MKQA: A Linguistically Diverse Benchmark for Multilingual Open Domain Question Answering

Shayne Longpre
Apple Inc.
slongpre@mit.edu

Yi Lu
Apple Inc.
ylu7@apple.com

Joachim Daiber
Apple Inc.
jodaiber@apple.com

Abstract

Progress in cross-lingual modeling depends on challenging, realistic, and diverse evaluation sets. We introduce Multilingual Knowledge Questions and Answers (MKQA), an open-domain question answering evaluation set comprising 10k question-answer pairs aligned across 26 typologically diverse languages (260k question-answer pairs in total). Answers are based on heavily curated, language-independent data representation, making results comparable across languages and independent of language-specific passages. With 26 languages, this dataset supplies the widest range of languages to-date for evaluating question answering. We benchmark a variety of state-of-the-art methods and baselines for generative and extractive question answering, trained on Natural Questions, in zero shot and translation settings. Results indicate this dataset is challenging even in English, but especially in low-resource languages.[1]

1 Introduction

Training and evaluation data for question answering (QA) is severely lacking outside of high-resource languages like English. As unsupervised, transfer learning and zero/few-shot methods narrow the multilingual performance gap with English (Conneau et al., 2020; Lee and Lee, 2019; Cui et al., 2019a; Lewis et al., 2020), their real progress is hard to measure without challenging, realistic, and linguistically diverse evaluation sets. Existing multilingual QA datasets are realistic and challenging, but they lack linguistic diversity, comparable evaluation between languages, and are often limited to passages provided with the dataset (see Table 2).

We introduce Multilingual Knowledge Questions and Answers (MKQA) for evaluation of open-domain question answering. MKQA selects 10k realistic English queries from the Natural Questions dataset (NQ, Kwiatkowski et al., 2019) and human translates them into 25 additional languages and dialects. Accompanying these query translations we replace NQ's passage embedded answer spans with high-quality, language- and retrieval-independent answer annotations, linked directly against Wikidata entities and a limited set of well-defined value types (numbers, dates, strings, etc.).[2]

See one full example in Table 1. More flexible than existing multilingual datasets, MKQA's grading procedure ensures these labels are sufficient to evaluate any QA method, including knowledge graph and generative approaches. The objective of this evaluation set is to facilitate fair comparison between languages, without imposing assumptions on the underlying QA approach. We see MKQA as a useful tool enabling practitioners to benchmark a variety of multilingual open domain question answering methods against the widest range of available languages yet. Below, we discuss its central properties as an evaluation benchmark.

Realistic and Reliable Annotations Of crucial importance to any evaluation set is (a) how well it reflects realistic, real-world settings, and (b) the reliability of its annotations. To ensure the English queries, which form the basis of our dataset, are realistic, we use Natural Questions, formulated by real users, independent of passages or answers. To ensure these queries are realistic in other languages we employ expert bilingual translators, guided by strict localization criteria. We confirm that a large majority of these queries are **geographically invariant**, meaning that their answer is not culturally or geographically dependent (we found that less

---

[1] MKQA data and evaluation scripts are available at https://github.com/apple/ml-mkqa.

[2] Wikidata is a collaboratively edited open knowledge graph: https://www.wikidata.org/.

Language	Language Code	Question	Answer and Aliases	Answer Type	Entity ID
English		who produces the most pistachios in the world	Iran, Republic of Iran, ir, Islamic Republic of Iran, Persia, Islamic Rep. Iran	Entity	Q794
Arabic	ar-SA	من أكبر منتج للفستق في العالم	إيران، ايران، الجمهورية الإسلامية الإيرانية، ایران، الجمهورية الإسلامية في إيران، الجمهور الإسلامية الإيرانية، جمهورية إيران الإسلامية، جمهورية إيران الإسلامية، جمهورية إيران، الجمهورية الأسلامية الايرانية، جمهوريّة إيران الإسلامية، الجمهورية الإسلامية الايرانية، ايران الأكبر		
Danish	da-DK	hvem producere flest pistacienødder i verden	Iran		
German	de-{DE,,AT,CH}	Wer produziert weltweit die meisten Pistazien	Iran, Islamische Republik Iran		
Spanish	es-ES	¿Quién produce la mayor cantidad de pistachos del mundo?	Irán, República Islámica de Irán, Iran, Republica Islamica de Iran		
Finnish	fi-FI	Kuka tuottaa eniten pistaasipähkinöitä maailmassa	Iran		
French	fr-FR	Qui produit le plus de pistaches dans le monde	Iran, Perse, République islamique d'Iran		
Hebrew	he-IL	מי מייצר הכי הרבה פיסטוקים בעולם	איראן		
Hungarian	hu-HU	Ki termeli a legtöbb pisztáciát a világon?	Irán, Iráni Iszlám Köztársaság, Perzsia		
Italian	it-IT	Chi produce più pistacchi al mondo	Iran, Repubblica Islamica dell'Iran		
Japanese	ja-JP	世界で一番ピスタチオを生産しているのは誰ですか	イラン, イラン イスラム共和国		
Khmer	km-KH	អ្នកណាផលិត pistachios ច្រើនជាងគេបំផុតនៅលើពិភពលោក?	អ៊ីរ៉ង់		
Korean	ko-KR	전 세계에서 누가 가장 많은 피스타치오를 생산하나요	이란		
Malay	ms-MY	siapa menghasilkan pistachios paling banyak di dunia	Iran		
Dutch	nl-NL	wie produceert de meeste pistachio nootjes ter wereld	Iran, Islamitische Republiek Iran, IR, IRN		
Norwegian	nb-NO	hvem lager mest pistasjnøtter i verden	Iran		
Polish	pl-PL	kto produkuje najwięcej pistacji na świecie	Iran		
Portuguese	pt-BR	quem produz mais pistaches no mundo	Irão, ایران República Islâmica do Irão, Republica Islamica do Irao		
Russian	ru-RU	кто в мире производит больше всего фисташек	Иран, Исламская Республика Иран, Персия, ИРИ, Iran, Иран (Исламская Республика), Название Ирана		
Swedish	sv-SE	Vem producerar mest pistagenötter i världen?	Iran		
Thai	th-TH	ใครผลิตถั่วพิตาชิโอมากที่สุดในโลก	ประเทศอิหร่าน		
Turkish	tr-TR	Dünyanın en çok antep fıstığını kim üretiyor?	Iran, Iran İslam Cumhuriyeti		
Vietnamese	vi-VN	nước nào sản xuất nhiều quả hồ trăn nhất thế giới	Iran		
Chinese (Simplified)	zh-CN	谁生产了世界上最多的开心果	伊朗		
Chinese (Hong Kong)	zh-HK	誰是世界上開心果產量最高的人	伊朗, 伊朗伊斯蘭共和國		
Chinese (Traditional)	zh-TW	全世界開心果的主要生產者是誰	伊朗		

Table 1: **Questions and answers in all supported languages for one instance in MKQA.** The IETF BCP- 47 language codes specify the language and locale. The Entity ID corresponds to Wikidata (see for instance `https://www.wikidata.org/wiki/Q794`).

than 4% of answers are rendered incorrect by geographical and cultural context, for more details see Section 4.2). To ensure annotation reliability, we enforce minimum inter-grader agreement, conduct quality checks, and re-annotation from expert graders where necessary. Further, the Wikidata entity identifiers (QIDs) ground the answer annotations in structured data. This can be used for other knowledge graph-specific metrics, to retrieve other valid answer strings, and trivial entity translation into hundreds of languages beyond the scope of MKQA.

Parallel Questions Our evaluation set is fully aligned, or ''parallel'', across all available languages, meaning the same examples exist in all languages. This is accomplished by a mixture of expert human translation and using multilingual data from Wikidata. This property enables direct comparison between all 26 languages for fully cross-lingual or zero-shot systems. While Clark et al. (2020) point out the natural query distribution varies by language and geography, we reserve our assessment to geographically invariant queries for the purpose of more fair comparison between methods.

