XOR-TYDI QA Datasheet

Akari Asai[♣], **Jungo Kasai**[♣], **Jonathan H. Clark**[♦], **Kenton Lee**[♦], **Eunsol Choi**[♥], **Hannaneh Hajishirzi**[♣♠]
[♣]University of Washington [♦]Google Research
[♥]The University of Texas at Austin [♠]Allen Institute for AI
{akari, jkasai, hannaneh}@cs.washington.edu
{jhclark, kentonl}@google.com, eunsol@cs.utexas.edu

1 Motivation for Datasheet Creation

Why was the dataset created? We bring together for the first time information-seeking questions, open-retrieval QA, and multilingual QA to create a multilingual open-retrieval QA dataset that enables cross-lingual answer retrieval. This task framework reflects well real-world scenarios where a QA system uses multilingual document collections and answers questions asked by users with diverse linguistic and cultural backgrounds (Fig. 1). Despite the common assumption that we can find answers in the target language, web resources in non-English languages are largely limited compared to English (*information scarcity*), or the contents are biased towards their own cultures (*information asymmetry*). To solve these issues, XOR-TYDI QA (Asai et al., 2020) provides a benchmark for developing a multilingual QA system that finds answers in multiple languages.

Has the dataset been used already? All papers reporting results on XOR-TYDI QA are required to submit their system outputs to https://nlp.cs.washington.edu/xorqa/.

Who funded the dataset? XOR-TYDI QA was funded by Google, ONR (N00014-18-1-2826), DARPA (N66001-19-2-403), NSF (IIS1252835, IIS-1562364), the Allen Distinguished Investigator Award, the Sloan Fellowship, and the Nakajima Foundation Fellowship.

2 Dataset Composition

What are the instances? At their core are information-seeking questions from native speakers of one of the 7 typologically diverse languages.

How many instances are there? Shown in Table 1 are the data sizes for the 7 languages. In total, about 40k questions are newly annotated with answers from English Wikipedia (cross-lingual data). In-language data are annotated with answers from the target language Wikipedia articles

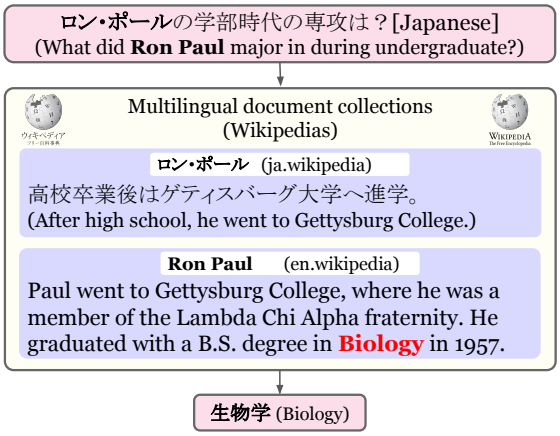

Figure 1: Overview of XOR QA. Given a question in a target language (e.g., Japanese, Korean, or Russian), the model finds an answer in either English or the the target language Wikipedia and returns an answer in the target language.

	Cross-lingual			In-language		
	Train	**Dev**	**Test**	**Train**	**Dev**	**Test**
Ar	2574	351	137	15828	357	1133
Bn	2582	312	128	2428	115	138
Fi	2088	362	530	7680	253	1197
Ja	2288	296	449	5527	140	869
Ko	2469	299	647	1856	74	507
Ru	1941	255	235	7349	313	1125
Te	1308	238	375	5451	113	712
Tot.	15250	2113	2501	46119	1365	5681

Table 1: Dataset size of the XOR-TYDI QA corpus. **Cross-lingual** data come from our re-annotated questions that did not originally have same-language answers in TYDI QA. **In-language** data are taken directly from answerable questions in TYDI QA. During evaluation, we exclude the questions for which we cannot find any minimal answer annotations (about 28% of the questions).

