LAReQA: Language-Agnostic Answer Retrieval from a Multilingual Pool

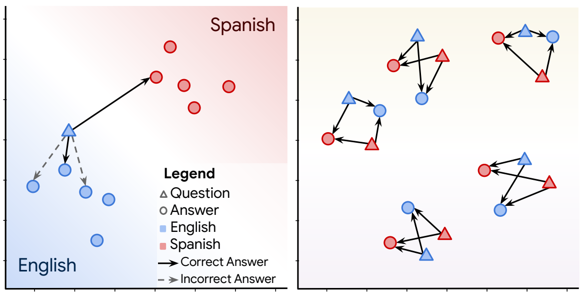
Uma Roy[‡*] **Noah Constant**[*] **Rami Al-Rfou** **Aditya Barua** **Aaron Phillips** **Yinfei Yang**
Google Research
uma.roy.us@gmail.com
{nconstant,rmyeid,adityabarua,aaronbphillips,yinfeiy}@google.com

Abstract

We present LAReQA, a challenging new benchmark for language-agnostic answer retrieval from a multilingual candidate pool. Unlike previous cross-lingual tasks, LAReQA tests for "strong" cross-lingual alignment, requiring semantically related *cross*-language pairs to be closer in representation space than unrelated *same*-language pairs. This level of alignment is important for the practical task of cross-lingual information retrieval. Building on multilingual BERT (mBERT), we study different strategies for achieving strong alignment. We find that augmenting training data via machine translation is effective, and improves significantly over using mBERT out-of-the-box. Interestingly, model performance on zero-shot variants of our task that only target "weak" alignment is not predictive of performance on LAReQA. This finding underscores our claim that language-agnostic retrieval is a substantively new kind of cross-lingual evaluation, and suggests that measuring both weak and strong alignment will be important for improving cross-lingual systems going forward. We release our dataset and evaluation code at https://github.com/google-research-datasets/lareqa.

1 Introduction

Recent progress in self-supervised pretraining for language understanding has enabled training large multilingual models on 100+ languages at the same time, as in multilingual BERT (mBERT) and XLM-R (Devlin et al., 2019; Conneau et al., 2019). These models, despite being trained without any explicit objective of cross-lingual alignment, are surprisingly effective for cross-lingual transfer (Pires et al., 2019; Wu and Dredze, 2019; Conneau et al., 2019), suggesting that the models may have learned to "factor out" language and embed inputs into a language-agnostic space.



(a) **Weak Alignment** (b) **Strong Alignment**

Figure 1: A weakly aligned multilingual embedding space enables zero-shot transfer between languages, but incorrect answers in the same language are preferred over correct answers in a different language. A strongly aligned embedding space "factors out" language, so the most semantically relevant pairs are always the closest, regardless of language.

At the same time, Wu and Dredze (2019) observe that mBERT representations at all layers are highly accurate (>96%) at classifying language ID, which the authors find "surprising given the model's zero-shot cross-lingual abilities". This raises an interesting question. To what degree are models like mBERT and XLM-R language agnostic or easily adaptable to be so? Have they effectively disentangled the language-specific signal from the underlying semantic content, with each occupying a separate subspace? Can the learned representations be adapted by a lightweight alignment procedure to be *truly* language agnostic?

To address these questions, we introduce a challenging new task, LAReQA: Language Agnostic Retrieval Question Answering, which requires models to retrieve relevant cross-lingual answers from a multilingual candidate pool. Perform-

[‡]Work done as a Google AI Resident.
[*]Corresponding authors: uma.roy.us@gmail.com, nconstant@google.com

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 5919–5930,
November 16–20, 2020.

ing well on this task demands a stronger degree of cross-lingual alignment than previous cross-lingual evaluations like XNLI (Conneau et al., 2018). Concretely, we propose to distinguish "weak" vs. "strong" alignment, defined as follows and illustrated in Figures 1a–1b:

Weak Alignment For any item in language L_1, the nearest neighbor in language L_2 is the most semantically relevant item. (The specific notion of "relevance" may vary across tasks.)

Strong Alignment For any item, all semantically relevant items are closer than all irrelevant items, regardless of their language. Crucially, relevant items in different languages are closer than irrelevant items in the same language.[1]

To our knowledge, LAReQA is the first cross-lingual benchmark to target strong alignment.[2] Building on top of multilingual BERT, we develop and test several baseline models on LAReQA. We find that mBERT already exhibits strong alignment between *some* language pairs, but that this can be improved significantly by leveraging machine translation to extend the set of training examples and encourage cross-lingual alignment.

One observation that emerges from our experiments is that strong alignment comes at a cost. Specifically, our baseline that reaches the best LAReQA performance lags behind other baselines on the narrower task of retrieving relevant answers that match the question language.

