.03080.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. arXiv preprint arXiv:2004.04906.

Omar Khattab, Christopher Potts, and Matei Zaharia. 2020. Relevance-guided supervision for openqa with colbert. CoRR, abs/2007.00814.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. Transactions of the Association of Computational Linguistics.

Jinhyuk Lee, Seongjun Yun, Hyunjae Kim, Miyoung Ko, and Jaewoo Kang. 2018. Ranking paragraphs
Figure 5: (a)t-SNE projection of 500 English question embeddings and 500 Finnish question embeddings. The well-mixed scatters indicate the existence of similar questions across English and Finnish. (b) No significant correlation between sentence embedding similarity and translation BLEU score is observed.
A Multilingual Question Answering
Examples

We provide more multilingual question answering examples with analysis in Fig. 6 and Fig. 7.

B Crafting Multilingual QA Test set

B.1 Finding similar questions

TyDi QA (Clark et al., 2020a) is a benchmark for information-seeking question answering in typologically diverse languages, including English and Russian. We first attempt to find a subset of questions that are present in multiple languages but with potentially controversial and different answers. To identify similar questions across different languages, we compare their contextual sentence embeddings (Reimers and Gurevych, 2019). For each English question $Q^{en}$ and each Russian question $Q^{ru}$, we obtain their respective sentence embeddings $e^{en}, e^{ru} \in \mathbb{R}^{768}$. We then compute the pairwise cosine similarities of these embeddings to find similar questions.

Similarly, we project 500 English question embeddings and 500 Russian question embeddings via t-SNE as shown in Fig. 5a. Through this visualization, we see that there exist some embeddings that are quite close to each other. The inseparability of the two clusters in this two-dimensional space indicates that some questions share considerable semantic and syntactic similarities.

We then filter the question pairs whose embeddings have cosine similarity greater than 0.7. To evaluate if it is effective to use the sentence embeddings cosine similarity as a metric for question similarity, we translate the Russian questions into English (Tiedemann and Thottingal, 2020). We hypothesize that a higher cosine similarity between a pair of English and Russian question embeddings would lead to a higher BLEU score after we translate. However, Fig. 5b does not indicate this trend.

There is little correlation between sentence embedding cosine similarity and the BLEU score. This result is likely due to the short length of questions and a small overlap of the same named entities in questions across the two languages. We observe a similar trend between English and Finnish as well as shown in Fig. 5.

Nonetheless, using the cosine similarity as a metric gives us some questions on similar topics. We illustrate a few examples in Table. 1. We find that questions about similar topics tend to have higher cosine similarities. For example, both the English and the Russian questions ask about wars in history or sports clubs. Through these limited number of examples, we already see some different cultural divergences in the two subsets. For example, the question “When was the KPSS founded?” is equivalent to “When is the Communist Party of the Soviet Union established?” This question is more culturally specific to Russia and will be less likely to appear in the English question subset.
The history of road transport started with the development of roads by humans and their beasts of burden. The first forms of road transport were animal trails, such as the Hardwick Trace, in the Penn State Forest. In the 17th and 18th centuries, roads were developed based on humans and animal traffic. The first improved trails would have been those in the North American states, like the Pennsylvania Trace, which was developed through swampy terrain, consisting of crossing and bridge over slopes. As traffic increased, the roads were often flattened or widened to accommodate human and animal traffic.

In the history, roads and information flow are close to each other, and people came and went frequently. At that time, the roads were the main arteries of communication, and trade goods were transported on them. The roads were often flattened or widened to accommodate human and animal traffic.

When roads and information flow are close to each other, and people came and went frequently. At that time, the roads were the main arteries of communication, and trade goods were transported on them. The roads were often flattened or widened to accommodate human and animal traffic.

In 1776, the history of road transport started with the development of roads by humans and their beasts of burden. The first forms of road transport were animal trails, such as the Hardwick Trace, in the Penn State Forest. In the 17th and 18th centuries, roads were developed based on humans and animal traffic. The first improved trails would have been those in the North American states, like the Pennsylvania Trace, which was developed through swampy terrain, consisting of crossing and bridge over slopes. As traffic increased, the roads were often flattened or widened to accommodate human and animal traffic.
### Error Type | Question | Context | Predicted Answer in English (for demonstration only) |
|---|---|---|---|
| EN | What is the oldest collection of poetry in the world? | The oldest surviving speculative fiction poem is the "Tale of the Shipwrecked Sailor", written in "Hieratic" and ascribed to the ancient Egyptian god Thoth. | "Type of the Shipwrecked Sailor" |
| TE | దీవు నిర్మాణం ఎవరు కాదు? | దీవు నిర్మాణం శతాబ్దిల వింతల వరకు వయసు లో శతాబ్దిల వింతల వరకు వయసు లో శతాబ్దిల వింతల వరకు వయసు లో శతాబ్దిల వింతల వరకు వయసు లో శతాబ్దిల వింతల వరకు వయసు లో శతాబ్దిల వింతల వరకు వయసు లో శతాబ్దిల వింతల వరకు వయసు లో శతాబ్దిల వింతల వరకు వయసు లో శతాబ్దిల వింతల వరకు వయసు లో శతాబ్దిల వింతల వరకు వయసు లో | 360 BC. |
| TE | తిరిగితే సమాధి అభివృద్ధి కాంతిలో కాదు? | తిరిగితే సమాధి అభివృద్ధి కాంతిలో కాదు? | 360 BC. |

### Data Type: Multilingual Question Answering Examples with Analysis

**Figure 7**