L	Original Question	Passage Answer	Minimal Answer	Final Answer
Ko	1993년 프랑스 총리는 누구인가요? (Who was the French Prime Minister in 1993?)	Mayor of Neuilly-sur-Seine from 1983 to 2002, he was Minister of the Budget under Prime Minister Édouard Balladur (1993–1995).	Édouard Balladur	에두아르 발라뒤르
Ru	Какая средняя зарплата в Краснодаре на сегодняшний день? (What is the average wage in **Krasnodar**?)	Krasnodar has the lowest unemployment rate among the cities of the Southern Federal District at 0.3% of the total working-age population. In addition, Krasnodar holds the first place in terms of highest average salary—21,742 rubles per capita.	21,742 rubles	21,742 рубля
Ja	速水堅曹はどこで製糸技術を学んだ？ (Where did **Kenso Hayami** learn the silk-reeling technique?)	藩営前橋製糸所を前橋に開設。カスパル・ミュラーから直接、器械製糸技術を学び (he founded Hanei Maebashi Silk Mill and learned instrumental silk reeling techniques directly from Caspal Müller)	–	藩営前橋製糸所 (Hanei Maebashi Silk Mill)

Table 2: Examples newly annotated for Korean (Ko) and Russian (Ru) questions. The bottom example is an answerable question from TyDi QA for which only Japanese Wikipedia includes the correct answer.

and taken from TyDi QA.

What data does each instance consist of? Each instance contains a question written by a native speaker who is actually interested in knowing an answer. We annotate each question with paragraphs from Wikipedia articles that has answer content. And these paragraphs are further annotated with minimal answer spans that answer the question. See Table 2 for examples.

Each data point contains a {query, answer span} written in the same language, L_i, and the L_i is one of 7 typologically diverse languages. For the newly annotated portion, we also have {English query, English answer span, a English Wikipedia paragraph where the answer span is found}, and for the existing TyDiQA annotation, we have {a Wikipedia paragraph in L_i where the answer was found}.

Does the data rely on external resources? All necessary data are available in the release, including questions with in-langauge answers from TyDi QA.

Are there recommended data splits or evaluation measures? To facilitate a wide range of research with Xor-TyDi QA, we defined three tasks in the order of increasing complexity: XOR-Retrieve, Xor-EnglishSpan, and Xor-Full. In XOR-Retrieve, a system retrieves paragraphs from English Wikipedia that contain information to answer the question posed in the target language. This is evaluated with recall-based measures. Xor-EnglishSpan takes one step further and finds a minimal answer span from the retrieved English paragraphs. This is evaluated with F1 and exact matching scores. Finally, Xor-Full expects a system to generate an answer end to end in the target language by consulting both English and the target language's Wikipedia. We measure F1, exact matching, and BLEU scores in the target language for evaluation. We use the same random data split for all three tasks, and the answer annotations for the test data are all hidden from public.

3 Data Collection Process

How was the data collected? Our annotation pipeline consists of four steps: 1) collection of realistic questions that require cross-lingual references by annotating questions from TyDi QA without a same-language answer; 2) question translation from a target language to the pivot language of English where the missing information may exist; 3) answer span selection in the pivot language given a set of candidate documents; 4) answer verification and translation from the pivot language back to the original language. The first step was done by collecting questions without answer annotations from an existing dataset, TyDi QA. We proceeded with the second and fourth steps by collaborating with our university

volunteers and a third party translation service, Gengo.[1] We performed crowdsourcing via Amazon Mechanical Turk[2] for the third step of answer annotation.

Who was involved in the collection process and what were their roles? As mentioned above, our university volunteers, the Gengo translation service, and crowdworkers from Amazon Mechanical Turk were involved in the collection process.

Over what time frame was the data collected? The dataset was collected over a period of March 2020 through September 2020, during which we constantly monitored the progress and data quality.

Does the dataset contain all possible instances? Ultimately, the goal of multilingual question answering research is to build a system that can answer any question from anyone in any language. For this reason, our dataset is obviously a sample from the infinite space of potential languages, people, and questions.

If the dataset is a sample, then what is the population? We sample 7 diverse languages from 7 distinct language families: Arabic, Bengali, Finnish, Japanese, Korean, Russian, and Telugu. For each language, questions are written by native speakers who read a randomly chosen Wikipedia article as a prompt. These 7 languages represent a great degree of typological diversity that will be key in deploying QA systems to many languages in the world. For example, Japanese is an agglutinative language with SOV word order and four distinct alphabets. Bengali is a morphologically-rich language that has inflection, affixation, compounding, and reduplication. Moreover, Bengali Wikipedia has fewer than 100k Wikipedia articles as of fall 2020, serving as a benchmark for low-resource question answering.