Retrieval-Independent Annotations Existing training and evaluation sets are oriented to ''extractive'' QA, providing specific passages and passage-dependent answer annotations (Clark et al., 2020; Lewis et al., 2020; Artetxe et al., 2020b; Liu et al., 2019a). These types of annotations are of limited use with varying retrieval systems, knowledge graph approaches, and even generative approaches because the answers are tied to the particular phrasing of their passage. Translating annotations from English passages may also introduce ''translationese artifacts'' as the translation is implicitly influenced by the original English structure (Artetxe et al., 2020a). These artifacts render the task easier for methods relying on English supervision or machine translation techniques. As we shall discuss in Section 3, the MKQA collection procedure yields primarily entity and structured ''atomic'' answer types. We contend retrieval-independent (and particularly entity-oriented) annotations minimize the risk of translation artifacts, and remove limitations on the underlying QA approach.

Linguistic Diversity Lastly, MKQA has broad linguistic diversity, covering 26 languages and dialects from 14 language family branches. Languages from MKQA cover half of the world populations' native language, and more than 90% of the world population lives in a country where one of these languages is an official language (see Section 4.1 for more details). It is to our knowledge both the largest and most linguistically diverse open-domain QA evaluation set currently available (see Table 2 and 3).

Multilingual QA Evaluation Set	Answer Independence	Parallel Questions	Language Fam. Branches	Languages	Total Examples
XQA (Liu et al., 2019a)	✓	✗	5	9	28k
MLQA (Lewis et al., 2020)	✗	✓	6	7	46k
XQuAD (Artetxe et al., 2020b)	✗	✓	11	11	13k
TyDi (Clark et al., 2020)	✗	✗	11	11	204k
Xor-QA (Asai et al., 2021)	✗	✗	7	7	40k
MKQA (This work)	✓	✓	14	26	260k

Table 2: **Comparison of multilingual QA evaluation sets.** Answer independence indicates whether the gold answer is independent of a retrieved document, and parallel questions indicates whether examples are the same across languages.

MKQA makes two important contributions to the field of multilingual question answering:

- Our *answer collection procedure* renders the evaluation set highly reliable, independent, and unbiased towards the QA technique used. This unique setup allows us to fairly compare the performance of techniques as distinct as knowledge graph-based, dense and sparse retrieval and generative QA techniques on a large number of languages (see Section 5).

- Our dataset provides *fully aligned* examples in the largest yet number of typologically diverse languages, enabling comparable evaluation across many languages.

We find MKQA is innately more challenging than Natural Questions from which it was derived, due to the multi-stage re-annotation process. The best model obtains only 52.3% F1 in English, and only 5.7% above a naive baseline on the lowest resource language. Given these qualities, our dataset facilitates broad and reliable evaluation of multilingual, open-domain question answering.

2 Related Work

Cross-Lingual Modeling Recent work trains cross-lingual representations with unsupervised language modeling over many languages, including Multilingual BERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), and Multilingual T5 (Xue et al., 2021). Transfer learning techniques are often applied to these cross-lingual representations to overcome the dearth of non-English data (Cui et al., 2019a; Hsu et al., 2019; Lee and Lee, 2019; Kumar et al., 2019). Recent investigations into cross-lingual modeling have revealed "translation artifacts" in datasets where

machine translation systems are used, or human translation tasks are not carefully curated (Artetxe et al., 2020a; Wintner, 2016; Rabinovich and Wintner, 2015). "Translationese" results in hidden linguistic cues in translated text that render the task easier than a natural translation.

English QA Resources A majority of question answering research focuses on English, which offers ample selection of evaluation datasets, including SQuAD (Rajpurkar et al., 2016), Trivia-QA (Joshi et al., 2017), and Natural Questions (Kwiatkowski et al., 2019). Open Domain QA, pioneered by Green et al. (1986), is the task of answering open questions using external knowledge sources. A common approach is to combine retrieval and extractive techniques (Chen et al., 2016, 2017; Dhingra et al., 2017; Cui et al., 2017).

Monolingual QA Resources Non-English question answering resource options remain comparatively rare, with most options spanning only one other language, and rarely low-resource languages. DuReader (He et al., 2018), CMRC (Cui et al., 2019b), and DRCD (Shao et al., 2018) all offer high-quality Chinese QA datsets. Similarly, XCMRC (Liu et al., 2019b) and BiPar (Jing et al., 2019) present parallel, cross-lingual QA datasets between English and Chinese. Exploring slightly less resource-rich languages, numerous works have derived new datasets from SQuAD, employing varying degrees of human or semi-automatic translation techniques to non-English target languages: ARCD for Arabic (Mozannar et al., 2019), KorQuAD-1.0 for Korean (Lim et al., 2019), and MMQA for Hindi (Gupta et al., 2018).

Multilingual QA Resources Table 2 compares the largest publicly available multilingual question answering evaluation sets. The table highlights the

following properties of each dataset: whether the available gold answers are independent of retrieved documents, whether examples are aligned across languages, and the number of languages and examples provided. MLQA (Lewis et al., 2020) and XQuAD (Artetxe et al., 2020b) are examples of SQuAD-style extractive datasets, employing human translators to create parallel examples. Both MLQA and XQuAD ensure that all answers are answerable (discarding ''No Answer'' examples), and derive answers from provided documents. XQA (Liu et al., 2019a), one of the few retrieval-independent QA datasets, offers cloze-style questions, leveraging Wikipedia's daily questions and entity answers to populate document-independent answers. TyDi (Clark et al., 2020), like MKQA, focuses on typological diversity in its wide language selection. While TyDi offers a more natural distribution of questions, its annotations are based on the retrieval system used by the authors (Google search); hence their answers are actually start and end indices for spans of text within a given passage. Xor-QA (Asai et al., 2021) explores cross-lingual subtasks by re-annotating 40k TyDi examples, over 7 languages, sourcing answers from English documents and translating them back to the target language. Many of these multilingual resources have been bundled into cross-lingual benchmarks, such as XTREME (Hu et al., 2020) and XGLUE (Liang et al., 2020).

2.1 Comparison to Native Speaker Datasets

There are key advantages to datasets such as TyDi (Clark et al., 2020) and Xor-QA (Asai et al., 2021), which use native speakers questions, particularly in the naturalness and cultural authenticity of the corpora. However, there are also key disadvantages to these datasets that MKQA circumvents with language alignment, to provide more challenging and fair model evaluations across languages.

TyDi (Clark et al., 2020) and MKQA both target high typological diversity, highlight the importance of sourcing realistic questions (with answers unseen), and incorporate a broader distribution of question types than competing datasets (including ''No Answer'' and ''Yes''/''No'' answers). There are three main differences between MKQA and TyDi: (a) question alignment across languages, (b) answer distribution, and (c) anno-

tation retrieval independence (closely tied with the notions of ''open'' and ''closed'' domain). TyDi provides a different set of natural questions per language, at the expense of direct comparability across languages. Not only are the TyDi questions different between languages, but the percentage of answerable passages varies dramatically, from 22% in Korean to 69% in Arabic. XorQA-TyDi (Asai et al., 2021) partially resolves this issue by sourcing answers from English documents, but this may in turn re-introduce cultural biases. This suggests that the conceptual difficulty of these questions may also vary dramatically, as consumers from different locales cater their questions based on their existing beliefs of the quality of the virtual assistants in their language. As a result, it is difficult to interpret the core reasons why multilingual system's performance varies between languages. To ensure this property, MKQA verifies its questions are predominantly geographically invariant, and thus the answers will not change due to geographical or cultural factors.

The second difference between datasets is the answer distribution. MKQA answers (a) are predominantly entities (42.2%) or atomic answers such as dates, binary, or numbers with units, and (b) use a different definition of ''Unanswerable''. Xor-QA focuses only on answerable queries, TyDi's definition conditions on the presence of the answer in the passage, whereas MKQA's definition is based on the ability of a human to find a succinct answer to a question on the web, that is, whether it is human answerable. As a result, our annotations are not limited by the quality of selected passages, and provide higher answer coverage (67.58% as opposed to the TyDi language average of 38%).