Our main contributions are: (1) We propose a new framework for classifying different degrees of cross-lingual alignment. (2) We propose a challenging new benchmark to evaluate language bias in representations, setting stricter notions of cross-lingual embedding space alignment. (3) We investigate the potential for multilingual BERT to achieve "strong" cross-lingual alignment, including various fine-tuning techniques to improve alignment. (4) We publish our trained models and LAReQA benchmarking code for others to reproduce.[3]

2 Looking for Answers across Languages

In this section, we present the practical task of **answer retrieval from a multilingual candidate pool**, and argue that this task goes beyond existing cross-lingual benchmarks in demanding models with strongly aligned multilingual representations. The task can be summarized as: given a question in one language and potential answers in many different languages, retrieve the best answer for the question, *regardless* of language. We begin by describing why this task is both useful and challenging. Subsequently, we compare this task with existing cross-lingual tasks, and show how they differ in their ability to measure "language bias".

2.1 Practical Value of Cross-lingual Retrieval

Finding relevant answers to questions from a large multilingual candidate pool is not a contrived task. User-generated content on the web is increasingly multilingual[4], and the best answer to a given question may be written in a different language than the question. Current information retrieval systems typically do not surface such "cross-lingual" results. This shortcoming is particularly problematic for speakers of languages with a smaller web presence, who may be forced to issue queries in a less familiar language in order to find useful results.

If search engines were language agnostic, retrieved results would come from a wide range of languages. Of course, some results would have to be machine translated to be made interpretable to a given user. But in many cases, even a poorly translated relevant result is more helpful than a less relevant native result.

One domain where cross-lingual retrieval is particularly valuable is in searching over user-generated content such as reviews of products and businesses. For example, suppose a Thai speaker wants to know if a local library offers private meeting rooms, and this question is answered by an existing Arabic user review of the library. Being able to respond to the Thai question by surfacing (a translation of) the relevant Arabic review unlocks content that was previously inaccessible.

There are various options for how to implement such a cross-lingual retrieval system in practice, not all of which require a model to support cross-language matching. One solution would be to store

[1] Stricter notions of cross-lingual alignment are possible, such as requiring that model representations remove any trace of the original text language, preventing language ID from being reconstructed. We treat these as sub-types of "strong" alignment, but leave their investigation for future work.

[2] Pires et al. (2019) develop a related heuristic by calculating the average vector delta between two languages, and testing how well translation targets can be retrieved by finding the closest neighbor to the source plus the delta.

[3] Our trained models are available at https://tfhub.dev/s?q=lareqa. Our dataset and evaluation code are available at https://github.com/google-research-datasets/lareqa.

[4] For example, see https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia.

English translations alongside results in the index, and translate all queries to English before performing search. Alternatively, one could pre-translate results into all languages ahead of time. However as these solutions require extra storage, we believe it is worth considering the simpler approach of retrieving multilingual results directly with a cross-lingual model.

2.2 Language Bias

From the modeling perspective, one of the main challenges in retrieving relevant answers from across languages is avoiding "language bias", where a model prefers one language over another. It's clear why this bias is harmful: if the model prefers answers in a given language, it is prone to retrieve *irrelevant* results in that language over relevant results from another language.

The main type of language bias we observe in our experiments is **same-language bias**—the tendency for models to prefer answers that match the question language. This is illustrated in Figure 1a, where the embeddings cluster primarily by language, and incorrect same-language candidates are preferred over any cross-language candidate. For a multilingual model to avoid same-language bias, it must align text from different languages under a language-agnostic embedding space, as in 1b.

2.3 Taxonomy of Cross-lingual Tasks

Existing cross-lingual tasks—including all tasks in the recent XTREME suite (Hu et al., 2020)—fall into two categories, as described below. Neither type allows us to diagnose language bias and test for language-agnostic embeddings. The key missing piece is that none of these tasks require the model to make a choice among candidates in different languages.

Monolingual Tasks in Many Languages Most existing cross-lingual benchmarks are formed by collecting monolingual tasks across various languages. Often, these evaluations are framed in terms of zero-shot or few-shot transfer learning, with the assumption that a practitioner only has access to task-specific training data in a single language. For instance, XNLI (Conneau et al., 2018) tests how well a model fine-tuned on an English classification task (natural language inference) can generalize to non-English versions of the same task. Similarly, MLDoc (Schwenk and Li, 2018) and PAWS-X (Yang et al., 2019b) test cross-language transfer on document classification and paraphrase identification respectively.

Several recent cross-lingual QA benchmarks can also be described as cross-lingual collections of monolingual tasks. For example, XQuAD (Artetxe et al., 2019) extends the popular SQuAD (Rajpurkar et al., 2016) benchmark to cover QA pairs in 11 languages, but models are only tested on finding answers in contexts that match the question language.[5] TyDi QA (Clark et al., 2020) and MLQA's (Lewis et al., 2019) "cross-lingual transfer" task also fall under this category.

Cross-lingual Tasks with Monolingual Candidates A second class of cross-lingual tasks tests whether a model can, given an input in language X, identify a target in language Y. However, crucially, the set of candidates is restricted to language Y. Thus, while the task is inherently cross-lingual, it does not test for language bias.

BUCC (Zweigenbaum et al., 2017) is a task of this type. Given an English sentence, the task is to retrieve the corresponding translation from a monolingual pool of candidates in another language. Similarly, Tatoeba (Artetxe and Schwenk, 2019) tests retrieval of translation pairs between English and 112 languages, but is restricted to monolingual pools. Bilingual lexicon induction or BLI (Glavaš et al., 2019) is a similar task of cross-lingual retrieval from a monolingual pool, but targeting words rather than sentences.[6]

MLQA (Lewis et al., 2019) is an extractive question answering task, and in one variant of the task, "*generalized* cross-lingual transfer", the question and answer are drawn from different languages. However, even in this case, the candidate answers are restricted to spans within a specific (monolingual) paragraph of context, so there is no way to assess whether the model is biased in preferring answers in one language over another.