4 Data Preprocessing

What preprocessing / cleaning was done? Our questions are originally sampled from TyDi QA (Clark et al., 2020). In particular, we randomly sample 5000 questions without any passage answer annotations (unanswerable questions) from the TyDi QA train data, and split it into training (4,500) and development (500) sets. We downloaded Wikipedia dumps for the 7 languages and English, which are archived by TyDi QA authors.[3] We extract plain text from the Wikipedia dump data by wikiextractor,[4] and then we process the data by following common practice in open domain QA work (Chen et al., 2017).

To clean the final data, we trained undergraduate students who are native English speakers to verify the annotated paragraphs and short answers. Only 8% of the answers were marked as incorrect through the verification phase and were later corrected by our pool of high-quality crowdworkers who yielded less than 1% annotation error.

Was the raw data saved in addition to the cleaned data? As we re-annotate those noisy annotations, we do not include the raw data in our main dataset to avoid confusion. Those raw data will be available upon request. In the raw annotation data, we preserve the comments and feedback provided by our annotators and students who verified the data, which often provide useful feedback on why the data were flagged as noisy.

Does this dataset collection/preprocessing procedure achieve the initial motivation? XOR-TyDi QA is a dataset that allows researchers to explore phenomena specific to cross-lingual open-retrieval QA and to work towards building a real-world open-retrieval QA system for diverse languages. This is the first attempt to construct a large-scale cross-lingual open-retrieval QA dataset, and we limit the retrieval target to English (i.e., a system is required to search English or the target language Wikipedia) to collect high-quality data at scale from crowdworkers. Although English Wikipedia is by far the largest among all of the Wikipedias and we observe our questions often become answerable given English Wikipedia information, searching more Wikipedias to answer questions about certain cultures (e.g., Japanese Wikipedia is likely to more information about a Japanese manga than any other Wikipedias). We also sample 7 languages from TyDi QA based on the cost and availability of translators. We hope future work will extend the retrieval target as well as the target languages to a more diverse set of languages.

[1] https://gengo.com/

[2] https://www.mturk.com/

[3] https://storage.googleapis.com/tydiqa/tydiqa.pdf

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

5 Dataset Distribution

How is the dataset distributed? It is available for downloads at https://nlp.cs.washington.edu/xorqa/.

When was it released? Oct. 2020.

What license (if any) is it distributed under? XOR-TYDI QA is distributed under the CC BY-SA 4.0 license.[5]

Who is supporting and maintaining the dataset? XOR-TYDI QA will be maintained by the first two authors of the paper: Akari Asai and Jungo Kasai. All updates will be posted on the dataset website.

6 Legal and Ethical Considerations

Were workers told what the dataset would be used for and did they consent? Crowdworkers consented to have their responses used in this way through the Amazon Mechanical Turk Participation Agreement.

If it relates to people, could this dataset expose people to harm or legal action? Our dataset can include incorrect information to the extent that Wikipedia can have wrong information about people. Nonetheless, we performed extensive quality control and answer verification to minimize the risk of harming people.

If it relates to people, does it unfairly advantage or disadvantage a particular social group? One fundamental problem with the recent question answering benchmarks is that most of their questions are written by native English speakers and overly represent English-centric topics, such as American politics, sports, and culture. As such, models trained and developed on those datasets are likely to fail to serve people with diverse language and cultural backgrounds. XOR-TYDI QA partially remedies this long-standing problem by annotating questions from native speakers of diverse languages. Thus, we encourage researchers and developers to benchmark on XOR-TYDI QA to mitigate the potential bias and unfairness of QA systems. We acknowledge, however, this dataset still covers a very limited subset of languages in this world.

[5] https://creativecommons.org/licenses/by-sa/4.0/legalcode

References

Akari Asai, Jungo Kasai, Jonathan H. Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2020. XOR QA: Cross-lingual open-retrieval question answering.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *ACL*.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *TACL*.