Finally, while MKQA does not expect an answer to be derived from a single source document, TyDi is an extractive QA dataset. Consequently, its answer annotations are defined as spans, tied directly to particular Wikipedia documents and fixed index from which they were retrieved. As an evaluation set we contend the flexibility of document-independent answers is critical to not restrain what approaches can be evaluated in future research.

3 Dataset Collection

We aim for certain properties of our evaluation set: (i) realistic questions, (ii) reliable annotations

(e.g., via inter-annotator agreement), and (iii) a flexible task setup that makes as few assumptions as possible about the underlying modeling techniques, enabling fair comparison between any approach.

3.1 Query Selection

Our evaluation set collection pipeline begins with the Answer Curation steps outlined in Figure 1. These are designed to yield high-concensus answer labels, with normalized textual formats, expressive alias sets for robust comparison, and grounding in structured information for entity disambiguation or more informative analysis. For the first step, we sample 10,000 queries from Natural Questions (NQ) (Kwiatkowski et al., 2019), as this is one of the few QA datasets based on realistic queries, generated by information seeking users.

3.2 Raw Answer Collection

At the raw answer collection stage, 5 annotators are independently shown the query and asked to search the web to either copy *or generate* an ideal answer. They are asked to select an answer type (radio buttons) from the options shown below, and input the answer (text box) according to format instructions per answer type. The formatting constraints allow us to automatically link WikiData entities for the units in ''number with units'' and to gather well-structured data for answers such as dates, to save annotator time.

For each query, the graders select a typed answer from the following taxonomy:

- **Atomic value:** This category includes dates, numbers and number ranges with or without a unit (meters, years, . . .).

- **Entities:** Entities are annotated with Wikidata QIDs and include generic entities, people, objects, and most locations.

- **Yes/No:** Type representing yes/no answers.

- **Short answer:** Answers which cannot be encapsulated in an atomic value, entity or binary (yes/no) answer, but are still a short phrase.

- **Long answer:** The long answer category indicates no simple factual answer or short phrase answers this question and a longer or visual explanation is required. During evaluation we treat these as ''Unanswerable'' for simplicity.

- **Unanswerable:** This category indicates that the query is not answerable, potentially because it is ill-formed or because no clear answer is available.

3.3 Answer Resolution

Given the query and a candidate answer from the previous stage, annotators are next asked to normalize date/number formats and resolve the answer text against Wikidata entities, where feasible. To resolve short textual answers against Wikidata entities, we apply an internal entity linking system to the answer string to generate Wikidata candidate entities.[3] The top 10 entity suggestions and their descriptions, along with the original query and short answer are then presented to 3 graders, who are asked to pick the correct reference entity or ''None of the above.'' In cases where graders do not achieve sufficient agreement or where the correct entity is not in the list, a domain expert (one of the MKQA authors/designers) provides the correct reference. Overall, this step enables us to disambiguate homonyms and collect valid answer synonyms/aliases, for more robustly measuring annotator agreement and prediction accuracy.

3.4 Answer Verification

Up until this stage, 5 raw answers were collected per query, and subsequently format normalized and resolved against Wikidata. In the fourth stage of Answer Curation (in Figure 1) any normalized answer given by at least 2 annotators is admitted to the final set as a gold answer. For those annotations that did not achieve the required agreement from at least two annotators, a domain expert (one of the MKQA authors/designers) with access to all 5 preliminary annotations is tasked to provide a final decision. This second manual round was afforded as much time per decision as necessary to obtain a satisfactory answer. The instructions permit the selection of existing normalized answer(s), modifying them slightly, or overriding them if necessary.

3.5 Answer Localization

In the last two stages of MKQA curation shown in Figure 1 we translate, or ''localize'', the English queries and answers into the target languages. Given the special care we took to avoid them in

---

[3]This step can be replicated using an off-the-shelf entity linker such as spaCy available at `https://spacy.io/api/entitylinker`.
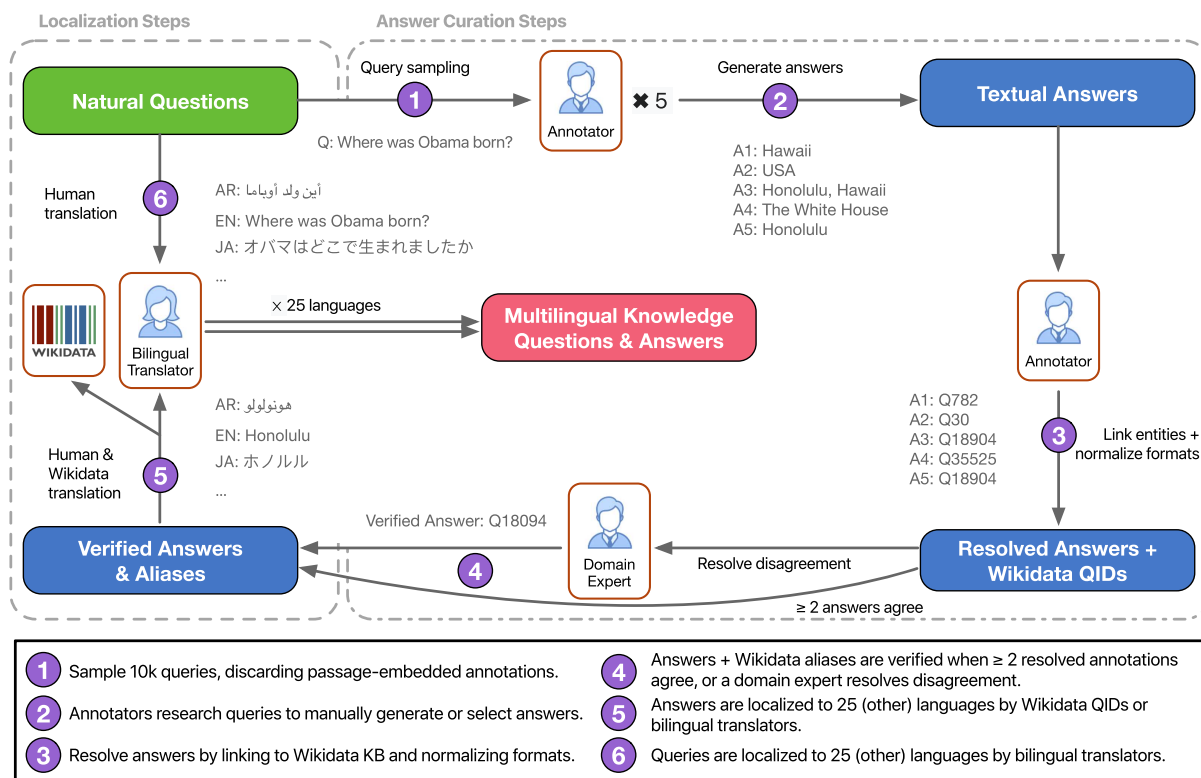
Figure 1: **Data Collection Process.** A depiction of the 6 sequential steps in our data collection pipeline. The first four steps involve Answer Curation, and the last two localize questions and answers into 26 target languages.

our methodology, and since we only localize short answers and queries (no context passages), we believe translation artifacts are likely to be minimal in MKQA.

Verified answers are localized into the target language by a combination of methods. For Wikidata-resolved answers, we leverage Wikidata's names and aliases for the target language. These names and aliases are transcribed in the native alphabet where appropriate, reflecting the expected answer in each language. Atomic answer types, including numeric, number with entity, and date types were also translated by this method, maintaining Arabic numerals for all languages, but naturalizing unit terms such as "November", "century", "b.c", "acres", and "light years". For date types specifically, for every combination of year, month, and day, we generate template answers in each language, accommodating both American and European date formats, as well as numeric and written out versions for months.