3 LAReQA

Having motivated the need for a cross-lingual benchmark that asks models to choose *between*

[5] Due to its parallel construction, it is possible to construct "mixed-language" QA pairs from XQuAD. This is the approach we take in Section 3.

[6] One could construct versions of BUCC, Tatoeba and BLI that test for strong alignment, by switching to multilingual candidate pools. It would be interesting to compare these benchmarks to LAReQA in future work. Note, however, that the resulting tasks are somewhat "unnatural", in that there is typically no need to consider same-language candidates when mining for translation pairs or building a bilingual lexicon.

languages, we now present a concrete case of such a cross-lingual evaluation, **LAReQA: Language-Agnostic Retrieval Question Answering**.

3.1 Constructing LAReQA

Our goal is to construct a QA retrieval task over a large multilingual pool where many or most of the target answers are found in a different language than the question itself. To achieve this, we take the existing cross-lingual *extractive* QA tasks XQuAD and MLQA and convert them into *retrieval* tasks: **XQuAD-R** and **MLQA-R**. These sets are designed so as to include parallel QA pairs across languages, allowing us to match questions with answers from different languages. We release XQuAD-R at `https://github.com/google-research-datasets/lareqa`.

XQuAD is constructed by taking 240 paragraphs from the SQuAD v1.1 dev set and professionally translating the questions and contexts into 10 languages. Thus each question appears in 11 different languages and has 11 parallel correct answers. MLQA is constructed by using LASER (Artetxe and Schwenk, 2019) to mine parallel sentences from Wikipedia, which annotators then use to generate questions. Unlike XQuAD, the questions in MLQA have a variable number (2–4) of parallel correct answers across the corpus. Additionally, MLQA only covers 7 of the 11 XQuAD languages, and contexts surrounding the answer are not guaranteed to be parallel. See Artetxe et al. (2019) and Lewis et al. (2019) for more details on these sets.

To convert these span-tagging tasks into retrieval tasks, we follow the procedure from ReQA (Ahmad et al., 2019). Specifically, we break each contextual paragraph into sentences[7], and include all sentences across the dataset as candidate answers. A sentence is considered a correct answer to a question if it contains the target answer span for either that question or an equivalent question in another language (as identified by `qas_id`).[8] Table 1 shows the number of questions and candidates per language in XQuAD-R and MLQA-R. We use the MLQA dev set rather than the larger test set to keep the speed of evaluation reasonable. While the contexts for XQuAD are parallel across languages, differences in sentence breaking lead to variations in the number of candidates per language.[9]

	XQuAD-R		MLQA-R	
	questions	candidates	questions	candidates
ar	1190	1222	517	2545
de	1190	1276	512	2362
el	1190	1234	-	-
en	1190	1180	1148	6264
es	1190	1215	500	1787
hi	1190	1244	507	2426
ru	1190	1219	-	-
th	1190	852	-	-
tr	1190	1167	-	-
vi	1190	1209	511	2828
zh	1190	1196	504	2322

Table 1: Numbers of questions and candidates per language in XQuAD-R and MLQA-R.

3.2 Evaluation

For our primary evaluation, we use the standard information retrieval metric "mean average precision" (mAP), which measures a model's ability to rank relevant results over irrelevant ones. This metric is suitable when there are multiple relevant results for a given query. In our case, an XQuAD-R question will have 11 relevant answers, while an MLQA-R question will have 2–4 relevant answers.

Formally, given a set of questions Q and a ranking function over all candidates, mean average precision is defined as in Equation 1, where R_i is the number of correct answers for question q_i, $P@j(q_i)$ is the Precision@j for q_i and $\mathrm{rel}(i, j)$ is an "indicator" function with value 1 if the j-th ranked candidate for q_i is correct, 0 otherwise.

$$\mathrm{mAP} = \frac{1}{T} \sum_{q_i \in Q} \frac{1}{R_i} \sum_{j=1}^{K} P@j(q_i) \times \mathrm{rel}(i, j) \quad (1)$$

The mAP metric falls between 0 and 1. Any model that ranks all C correct answers in the top C positions (regardless of order) will achieve a perfect 1.0. Note, such a model cannot have strong language bias, as it needs to rank correct answers

[7] Sentence boundaries are generated by an internal sentence breaker. For Thai we use `https://pypi.org/project/thai-segmenter`.

[8] For both XQuAD and MLQA, there were no cases where an answer span crossed a sentence boundary. One sentence can be the correct answer for multiple questions (with different `qas_id`), as long as it contains the relevant target answer spans. We include all sentences from contextual paragraphs as candidates, even those that do not answer any question.

[9] Thai is an outlier, with around 70% the sentences per paragraph as the other languages. This is likely due to erroneous or ambiguous sentence breaking. Note, Thai lacks explicit sentence boundary markers, and human agreement on sentence breaking is much lower than English (Aroonmanakun, 2007).

