Pivot Through English: Reliably Answering Multilingual Questions without Document Retrieval

Ivan Montero[♠] **Shayne Longpre**[♣]
Ni Lao[♣] **Andrew J. Frank**[♣] **Christopher DuBois**[♣]

[♠]Paul G. Allen School of Computer Science & Engineering, University of Washington
[♣]Apple Inc.
ivamon@cs.washington.edu
{slongpre, ni_lao, a_frank, cdubois}@apple.com

Abstract

Existing methods for open-retrieval question answering in lower resource languages (LRLs) lag significantly behind English. They not only suffer from the shortcomings of non-English document retrieval, but are reliant on language-specific supervision for either the task or translation. We formulate a task setup more realistic to available resources, that circumvents document retrieval to reliably transfer knowledge from English to lower resource languages. Assuming a strong English question answering model or database, we compare and analyze methods that pivot through English: to map foreign queries to English and then English answers back to target language answers. Within this task setup we propose Reranked Multilingual Maximal Inner Product Search (RM-MIPS), akin to semantic similarity retrieval over the English training set with reranking, which outperforms the strongest baselines by 2.7% on XQuAD and 6.2% on MKQA. Analysis demonstrates the particular efficacy of this strategy over state-of-the-art alternatives in challenging settings: low-resource languages, with extensive distractor data and query distribution misalignment. Circumventing retrieval, our analysis shows this approach offers rapid answer generation to many other languages off-the-shelf, without necessitating additional training data in the target language.

1 Introduction

Open-Retrieval question answering (ORQA) has seen extensive progress in English, significantly outperforming systems in lower resource languages (LRLs). This advantage is largely driven by the scale of labelled data and open source retrieval tools that exist predominantly for higher resource languages (HRLs) — usually English.

To remedy this discrepancy, recent work leverages English supervision to improve multilingual systems, either by simple translation or zero shot transfer (Asai et al., 2018; Cui et al., 2019; Charlet et al., 2020). While these approaches have helped generalize reading comprehension models to new languages, they are of limited practical use without reliable information retrieval in the target language, which they often implicitly assume.

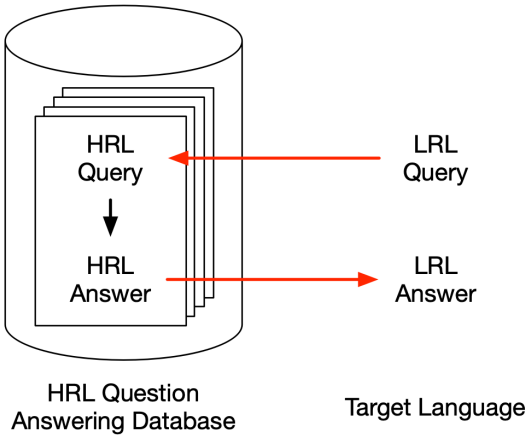

Figure 1: **Cross-Lingual Pivots (XLP):** We introduce the "Cross Lingual Pivots" task, formulated as a solution to multilingual question answering that circumvents document retrieval in low resource languages (LRL). To answer LRL queries, approaches may leverage a question-answer system or database in a high resource language (HRL), such as English.

In practice, we believe this assumption can be challenging to meet. A new document index can be expensive to collect and maintain, and an effective retrieval stack typically requires language-specific labelled data, tokenization tools, manual heuristics, and curated domain blocklists (Fluhr et al., 1999; Chaudhari, 2014; Lehal, 2018). Consequently, we discard the common assumption of robust non-English document retrieval, for a more realistic one: that there exists a high-quality English database of query-answer string pairs. We motivate and explore the Cross-Lingual Pivots (**XLP**) task (Section 2), which we contend will accelerate progress in LRL question answering by reflecting these practical considerations. This pivot task is

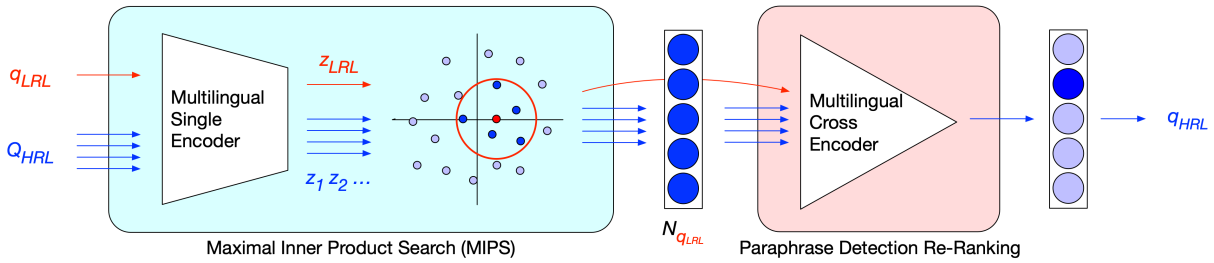

Figure 2: **Reranked Multilingual Maximal Inner Product Search (RM-MIPS):** For the Cross-Lingual Pivots task, we propose an approach that maps the LRL query to a semantically equivalent HRL query, finds the appropriate HRL answer, then uses knowledge graph or machine translation to map the answer back to the target LRL. Specifically, the first stage (in blue) uses multilingual single encoders for fast maximal inner product search (MIPS), and the second stage (in red) reranks the top k candidates using a more expressive multilingual cross-encoder that takes in the concatenation of the LRL query and candidate HRL query.

similar to "translate test" and "MT-in-the-middle" paradigms (Hajič et al., 2000; Zitouni and Florian, 2008; Schneider et al., 2013; Mallinson et al., 2017) except for the availability of the high-resource language database, which allows for more sophisticated pivot approaches. Figure 1 illustrates a generalized version of an XLP, where LRL queries may seek knowledge from any HRL with its own database.