In cases where a Wikidata link could not be found, or where answers were not available for a given language code, professional bilingual human translators were used to provide the native equivalent. For this task, human translators are given access to the English query, the English answer, and where available the Wikidata link and Wikipedia page for the entity. We found localization quality improved when bilingual translators are shown several examples prior to grading, covering each of the localization options:

Localization Options:

- **Transliteration** is a type of conversion of a text from one script to another that involves swapping letters (thus trans- + liter-) in predictable ways (such as $\alpha \rightarrow$ a, $\chi \rightarrow$ ch, or æ$\rightarrow$ ae).

- **Translation** is the communication of the meaning of a source-language text by means of an equivalent target-language text.

- **Unchanged** is selected if the entity name does not need to be localized as it is commonly used as is.

- **Mix transliteration/translation/unchanged** if the entity is localized using more than one technique.

1394

3.6 Query Localization

The final stage of MKQA construction, as shown in Figure 1, is query localization. As with answer localization, bilingual translators were asked to translate each query ensuring the query's meaning is maximally preserved, while naturally phrased. Translators were further instructed to use localized names of named entities if they exist in the target language and to transliterate names otherwise. Our translators, who are native speakers of the target language, are verified to live in the targeted region and are required to pass an entrance exam to verify a high level of fluency in English. Translators received a standard hourly wage varying with the target region and were not compensated per completed task, as is usual with alternative public services such as Amazon Mechanical Turk. On average, around 16 translators participated in the translation of the 10k source queries from English into each target language.

4 Dataset Quality and Analysis

Given our dataset collection and methodology, we evaluate the effect of our choices, and the properties of the final set, including the selected languages, annotation quality, geographical invariance, and answer type distribution as compared to NQ.

4.1 Language Selection

We select a set of languages meeting both academic and practical considerations, by maximizing typological diversity as well as the share of the world population that understand at least one of the languages in the set. Table 3 shows the languages selected for our dataset with the corresponding branch of their language family. We also show the language's reach, that is, the percentage of the world population that speaks the language either as a first or second language (based on Ethnologue data, Simons and Fennig, 2018). Since combined first- and second-language speaker statistics are not readily available, it is not straight-forward to accurately determine what share of the world population can be covered by the languages in this set (e.g., a native speaker of German may also be fluent in English). A practical option is to calculate the share of the world population that lives in a country where one of the languages in our set is recognized as an official language. By this measure, 90.62% of the world population

Family	Branch	Language	Reach
Indo-European	Germanic	English	16.46%
		German	1.70%
		Dutch	0.38%
		Swedish	0.17%
		Danish	0.08%
		Norwegian	0.07%
	Italic	Spanish	6.99%
		French	3.59%
		Portuguese	3.28%
		Italian	0.87%
	Balto-Slavic	Russian	3.35%
		Polish	0.58%
Sino-Tibetan	Sinitic	Mandarin	14.54%
		Cantonese	1.10%
Afro-Asiatic	Semitic	Arabic	4.44%
		Hebrew	0.12%
Austronesian	Malayo-Poly.	Malay	3.47%
Japonic	Japonic	Japanese	1.64%
Austroasiatic	Vietic	Vietnamese	1.00%
	Khmer	Khmer	0.21%
Turkic	Com. Turkic	Turkish	1.10%
Kra–Dai	Tai	Thai	0.78%
Koreanic	Han	Korean	1.03%
Uralic	Finnic	Finnish	0.07%
	Ugric	Hungarian	0.17%

Table 3: **Languages with their corresponding language families and speakers.** *Reach* indicates the combined number of first-language (L1) and second-language (L2) speakers as a percentage of the world population (Ethnologue, Simons and Fennig, 2018).

live in a country with an official language covered by the languages in our set.[4] With the large number of diverse language families covered and the reach of the selected languages, MKQA addresses both academic and practical requirements for a wide and diverse question answering benchmark. Finally, we note that the Wikidata IDs provided for a large portion of our gold answers allow these answers to be further localized into Wikipedia languages beyond those in MKQA, should practitioners wish to expand their analysis.

---

[4]We determine this percentage based on Wikidata as the combined population (Wikidata property ''P1082'') of all countries that have an official language (Wikidata property ''P37'') in our dataset divided by the combined population of all countries in Wikidata.

| Language | Acceptance Rate | |
	Query Translation	Answer
English	–	97.03%
German	99.01%	91.08%
Spanish	99.01%	92.07%
Thai	96.04%	91.09%
Chinese (simpl.)	92.24%	89.32%

Table 4: **Query translation and retrieval-agnostic answer quality in various languages.** Query translation acceptance rate is the percentage of query translations judged as acceptable. Answer acceptance rates is the percentage of answers graders found acceptable in response to the translated target-language query.

4.2 Translation and Answer Quality

The quality and reliability of our dataset is highly dependent on two factors: (a) how well our professional translators were able to translate the English queries into each target language, and (b) how well our language-independent answer representations transfer to each target language.

We run a small-scale grading experiment, grading just above 1% of the total data, to estimate the quality of the query translations and how well the meaning of our language-independent answer annotation is preserved across languages (geographical invariance). We present graders with the localized query and its answer annotations and ask them to judge whether (a) the localized query is an acceptable translation of the original English query, and (b) whether the provided answer (entities are shown with their QID and description, and a short explanation is added to each other answer type) is acceptable for the translated target-language query. In addition, we also ask graders to judge the answer quality for the original English queries as a baseline.

Table 4 shows the acceptance rates for query translations and answers for a small selection of languages. The table shows that query translations are consistently judged as acceptable in German, Spanish, and Thai, while the quality for Chinese translations was judged as lower in comparison. Most translation issues are related to the localization of entities and to domain-specific terms (e.g., sports terminology such as ''receptions'' in football). As expected, the acceptability

of answers is judged to be higher for English than other languages but it is still at or above 90% even for languages as linguistically distant from English as Thai. Note that errors in answer acceptance rate and query translation acceptance rate heavily overlap since incorrect query translations will most likely mean that the existing language-independent answer will not match. Answer quality issues fall into the following categories (illustrated with German examples):

(1) Answer differs based on cultural context (44%) This includes cases where the localized version of an entity may have different properties. For example the English-language TV show ''Man vs Food'' has 8 seasons while the German version has 5. Similarly, a character in a movie such as ''Finding Nemo'' may be voiced by a different voice actor in the German version of the same movie.

(2) Generic annotation issues (33%) The second biggest source of errors are answer quality issues that will hold across languages. Examples include answers that are time-sensitive such as the answer to the question ''when was the oldest person in the world born'' and questions with ambiguous answers in the data such as ''is northern ireland a part of great britain.''

(3) Entities transliterated incorrectly (11%) Names for entities may be transliterated incorrectly if they do not exist in the target language (''who wrote the book clear and present danger'').

(4) Generic translation artifacts (11%) Generic translation errors may lead to a mismatch between the question and the language-independent answer. In one example the English ''words to'' meaning ''lyrics'' was translated into German as the literal ''Worte'' which would be an uncommon phrasing in a question about lyrics.