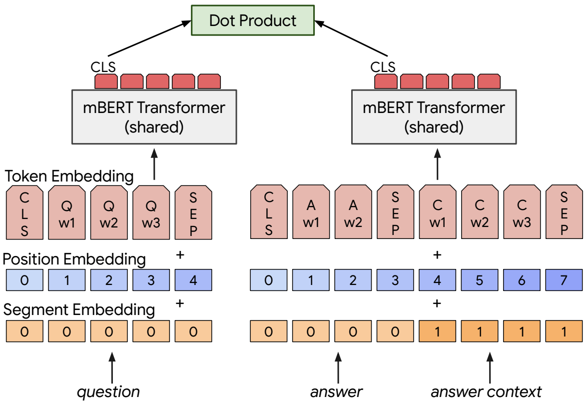


Figure 2: mBERT dual encoder architecture.

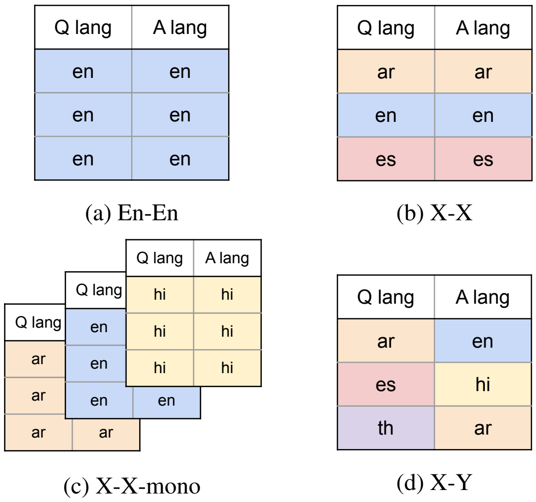


Figure 3: Sample batches for each baseline.

in *every* language highly. On the other hand, being free from language bias is not sufficient for high mAP. As a trivial example, a model that ranks candidates randomly will have a low mAP. In sum, performing well on LAReQA mAP requires both strong QA retrieval quality, as well as an absence of language bias.

4 Baseline Models

We consider several neural baseline models for evaluation on LAReQA. All our baselines are "dual encoder" models (Gillick et al., 2018), encoding questions and candidate answers separately. Unlike full cross-attention models, this architecture enables retrieval through approximate nearest neighbor search, and thereby scales well to large-scale retrieval problems. For more discussion of dual encoders for deep retrieval, see Gillick et al. (2018) and Ahmad et al. (2019).

Our baselines leverage multilingual BERT (Devlin et al., 2019), or "mBERT", for cross-lingual pretraining. These baselines allow us to test (i) how well mBERT already aligns languages in a language-agnostic space, and (ii) the degree to which it can be adapted to do so.

We initialize each tower of a dual encoder model with pretrained mBERT, sharing weights between the question and answer encoding towers[10], as in Figure 2. To obtain final question and answer encodings, we normalize the BERT CLS token to unit L2 norm. The model score for a QA pair $S(q,\, a)$ is the dot product of these encodings.

We fine-tune the mBERT towers for QA retrieval on SQuAD training data using an "in-batch sampled softmax" loss (Henderson et al., 2017), as this has been observed to converge quickly and perform well on retrieval tasks (Gillick et al., 2018). The loss, given in Equation 2, encourages the score of a correct answer pair $S(q_i,\, a_i)$ to be higher than scores for incorrect answer pairings from the mini-batch $S(q_i,\, a_j)$:[11]

$$-\frac{1}{K}\sum_{i=1}^{K}\left[S(q_i,a_i)-\log\sum_{j=1\, j\neq i}^{K}e^{S(q_i,a_j)}\right] \qquad (2)$$

We train four variants of our mBERT model (110M parameters), using different fine-tuning datasets and batching procedures. Each model is fine-tuned on 32 TPU-v3 cores, with a batch size of 2,048. The in-batch sampled softmax is calculated separately per core, over sub-batches of 64 QA pairs. We use the standard BERT learning rate schedule, with an initial learning rate of 1e-5, which performed the best among {1e-6, 3e-6, 1e-5, 3e-5}. We train all baselines for 100,000 steps and observe no overfitting.

Our first baseline **"En-En"** adapts mBERT to QA retrieval by fine-tuning on the 80,000 English QA pairs from the SQuAD v1.1 train set, with the ranking loss from Equation 2. This baseline tests how well mBERT can perform language-agnostic retrieval when only fine-tuning on English data.

Our second baseline **"X-X"** extends the same SQuAD train set by translating each example into the 11 XQuAD languages using an in-house trans-

[10] When feeding inputs to the answer encoding tower, we concatenate the answer sentence and its containing context paragraph ("answer context"), using BERT's segment ids to distinguish between the two.

[11] In practice, we scale the similarity scores by a trainable constant factor before computing the softmax, as we observed this led to faster convergence.