Translation artifacts are a recognized problem in multilingual datasets and manual grading of the data in Table 4 shows that the human translation step may introduce more or less query–answer discrepancies depending on the target language. In an alternative scenario, annotation could be performed directly on native queries from each language; however, such data is not readily available and might additionally suffer from other downsides such as relatively small user bases in less frequently spoken languages (see Section 2.1 for further discussion). Similar to our evaluation, the authors of NQ perform a manual precision
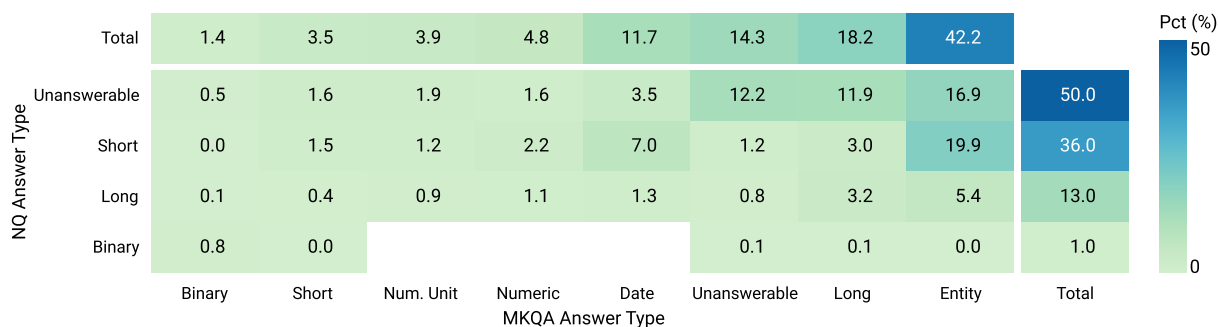
NQ Answer Type	Binary	Short	Num. Unit	Numeric	Date	Unanswerable	Long	Entity	Total
Total	1.4	3.5	3.9	4.8	11.7	14.3	18.2	42.2	
Unanswerable	0.5	1.6	1.9	1.6	3.5	12.2	11.9	16.9	50.0
Short	0.0	1.5	1.2	2.2	7.0	1.2	3.0	19.9	36.0
Long	0.1	0.4	0.9	1.1	1.3	0.8	3.2	5.4	13.0
Binary	0.8	0.0				0.1	0.1	0.0	1.0

MKQA Answer Type

Figure 2: **Answer Type Breakdown.** Compares the distribution of answer types between MKQA and Natural Questions (NQ) for the $10k$ examples in the evaluation set.

grading of their data and find an overall data precision of 84% for short answers. While we hope that future work can improve on data quality further, comparatively even for the language with the most severe translation artifacts in our evaluation, Simplified Chinese, the resulting data quality (answer acceptance rate of 89%) is still within an acceptable range. In addition, our dataset provides the only available source of question answering evaluation in many languages.

We encourage authors of future multilingual datasets that use any translation methods to report and detail their geographical invariance, as we have done, and to benchmark the reliability of examples and presence of translation artifacts.

4.3 Annotation Breakdowns

Next, we compare the distribution of answer types between the original NQ dataset, with those newly assigned in MKQA. As Figure 2 shows, 50% of NQ are completely ''Unanswerable'' by retrieved passages and another 13% require long passage answers. In the short answer setup for NQ both of these are considered unanswerable, amounting to 63% of all questions. In comparison, only 32.4% of examples are ''Unanswerable'' or ''Long'' answer type in MKQA. This is due to a shift in definition from whether a passage contains an answer, to whether a question is (succinctly) answerable by a human, with full web access. Given that the answer types in MKQA are not dependent on a learned retrieval system, they reflect the properties of the question only.

We later show that this ''unanswerable'' definition yields more challenging evaluation because (i) correctly answering questions is on average harder than learning when to abstain, and (ii) many

of the most difficult questions were unanswerable in NQ but are answerable in MKQA. This suggests the property of ''retrieval independent annotations'', currently not used in any other multilingual QA benchmarks except XQA, is highly desirable for (a) constructing more challenging QA evaluation sets, and (b) yielding annotations useful to evaluate any QA approach, not just extractive QA models.

We also encourage future QA benchmarks to mimic our multi-stage data collection framework in providing supplementary metadata per example (answer type and Wikidata QIDs). Beyond basic comparison of systems, our evaluation tools allow practitioners to perform further error analysis with more interpretable metrics.

5 Experiments

5.1 Task Definition

Given a question q^l in language l, the task is to produce a prediction $p^l \in \{$No Answer, Text Answer$\}$, where a Text Answer is a sequence of tokens in the corresponding language. p^l can be obtained by any method, extracted from a document, generated, or derived from a knowledge graph.

For evaluation using MKQA gold answers, every question q_i^l from $i \in [1, 10000]$ is accompanied by a set of valid annotations a_i^l per language. Every prediction p_i^l is scored based on exact match (EM) and token overlap F1, as with previous open-retrieval QA datasets. The official evaluation script also ingests a ''No Answer probability'' for each example. If the probability is above a chosen threshold value then the prediction defaults to No Answer instead of the provided Textual Answer. As this threshold varies from 0 to 1 the

predictions shift from entirely No Answer to all textual answers. We follow NQ in reporting the best F1 over the range of thresholds, to remove threshold tuning as a factor in evaluation. A best threshold is computed and applied per language, where each example receives a ''textual" (token overlap) F1 after language-specific normalization (removing whitespace, punctuation, and articles) is applied to both the prediction and gold answers. Finally, the official per-language F1 is computed as the mean of example F1s, and the official Macro Average F1 is the mean of per-language F1 scores.

5.2 Baseline Approaches

To benchmark our evaluation set, we combine state-of-the-art approaches in retrieval, machine translation, extractive QA, and generative QA. All retriever models are off-the-shelf, and all reader models are finetuned on Natural Questions, including XLM-ROBERTA LARGE (Conneau et al., 2020) and M-BERT (Devlin et al., 2019) for extractive QA, and mT5-LARGE (Xue et al., 2021) for generative QA.[5] In each case, tokenization is handled by the multilingual model used—sentencepiece for XLM-R and mT5-LARGE, WordPiece for M-BERT, each with vocabularies initialized from their specific pre-training implementations. Further, all query and prediction translations in our approaches use Zhang et al.'s (2020) open source many-to-many, encoder-decoder machine translation system, trained on the OPUS multilingual corpus, covering 100 languages.

Retrieval Corpora Our baselines operate on a Wikipedia document corpus from December 07, 2020, following previous work in open-domain question answering (Kwiatkowski et al., 2019; Asai et al., 2021; Clark et al., 2020). We use the language-specific Wikipedia corpora for Elasticsearch and the English versions for other baselines. Using Wikipedia as this base corpus is a pragmatic choice based on several aspects: 1) It provides comparability across baselines and previous work, and 2) compared to large web document corpora, such as Common Crawl, it requires less data cleaning and is computationally more tractable, which improves the replicability of our results and helps to ensure that the major variable being evaluated is model performance (rather than engineering effort). Hence, while we believe that

using a web-scale corpus, such as Common Crawl, would potentially enable even stronger baselines, we leave such experiments to future work.

Elasticsearch → XLM-R We benchmark a fully multilingual retriever approach using Elasticsearch followed by XLM-R as the extractive reader. Elasticsearch leverages language-specific tokenizers and analyzers with BM25 to search for native passages in the target language's Wikipedia dump. We used their built in language specific analyzers which include stopwords and stemmer in each language.[6] We took the Wikipedia dump from December 7, 2020, for each language as source documents. The languages Hebrew, Khmer, Korean, Malay, and Vietnamese are not part of the Elasticsearch baseline as they are not natively supported by Elasticsearch.

DPR → RoBERTa We benchmark an approach that utilizes state-of-the-art English retrieval and reader systems, enabled by translating the incoming query into English, and the outgoing prediction into the target language. We use off-the-shelf Dense Passage Retrieval (DPR, Karpukhin et al., 2020), followed by RoBERTa (Liu et al., 2019c) to extract a prediction.[7]

Gold NQ → Extractive QA For this set of baselines, optimal English retrieval is simulated via the passages provided with NQ. We illustrate baselines that leverage these provided ''Gold" English documents, machine translation, and extractive QA models. We vary the type of QA model (M-BERT vs. XLM-R) and the train/test approach, comparing common **zero shot**, **translate test**, and **translate train** approaches.

In **zero shot** transfer each multilingual model is finetuned with NQs' default English questions Q_{en} and passages P_{en}. At test time the model receives MKQA questions Q_{xx} in language xx, paired with English passages P_{en}.

For **translate test**, at train time the model uses NQ's default English. At test time, MKQA questions are translated into English $Q_{xx \to en}$, and the passage remains in English P_{en}. Passages remain in English for both training and inference.

---

[5]Note that we exclude the $10k$ examples used in our evaluation set from this training set.

[6]https://www.elastic.co/guide/en/elasticsearch/reference/current/analysis-lang-analyzer.html#arabic-analyzer.

[7]We use the trained ''Multiset" DPR model available in https://github.com/facebookresearch/DPR.

Retriever	Reader	Translation		Retrieval Metrics	Answerable Metrics		End-to-End Metrics	
		Query	Answer	R@1	Mean A $\in$ D F1	Mean A $\notin$ D F1	En F1	Mean F1
No Answer	–	–	–	–	–	–	32.4	32.4
MULTILINGUAL RETRIEVER								
Elasticsearch*	Xlm-R	–	–	42.57 ± 1.2	25.18 ± 3.8	7.24 ± 2.5	34.99	34.13± 0.4
TRANSLATE-TEST ENGLISH RETRIEVER								
DPR	RoBERTa	Test	Test	53.62 ± 2.2	20.33 ± 4.1	10.24 ± 1.8	45.19	36.81± 1.2
GOLD NQ PASSAGES								
Gold NQ	M-Bert	–			20.13 ± 5.5	7.56 ± 1.7		37.8± 2.0
Gold NQ	M-Bert	Test			28.10 ± 6.5	12.1 ± 2.1	51.97	41.4± 2.2
Gold NQ	M-Bert	Train	Test	80.22	<u>32.21</u> ± 6.0	<u>14.8</u> ± 1.9		<u>44.1</u>± 1.8
Gold NQ	Xlm-R	–			38.81 ± 3.2	20.05 ± 2.6		45.5± 1.4
Gold NQ	Xlm-R	Test			34.23 ± 5.0	16.38 ± 2.6	52.27	42.9± 2.1
Gold NQ	Xlm-R	Train			**40.28** ± 3.1	**20.93** ± 2.7		**46.0**± 1.4
GENERATIVE MODELS								
Query-only	mT5	–	–	–	–	–	43.8	35.0± 1.2
Gold NQ	mT5	–	–	80.22	36.8 ± 6.2	17.07 ± 2.6	47.6	38.5± 2.2

Table 5: **Results for each baseline, broken down by retrieval metrics (Recall @ K passages), answerable question metrics (F1 at the best confidence threshold), and end-to-end metrics (F1 at the best confidence threshold).** A naive approach, predicting exclusively No Answer, achieves a lower bound score of 32.42% F1. Translate-Train using NQs Gold passages and an Xlm-R reader outperforms all alternate settings. $A \in D$ denotes metrics for where the answer A exists in the top retrieved document D (exact match). $A \notin D$ denotes metrics for where the answer A does not exist in top retrieved document D (exact match). * Elasticsearch benchmark does not include Hebrew, Khmer, Korean, Malay, and Vietnamese.

For **translate train**, at train time, questions are translated into the target language $Q_{en \to xx}$. At test time the model is given queries in the target language Q_{xx} and passages P_{en} in the default English from NQ. Passages are always in English.

Query-only mT5 We benchmark a ''closed-book'', query-only generative QA approach, based on Roberts et al. (2020). This approach allows us to circumvent retrieval and machine translation entirely, using parametric knowledge within mT5 Large. Simply, the query is fed to the model, which is trained to generate the localized answer directly.

Gold NQ → mT5 We benchmark a stronger generative QA approach, that also has access to the English Gold NQ passages. Based on open-source implementations for MLQA and XQuAD datasets, the model is fed the non-English query, with (in this case) the English gold passage, and generates the predicted answer.[8]

5.3 Results

Table 5 presents retrieval and end-to-end metrics for each baseline, as the mean across all 26 languages. Retrieval metrics include recall at K, measuring if the correct answer appears anywhere in the top K retrieved passages, as traditionally used in information retrieval settings. Note that these metrics are computed by looking for an exact match of the text-normalized gold answer in the text-normalized passage. We find that translation followed by English DPR outperforms the Elasticsearch multilingual sparse retrievers. This is consistent with results observed in XOR-QA (Asai et al., 2021) which shows the surprising under-performance of multilingual retrievers. Errors are likely a combination of no answer being present in smaller non-English Wikipedia indexes, and the weak performance of sparse retrieval. The Gold NQ documents contain a valid answer 80.22% of the time. However, this is likely an upper bound, as these documents are often very long and noisy, such that NQ annotators often marked them as not containing an answer to the question, even though we find the gold answer string is present.

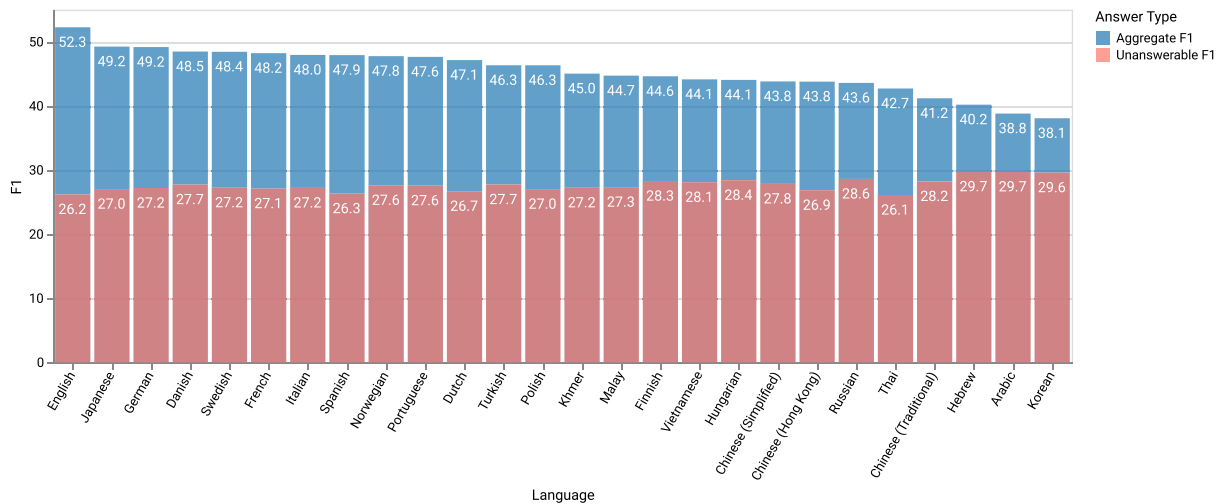For end-to-end metrics, we measure F1 just for English (''EN F1''), which omits the impact

Figure 3: **F1 by Language.** XLM-R Zero-Shot performance ranked by language. Unanswerable F1 (in red) corresponds to the proportion of the Aggregate F1 obtained from predicting No Answer. The Unanswerable proportion is calculated as the percentage of unanswerable examples (32.42%) multiplied by the Unanswerable F1.

of machine translation, and mean F1 over all 26 languages. The naive baseline of only predicting No Answer achieves a lower bound score of 32.42%. We chose to combine both Unanswerable and Long Answers into the No Answer category for evaluation to focus MKQA on short, factoid answers that can be evaluated automatically and robustly. Unsurprisingly, we observe models with access to NQ gold documents achieve the best results, with Translate Train XLM-R achieving the best mean F1 of 46.0±1.4. Among these methods, XLM-R outperforms M-Bert, and Translate-Train outperforms Translate-Test and Zero Shot. Generative approaches using MT5 perform fairly well, even under zero shot conditions (trained only on English), or without any passage provided (query-only).

We also measure the F1 scores for the subset of answerable questions to measure the ability of the retrievers and readers to find the right answer. We separately report the average all-language F1 for (i) questions in which a gold answer appears in the top retrieved document, and (ii) questions in which none are found. As expected, performance is much higher for both extractive and generative models where the retriever has succeeded. Translate Train with XLM-R still achieves the best performance. XLM-R also performs well on the correct outputs $(\mathbf{A} \in \mathbf{D})$ of the weakest retriever, Elasticsearch, though there are fewer of them. Comparing with end-to-end metrics, which includes unanswerable questions, answerable questions are more difficult to answer.