	XQuAD-R	MLQA-R
En-En	0.29	0.36
X-X	0.23	0.26
X-X-mono	0.52	0.49
X-Y	0.66	0.49
Translate-Test	0.72	0.58

Table 2: Mean average precision (mAP) of baseline models on XQuAD-R and MLQA-R.

	mAP −rand	mAP −same	% Δ	Rank (% Δ)
En-En	0.29	0.22	0.24	4
X-X	0.23	0.15	0.37	5
X-X-mono	0.52	0.47	0.10	2
X-Y	0.65	0.64	0.02	1
Translate-Test	0.69	0.60	0.13	3

Table 3: Performance on modified versions of XQuAD-R where one target answer is removed, either from the same language as the question (−same) or a random other language (−rand).

lation system. Within each example, the question and answer language are the same, giving 880,000 pairs total. If these pairs are shuffled and batched naively, as in Figure 3b, we expect the model to exhibit strong same-language bias, as all positive examples are within-language, while many in-batch negatives are cross-language. To avoid this bias, our third baseline **"X-X-mono"** trains on the same examples, but ensures that each batch is monolingual, as shown in Figure 3c.

Our fourth baseline **"X-Y"** is similar to X-X, but allows a question and answer to be translated into different languages, giving 9,680,000 examples. This setup is the first to directly incentivize the model to treat answers from other languages as correct, which we expect to further reduce same-language bias.

Our final baseline **"Translate-Test"** is not a proper text embedding model, as it relies on an external translation system at test time. Here, we simply translate any test data into English, and then score it with our En-En model.

Additionally, we compare the above baselines with Universal Sentence Encoder Multilingual QA (Yang et al., 2019a), which specifically targets cross-lingual QA retrieval. However, this model only supports 8 of the 11 XQuAD languages, and we found it was not competitive with our mBERT baselines, even when restricting evaluation to the supported languages. See Appendix A for details.

5 Results and Analysis

5.1 LAReQA Performance

We compare our five baseline models on the LAReQA task in Table 2. On both XQuAD-R and MLQA-R, the strongest model is the Translate-Test baseline. This is unsurprising in that LAReQA demands language-agnostic retrieval, and Translate-Test leverages an external machine translation system to actively "remove" the effects of language, by translating all test data to English.

Among the pure embedding approaches, X-Y does the best on XQuAD-R, and is tied for best on MLQA-R. The success of X-Y shows that training directly on "mixed-language" QA pairs is helpful for the end task of language-agnostic retrieval from a multilingual pool.

As expected, X-X-mono outperforms X-X, indicating that using a ranking loss with in-batch negatives is problematic when positives are within-language but negatives are mixed-language. Indeed, we will see shortly that X-X exhibits severe same-language bias.

For the remainder of the paper, we focus on XQuAD-R, as it is better balanced across languages than MLQA-R, and the two sets showed similar results.

5.2 Language Bias

We offer two additional analyses to more directly illustrate the language biases of our baselines. A third analysis, looking at the language distribution among top retrieved candidates, is given in Appendix B, and is consistent with the results here.

Remove One Target We rerun the XQuAD-R evaluation, but for each question, we remove one of its 11 target answers from the multilingual candidate pool. If a model is free of same-language bias, the effect of removing a single target should be constant, regardless of whether the removed target was in the same language as the question or not. Table 3 shows that in fact all our baselines perform better when a random *cross-language* target is removed (−rand), as compared to the same-language target (−same). Looking at the delta between these conditions, the X-Y baseline only displays a minimal bias, falling from 0.62 to 0.61 mAP. The most affected model is X-X, whose training procedure actively encouraged same-language bias. Interest-

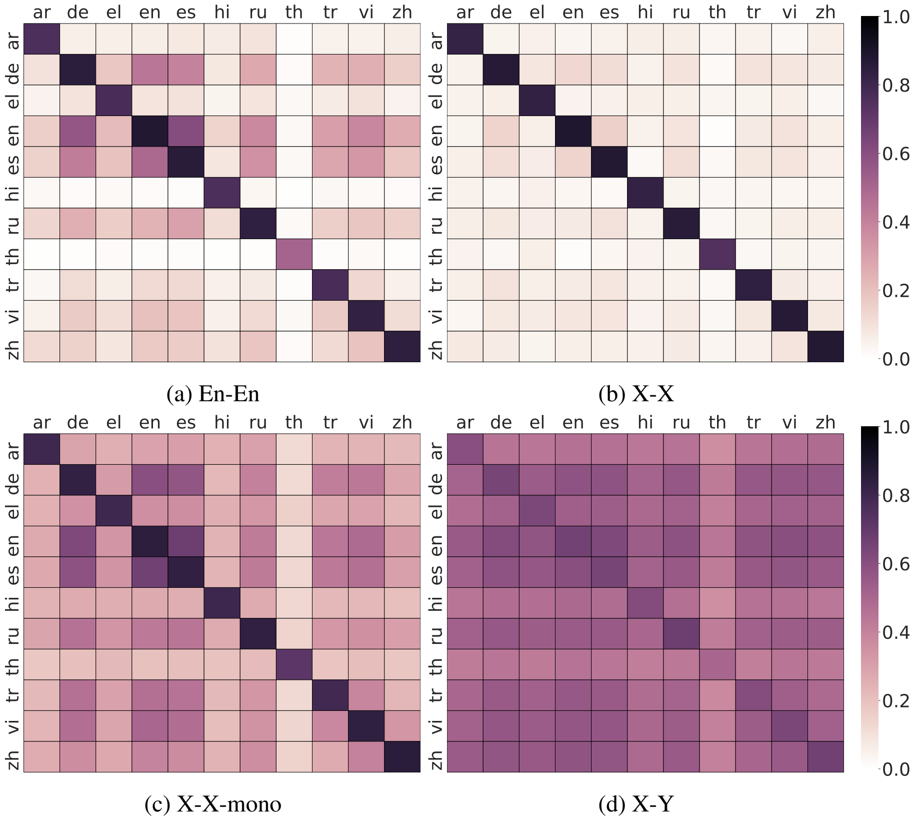


Figure 4: mAP on XQuAD-R broken down by question language (row) and answer language (column), when only one correct answer is included in the multilingual candidate pool.