Overall, these results show how collecting relevant passages remains a challenging bottleneck in multilingual open-retrieval QA. Multilingual retrievers, English state-of-the-art retrievers, and generative QA models all fail to overcome this problem, and even when gold passages are provided, multilingual readers and machine translation still fail to consistently produce localized answers (with generous evaluation settings).

In Figure 3 we compare cross-lingual performance between languages, ranked by F1 score. We plot XLM-R Zero Shot to minimize the noise from machine translation. As expected, the XLM-R model performs fairly well on English (52.3), and common non-English languages, including the most common Indo-European Germanic and Italic languages, but poorly on languages from lower-resourced families. Note that the minimum F1 score is 32.42%, where a threshold of 0 predicts No Answer to every question. Interestingly, as the Aggregate F1 decreases, the Unanswerable F1 rises on average from ∼27% to ∼29%, abstaining from an answer more often. Given the parallel questions property of MKQA, these metrics allow a practitioner to specifically identify languages with weak model performance, and answer abstention behavior for commonly used reader models, such as XLM-R. Even before considering a cultural shift in query distribution, these metrics allow us to isolate performance on geographically invariant queries, and general effectiveness of transfer learning for particular languages and training regimes.

5.4 Unanswerable vs. Long Answers

As discussed in Section 4.3, following the Short Answer setup for Natural Questions (Kwiatkowski et al., 2019) we define Unanswerable as a query without a short answer (i.e., examples with *long* or *unanswerable* answer types)—for our task. Although evaluating long answers is important, it is out of the scope of MKQA. The primary benefit of this decision is that it enforces the *retrieval-independent annotations* property of MKQA, since long answers have an unbounded number of correct answer strings. Here we investigate whether *long* and ''truly'' *unanswerable* examples in MKQA are treated differently by our baseline models.

To answer this question, we break down the larger Unanswerable set into the *long* and 'truly' *unanswerable* examples, comprising 56% and 44% respectively. We then compute the final performance (F1) by model type and by language for each of these two categories. We find the results vary according to the quality of the model and the language (as do performance on answerable queries), but the difference between the long answer and truly unanswerable scores are marginal. For instance, XLM-R Translate Train, using Gold NQ passages, achieves 84.2% F1 on *long*, and 84.7% on truly *unanswerable* examples, with a mean difference over all 26 languages of only 0.5%. These differences are similarly negligible across other baselines. This finding suggests standard open-domain QA systems, trained on short answer datasets like Natural Questions, have learned to consider long answers as unanswerable, and do not appear to find one set more challenging than the other.

6 Discussion

Difficulty of MKQA Our baselines represent a strong and diverse set of methods, that score competitively with state-of-the-art on similar open domain question answering datasets. Nonetheless, on English alone, the best system recieves an F1 score of only 52.3%, less than the same methods achieve on the open datasets Natural Questions and TriviaQA, or other standard benchmarks for this task. These comparative results demonstrate MKQA is highly challenging and leaves ample room for improvement in both English and the long tail of natural languages. In this section we

explain why, with a detailed comparison to its closest set, Natural Questions.

Why is MKQA so challenging for state-of-the-art approaches even for English open-domain QA? To shed light on this, we compare the difficulty of English-only annotations between Natural Questions (NQ) and MKQA. In Figure 4 we use the same BERT-LARGE English model (trained on NQ, using Gold NQ passages) and evaluate it on both sets of annotations. The ''F1 by Answer Type'' diagram shows unanswerable examples in MKQA (red line) are easier than the unanswerable examples in NQ (red dashed line), as the model maintains higher performance at all No Answer confidence thresholds. The opposite relationship is observed for answerable examples.

We hypothesize that this is due to the *Retrieval-Independence* property and high coverage of our re-annotation process (described in Section 3). Due to the annotation procedures NQ uses, there are several cases that can lead to a potential answer missing from the dataset: (a) the initial retrieval may have not produced a candidate, (b) the answer may have not been in Wikipedia, or (c) NQ graders may have missed a valid answer. MKQA annotations are not susceptible to (a) and (b) and likely less impacted by (c). Consequently, the most challenging questions migrated from unanswerable in NQ to answerable in MKQA, shifting the unanswerable distribution from 63% to 32% (as shown in Figure 2). Consider the following examples.

(a) NQ retrieval failure In this example, the NQ retrieved document does not contain an answer to the question, causing no long or short answer (*No Answer*) in NQ. There exists a better Wikipedia document (Wheel of Fortune) that does contain the MKQA answer ''Autumn Erhard''.

- **Q:** *Who won the most money on wheel of fortune?*

- **NQ URL:** Wikipedia: *American game show winnings records.*

- **NQ Answer:** *No Answer*

- **MKQA Answers:** *''Autumn Erhard''*

(b) No Wikipedia answer This is also an answerable query, labelled as no answer by NQ, because the answer is not found on Wikipedia (either by NQ or our best efforts). However, an
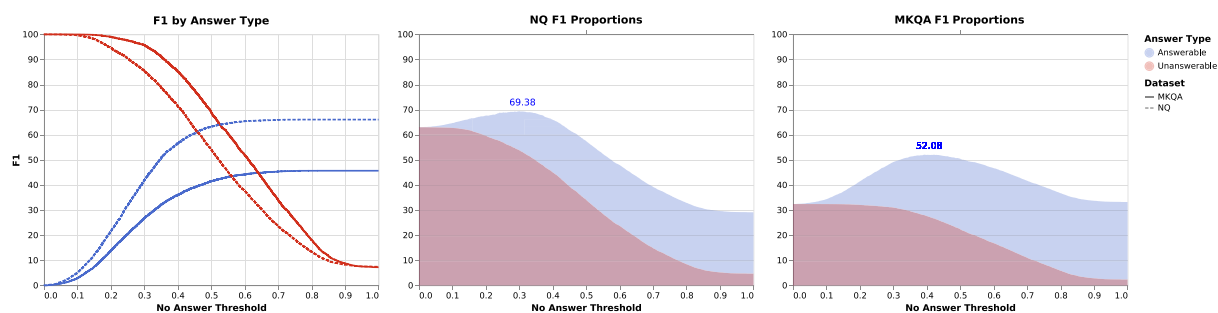
Figure 4: **Comparing MKQA and NQ English Annotations.** The performance of the same English BERT-LARGE model on each of Natural Questions (NQ) annotations and MKQA annotations, using the MKQA evaluation metrics. For all plots the y-axis is F1 score and the x-axis is the value of the threshold over No Answer probabilities. F1 by Answer Type (left diagram) compares the accuracy of the model on Answerable and Unanswerable examples for each dataset, showing Unanswerable examples are on average easier in MKQA, and Answerable examples are on average harder in MKQA. NQ F1 Proportions (middle) and MKQA F1 Proportions (right) show what proportion of the aggregate F1 score is derived from each Answer Type. These plots demonstrate MKQA is more difficult than NQ because there is a higher proportion of answerable questions, which are harder on average.

answer can be found by MKQA graders from other websites and sources.

- **Q:** *How many teeth does a saltwater crocodile have?*
- **NQ URL:** Wikipedia: *Saltwater Crocodile.*
- **NQ Answer:** *No Answer*
- **MKQA Answers:** *''66''*

(c) Annotator misses valid answer For this query, the answer is clearly visible in the provided Wikipedia article, but NQ's annotation process yields no answer.

- **Q:** *What language do they speak in the ukraine?*
- **NQ URL:** Wikipedia: *Languages of Ukraine.*
- **NQ Answer:** *No Answer*
- **MKQA Answers:** *''Ukrainian''*

Given the answer to these queries are not easily found in the corpus, by retrieval, or by human annotators, they are likely more challenging on average. As such, their label shift from no answer in NQ to answerable in MKQA likely explains why there is higher mean difficulty of answerable questions in MKQA, as observed in Figure 4. To understand the prevalence of each error type, we compute how often *any* MKQA answer appears in the retrieved document for which the NQ label says no answer exists. We find a valid answer appears in 70.4% of these documents, suggesting

category (c), annotator error, is the largest source of such unanswerable queries in NQ (and the largest source of improvement in label quality for MKQA).