ingly, even En-En shows a significant delta, indicating that simply fine-tuning mBERT on English QA is not sufficient to produce an embedding space that is strongly language agnostic.

Limit to One Target As a more in-depth analysis of language bias, we evaluate on retrieval from a multilingual pool containing *just one* correct answer. For XQuAD-R, since each question has 11 answers, this means evaluating on each target separately, with the other 10 targets removed from the pool. The heatmaps in Figure 4 show each baseline's mAP on single-answer retrieval, broken down by question language and answer language. Note, in this case mAP is equivalent to mean reciprocal rank (MRR)—the average inverse rank of the target answer. In line with our previous findings, all models display some degree of same-language bias, showing better performance on the diagonal, where Q and A languages match, than off-diagonal. The degree of bias matches the ranking from Table 3. X-Y displays the least bias, with most language pairs reaching over 0.4 mAP. As before, X-X is the most biased, but we also see significant bias in En-En and X-X-mono.

These results also shed light on how well mBERT supports strong cross-lingual alignment "out of the box". Interestingly, even En-En shows fairly strong alignment among typologically related languages (e.g. 0.61 mAP on English-to-German and 0.57 on English-to-Spanish). These findings parallel those of Pires et al. (2019) and Wu and Dredze (2019), who observe that mBERT zero-shot transfer is more effective among related languages. Our retrieval performance is lower between unrelated languages (e.g. 0.06 Arabic-to-Chinese), as well as on pairs where one of the languages is less well represented in mBERT's Wikipedia training data (Greek, Hindi and Thai).

While mBERT exhibits *some* strong cross-lingual alignment out of the box, our results show that this can be improved by using cross-lingual objectives, as in X-X-mono and X-Y. This finding echoes work by Artetxe and Schwenk (2019), Conneau and Lample (2019), Singh et al. (2019) and Siddhant et al. (2020) showing that cross-lingual training can improve zero-shot transfer.

One point worth highlighting is the trade-off between on-diagonal and off-diagonal performance. If we limit attention to the diagonal, the models rank X-X > X-X-mono > En-En > X-Y. Thus, it appears there is a "cost" to strong cross-lingual alignment. For a given application, it may be worth sacrificing same-language quality to achieve better cross-language performance. However this raises the question: Is there any training technique that can achieve strong cross-lingual alignment without degrading within-language performance?

5.3 Comparison to Standard Zero-Shot Cross-Lingual Transfer

To highlight the difference between LAReQA and standard zero-shot cross-lingual evaluations like XNLI, we construct a zero-shot version of our QA retrieval task. We process the XQuAD data as before, but instead of a shared multilingual candidate pool, we restrict candidates to those matching the question language. Thus, like XNLI and the original XQuAD task, we're testing a model's ability to generalize to monolingual tasks in new languages.

The performance of our baselines on this "zero-shot" retrieval from a monolingual pool is shown in Table 4. Remarkably, the model ranking under this task diverges from that under our proposed LAReQA task of retrieval from a *multilingual* pool. In particular, the X-X(-mono) baselines which were only trained on "within-language" examples now perform the best, beating the top LAReQA baselines Translate-Test and X-Y.

This result supports our claim that LAReQA tests for cross-lingual alignment in a way that existing zero-shot evaluations do not. Despite their strong language bias, visible in the dark diagonals in Figure 4, the X-X(-mono) models give excellent performance in the typical zero-shot cross-lingual transfer scenario. Yet, as we saw in in Table 2, these baselines are fundamentally ill-suited for retrieval from a multilingual pool, which demands strongly aligned multilingual embeddings. As an extreme case, X-X scored a mere 0.23 on LAReQA mAP, compared to the best embedding model X-Y and the best overall baseline Translate-Test with 0.63 and 0.70 respectively.

5.4 Embedding Spaces

Figure 5 plots the first two principal components of each baseline's embeddings of all XQuAD-R questions and candidates in English and Chinese (chosen as they are genetically unrelated languages with distinct scripts). The X-X embeddings show a dramatic separation between Chinese and English. This is a clear case of weak alignment, where a

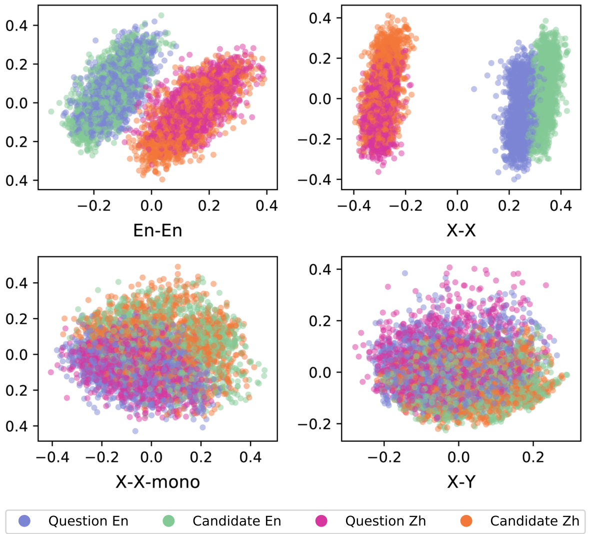


Figure 5: Model embeddings of all English and Chinese questions and candidates from XQuAD-R, visualized under 2D PCA.

model achieves good zero-shot performance (cf. Table 4) despite its embeddings being principally determined by language.

More generally, we observe that the more a model separates languages in embedding space, the worse it performs on LAReQA (cf. Table 2). This ordering is also reflected in the degree to which language ID is predictable from the embeddings. When we use a logistic regression to predict language (English vs. Chinese) from the question and candidate embeddings, the accuracies on a one-third holdout are X-X: 99.2%, En-En: 97.7%, X-X-mono: 87.5% and X-Y: 54.0%. This supports the claim that LAReQA is a better test of strong alignment than current zero-shot tasks.

6 Conclusion

LAReQA is a challenging new benchmark testing answer retrieval from a multilingual candidate pool. It goes further than previous cross-lingual benchmarks in requiring "strong" cross-lingual alignment, which is a step closer to truly language-agnostic representations.

We believe there is significant headroom for models to improve on LAReQA. Our best initial baseline sidesteps the alignment problem by simply translating all test data to English. Among embedding-based models, our strongest baseline ("X-Y") actively removes language bias by augmenting training data to include machine-translated cross-lingual examples. However, to achieve strong

	ar	de	el	en	es	hi	ru	th	tr	vi	zh	Avg	Rank Zero-shot	Rank LAReQA
En-En	0.76	0.87	0.78	**0.90**	0.87	0.76	0.85	0.51	0.77	0.84	0.85	0.80	4	4
X-X	**0.83**	0.87	0.84	0.89	0.88	0.83	0.86	0.75	0.84	0.87	0.88	0.85	2	5
X-X-mono	**0.83**	**0.88**	0.85	**0.90**	**0.89**	**0.84**	**0.87**	**0.76**	**0.85**	**0.88**	**0.89**	**0.86**	1	3
X-Y	0.75	0.83	0.79	0.85	0.83	0.76	0.82	0.69	0.78	0.80	0.82	0.79	5	2
Translate-Test	**0.83**	0.87	**0.86**	**0.90**	0.88	0.81	0.86	0.74	0.82	0.84	0.85	0.84	3	1

Table 4: mAP on the zero-shot version of XQuAD-R, retrieving a single answer from a *monolingual* pool that matches the question language.

alignment, this model sacrifices performance on both retrieval from a monolingual pool (Table 4), as well as retrieval of same-language candidates (Figure 4d diagonal). It is an interesting question for future work whether strong alignment always comes at a cost, or if better training techniques will lead to models that can improve on all these measures simultaneously.

Acknowledgements

Thank you to Mandy Guo and Gustavo Hernandez Abrego for discussion and initial analysis. We thank Sebastian Ruder and Melvin Johnson for helpful comments on an earlier draft of this paper. We also thank Rattima Nitisaroj for helping us evaluate the quality of our Thai sentence breaking.

References

Amin Ahmad, Noah Constant, Yinfei Yang, and Daniel Cer. 2019. ReQA: An evaluation for end-to-end answer retrieval models. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 137–146, Hong Kong, China. Association for Computational Linguistics.

Wirote Aroonmanakun. 2007. Thoughts on word and sentence segmentation in Thai. In *Proceedings of the Seventh Symposium on Natural language Processing, Pattaya, Thailand, December 13–15*, pages 85–90.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. *arXiv preprint arXiv:1910.11856*.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

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.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 7059–7069. Curran Associates, Inc.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Daniel Gillick, Alessandro Presta, and Gaurav Singh Tomar. 2018. End-to-end retrieval in continuous space. *arXiv preprint arXiv:1811.08008*.

Goran Glavaš, Robert Litschko, Sebastian Ruder, and Ivan Vulić. 2019. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 710–721, Florence, Italy. Association for Computational Linguistics.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *CoRR*, abs/1705.00652.

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*.

Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2019. MLQA: Evaluating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001.

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, Austin, Texas. Association for Computational Linguistics.

Holger Schwenk and Xian Li. 2018. A corpus for multilingual document classification in eight languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Aditya Siddhant, Ankur Bapna, Henry Tsai, Jason Riesa, Karthik Raman, Melvin Johnson, Naveen Ari, and Orhan Firat. 2020. Evaluating the cross-lingual effectiveness of massively multilingual neural machine translation. *arXiv preprint arXiv:1909.00437*.

Jasdeep Singh, Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2019. XLDA: Cross-lingual data augmentation for natural language inference and question answering. *CoRR*, abs/1905.11471.