The middle and right diagrams in Figure 4 normalize the answer types by their proportion within the dataset, so we can compare their relative contributions to the aggregate F1 (the sum of answerable and unanswerable). NQ labels enable a much higher aggregate F1 score (69.38% at the best threshold) than MKQA (52.08% at the best threshold) primarily due to the higher proportion of unanswerable examples—which are easier on average than answerable examples. By comparing the ratio of unanswerable to answerable examples attempted at the best thresholds in each of the middle and right diagrams (the blue regions vs. the red regions) we see that the MKQA task is more oriented to answering questions rather than abstaining.

Due to the *Parallel Question* property of MKQA, the dataset is similarly challenging in all 26 languages. There is also a noticeable gap between the performance on English and on lower-resourced languages (Figure 3). For Korean and Arabic the best F1 score is only 6% higher than the lower bound score of 32.42% obtained from predicting exclusively ''unanswerable.'' This demonstrates that existing transfer learning methods have significant deficits to overcome for low-resource multilingual QA to match English performance. MKQA offers a challenging benchmark to measure this cross-language progress specifically.

Future Work The parallel questions property of MKQA offers alternative task setups in addition to typical open domain question answering. Lewis et al. (2020) suggests a generalized cross-lingual transfer task (G-XLT) where the question and answer languages are intentionally different. Alternatively, future work might assume we are given the English question-answer pairs, and attempt to propagate these answers into other languages by localizing the questions and answers.

We anticipate that this dataset will enable industry practitioners and researchers to rapidly test and compare novel cutting-edge techniques for QA against existing techniques in a more fair, comparable, and precise manner than previous benchmarks. Additionally, we hope that the linguistic diversity and large number of languages will inspire more researchers to treat model performance across many (partially less-resourced) languages as an important and worthy goal in itself. As MKQA offers the **only** open-QA option for many of these languages, we also hope to spark important research in these monolingual, non-English settings.

7 Conclusion

In this work, we introduce a multilingual open domain question answering evaluation set. Its properties, including geographical invariance, language-parallel questions, retrieval-independent annotations, and linguistic diversity, set it apart from existing resources in terms of annotation quality, difficulty, and flexibility to evaluate new approaches. We encourage future multilingual benchmarks to adopt data collection and annotation principles to promote higher-quality, and informative evaluation practices. We evaluate several baselines, based on state-of-the-art methods, and demonstrate ample room for improvement both in English and in the tail of lower-resourced languages. We hope that this evaluation set enables wider exploration of cross-lingual and monolingual methods in non-English QA.

Acknowledgments

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020a. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684. https://doi.org/10.18653/v1/2020 .emnlp-main.618

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020b. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637. https://doi.org/10 .18653/v1/2020.acl-main.421

Akari Asai, Jungo Kasai, Jonathan H. Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2021. XOR QA: Cross-lingual open-retrieval question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 547–564. https://doi.org/10 .18653/v1/2021.naacl-main.46

Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016. A thorough examination of the CNN/Daily Mail reading comprehension task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2358–2367, Berlin, Germany. Association for Computational Linguistics. https://doi.org/10 .18653/v1/P16-1223

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver,

Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-1171

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. *Transactions of the Association for Computational Linguistics*, 8:454–470. https://doi.org/10.1162/tacl_a_00317

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2019a. Cross-lingual machine reading comprehension. 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 1586–1595.

Yiming Cui, Zhipeng Chen, Si Wei, Shijin Wang, Ting Liu, and Guoping Hu. 2017. Attention-over-attention neural networks for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 593–602, Vancouver, Canada. Association for Computational Linguistics.

Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2019b. A span-extraction dataset for Chinese machine reading comprehension. 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 5886–5891.

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.

Bhuwan Dhingra, Hanxiao Liu, Zhilin Yang, William Cohen, and Ruslan Salakhutdinov. 2017. Gated-attention readers for text comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1832–1846, Vancouver, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-1168

B. Green, A. Wolf, C. Chomsky, and K. Laughery. 1986. *BASEBALL: An Automatic Question Answerer*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Deepak Gupta, Surabhi Kumari, Asif Ekbal, and Pushpak Bhattacharyya. 2018. MMQA: A multi-domain multi-lingual question-answering framework for English and Hindi. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, and others. 2018. Dureader: A Chinese machine reading comprehension dataset from real-world applications. In *Proceedings of the Workshop on Machine Reading for Question Answering*, pages 37–46. https://doi.org/10.18653/v1/W18-2605

Tsung-Yuan Hsu, Chi-Liang Liu, and Hung-yi Lee. 2019. Zero-shot reading comprehension by cross-lingual transfer learning with multilingual language representation model. 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 5933–5940, Hong Kong, China. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin

Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *arXiv preprint arXiv:2003.11080*.

Yimin Jing, Deyi Xiong, and Zhen Yan. 2019. Bipar: A bilingual parallel dataset for multilingual and cross-lingual reading comprehension on novels. 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 2452–2462. `https://doi.org/10.18653/v1/D19 -1249`

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611. `https://doi .org/10.18653/v1/P17-1147`

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. `https://doi.org/10.18653/v1/2020 .emnlp-main.550`

Vishwajeet Kumar, Nitish Joshi, Arijit Mukherjee, Ganesh Ramakrishnan, and Preethi Jyothi. 2019. Cross-lingual training for automatic question generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4863–4872. `https://doi.org/10.18653/v1/P19 -1481`

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, 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 for Computational Linguistics, 7:453–466. `https://doi.org/10.1162 /tacl_a_00276`

Chia-Hsuan Lee and Hung-Yi Lee. 2019. Cross-lingual transfer learning for question answering. *arXiv preprint arXiv:1907.06042*.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330. `https://doi.org/10.18653/v1/2020.acl -main.653`

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018. `https://doi.org/10 .18653/v1/2020.emnlp-main.484`

Seungyoung Lim, Myungji Kim, and Jooyoul Lee. 2019. Korquad1. 0: Korean QA dataset for machine reading comprehension. *arXiv preprint arXiv:1909.07005*.

Jiahua Liu, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2019a. XQA: A cross-lingual open-domain question answering dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2358–2368.

Pengyuan Liu, Yuning Deng, Chenghao Zhu, and Han Hu. 2019b. XCMRC: Evaluating cross-lingual machine reading comprehension. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 552–564. Springer. `https://doi.org /10.1007/978-3-030-32233-5_43`

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy,

Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019c. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Hussein Mozannar, Elie Maamary, Karl El Hajal, and Hazem Hajj. 2019. Neural Arabic question answering. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 108–118. https://doi.org/10.18653/v1/W19-4612

Ella Rabinovich and Shuly Wintner. 2015. Unsupervised identification of translationese. *Transactions of the Association for Computational Linguistics*, 3:419–432. https://doi.org/10.1162/tacl_a_00148

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392. https://doi.org/10.18653/v1/D16-1264

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426. https://doi.org/10.18653/v1/2020.emnlp-main.437

Chih Chieh Shao, Trois Liu, Yuting Lai, Yiying Tseng, and Sam Tsai. 2018. DRCD: A Chinese machine reading comprehension dataset. *arXiv preprint arXiv:1806.00920*.

Gary F. Simons and Charles D. Fennig. 2018. Ethnologue: Languages of the world, twenty. *Dallas, Texas: SIL International. Online version:* http://www.ethnologue.com

Shuly Wintner. 2016. Translationese: Between human and machine translation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Tutorial Abstracts*, pages 18–19, Osaka, Japan. The COLING 2016 Organizing Committee.

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.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639. https://doi.org/10.18653/v1/2020.acl-main.148