Shijie Wu and Mark Dredze. 2019. Beto, Bentz, Becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2019a. Multilingual universal sentence encoder for semantic retrieval. *arXiv preprint arXiv:1907.04307*.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019b. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. 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 3678–3683.

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2017. Overview of the second BUCC shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 60–67.

A USE-QA

We test the Universal Sentence Encoder Multilingual QA (Yang et al., 2019a), which specifically targets cross-lingual QA retrieval on our LAReQA benchmark. The architecture and training details of USE-QA are provided in Yang et al. (2019a). We use the USE-QA model out of the box[12] without fine-tuning on any SQuAD data, as it was already trained specifically for retrieval QA. As USE-QA only supports 8 of the 11 XQuAD languages (ar, de, en, es, ru, th, tr, zh), we restrict our evaluation to these languages when comparing USE-QA to other models.

	XQuAD-R_{USE}
En-En	0.33
X-X	0.25
X-X-mono	0.55
X-Y	0.67
Translate-Test	0.73
USE-QA	0.51

Table 5: Mean average precision (mAP) of baseline models on XQuAD-R_{USE}, a version of XQuAD-R restricted to the 8 languages supported by USE-QA.

From Table 5, we can see USE-QA is not competitive with the mBERT baselines, despite being trained specifically for QA retrieval over a large in-house QA dataset. However, it may be possible to improve this performance by fine-tuning for SQuAD retrieval.

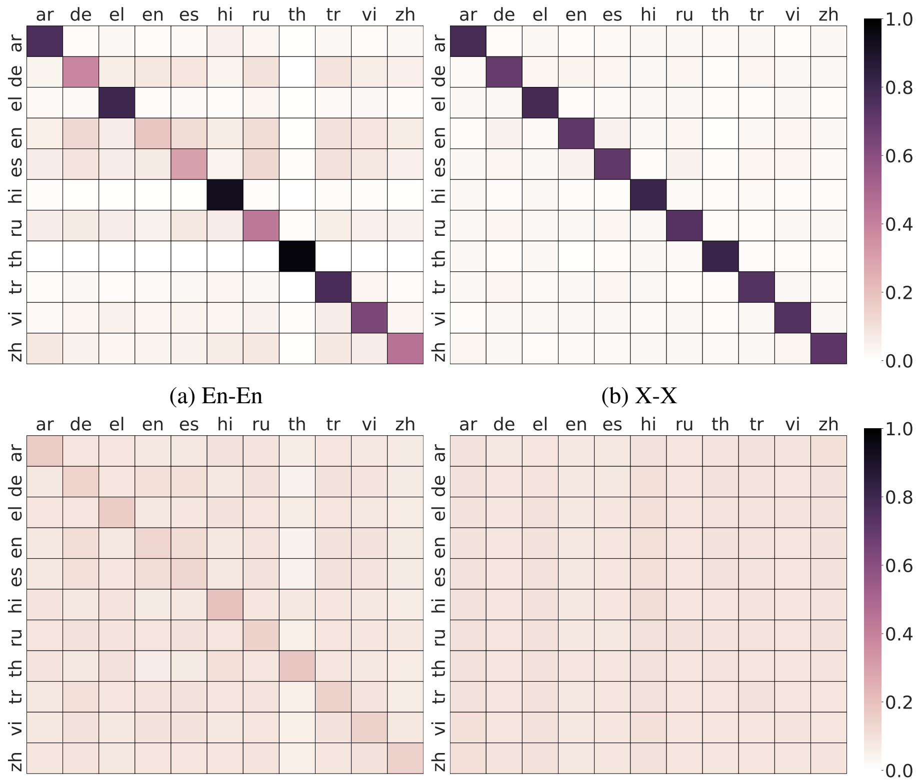
B Language Distributions of Top Results

Our core LAReQA mAP metric tests for both question answering (QA) matching ability, as well as absence of language bias. We can factor out QA performance and focus more directly on language bias by simply ignoring which answers are correct, and observing the distribution of languages that a model retrieves among its top-ranked candidates.

The heatmaps in Figure 6 show for each question language (row), the frequency of different answer languages (column) among the top 100 retrieved candidates, for each of our baseline models on the XQuAD-R dataset. The strong diagonal in X-X indicates that when the question is in a given language, nearly all of the top 100 retrieved results are in the same language. Overall, this measure of language bias is consistent with those discussed in Section 5.2, with the models ranking X-Y > X-X-mono > En-En > X-X.

Interestingly, X-Y performs almost perfectly on this “semantics-free” measurement of language bias. This is in contrast to the mAP performance of the same model in Figure 4d, where the retrieval of *correct* answers is somewhat improved when the Q and A languages match. Taken together, we can say that X-Y is nearly perfectly unbiased in which languages it retrieves *on the whole*, but is slightly biased as to which language pairs exhibit the strongest QA matching.

[12]https://tfhub.dev/google/universal-sentence-encoder-multilingual-qa



(a) En-En (b) X-X

(c) X-X-mono (d) X-Y

Figure 6: Proportion of top-100 retrieved answers that are in a given language (column), broken down by question language (row). Each row sums to 1.0.