For this task we combine and compare state-of-the-art methods in machine translation ("translate test") and cross-lingual semantic similarity, in order to map LRL queries to English, and then English answers back to the LRL target language. In particular we examine how these methods are affected by certain factors: (a) whether the language is high, medium or low resource, (b) the magnitude of data in the HRL database, and (c) the degree of query distribution alignment between languages (i.e., the number of LRL queries that have matches in the HRL database).

Lastly we propose an approach to this task, motivated by recent dense nearest neighbour (kNN) models in English which achieve strong results in QA by simply searching for similar questions in the training set (or database in our case) (Lewis et al., 2020). We leverage nearest neighbor semantic similarity search followed by cross-encoder reranking (see Figure 2), and refer to the technique as Reranked Multilingual Maximal Inner Product Search (**RM-MIPS**). Not only does this approach significantly improve upon "Translate Test" (the most common pivot technique) and state-of-the-art paraphrase detection baselines, our analysis demonstrates it is more robust to lower resource languages, query distribution misalignment, and the size of the English database.

By circumventing document retrieval and task-specific supervision signals, this straightforward approach offers reliable answer generation to many of the languages present in pretraining, off-the-shelf. Furthermore, it can be re-purposed to obtain reliable training data in the target language, with fewer annotation artifacts, and is complementary to a standard end-to-end question answering system. We hope this analysis complements existing multilingual approaches, and facilitates adoption of more practical (but effective) methods to improve knowledge transfer from English into other languages.

We summarize our contributions as:

- XLP: We explore a more realistic task setup for practically expanding Multilingual ORQA to lower resource languages.
- Comprehensive analysis of factors affecting XLP: (I) types of approaches (translation, paraphrasing) (II) language types, (III) database characteristics, and (IV) query distribution alignment.
- RM-MIPS: A flexible approach to XLP that beats strong (or state-of-the-art) baselines.

2 Task: Cross-Lingual Pivots

The Open-Retrieval Question Answering (ORQA) task evaluates models' ability to answer information-seeking questions. In a multilingual setting, the task is to produce answers in the same language as the query. In some cases, queries may only find answers, or sufficient evidence, in a different language, due to *informational asymmetries* (Group, 2011; Callahan and Herring, 2011). To address this, Asai et al. (2020) propose

Cross-Lingual Open-Retrieval Question Answering (XORQA), similar to the Cross-Lingual Information Retrieval (CLIR) task, where a model needs to leverage intermediary information found in other languages, in order to serve an answer in the target language. In practice, this intermediary language tends to be English, with the most ample resources and training data.

Building on these tasks, we believe there are other benefits to pivoting through high resource languages that have so far been overlooked, and consequently limited research that could more rapidly improve non-English QA. These two benefits are (I) large query-answer databases have already been collected in English, both in academia (Joshi et al., 2017) and in industry (Kwiatkowski et al., 2019), and (II) it is often very expensive and challenging to replicate robust retrieval and passage reranking stacks in new languages (Fluhr et al., 1999; Chaudhari, 2014; Lehal, 2018). [1] As a result, the English capabilities of question answering systems typically exceed those for non-English languages by large margins (Lewis et al., 2019; Longpre et al., 2020; Clark et al., 2020).

We would note that prior work suggests even without access to an English query-answer database, translation methods with an English document index and retrieval outperforms LRL retrieval for open-retrieval QA (see the end-to-end XOR-FULL results in Asai et al. (2020)). This demonstrates the persistent weakness of non-English retrieval, and motivates alternatives approaches such as cross-lingual pivots.

To remedy this disparity, we believe attending to these two considerations would yield a more realistic task setup. Like multilingual ORQA, or XORQA, the task of XLPs is to produce an answer $\hat{a}_{LRL}$ in the same "Target" language as question q_{LRL}, evaluated by Exact Match of F1 token-overlap with the real answer a_{LRL}. Instead of assuming access to a LRL document index or retrieval system (usually provided by the datasets), we assume access to an English database D_{HRL} which simply maps English queries to their English answer text. Leveraging this database, and circumventing LRL retrieval, we believe progress in this task will greatly accelerate multilingual capabilities of real question answering assistants.

[1] While it is straightforward to adapt question answering "reader" modules with zero-shot learning (Charlet et al., 2020), retrieval can be quite challenging. Not only is the underlying document index costly to expand and maintain for a new language (Chaudhari, 2014), but supervision signals collected in the target language are particularly important for dense retrieval and reranking systems which both serve as bottlenecks to downstream multilingual QA (Karpukhin et al., 2020). Additionally, real-world QA agents typically require human curated, language-specific infrastructure for retrieval, such as regular expressions, custom tokenization rules, and curated website blocklists.

3 Re-Ranked Multilingual Maximal Inner Product Search

For the first stage of the XLP task, our goal is to find an equivalent English query for a LRL query: "Query Matching". Competing approaches include Single Encoders and Cross Encoders, described further in section 4.2. Single Encoders embed queries independently into a latent vector space, meaning each query q_{EN} from the English database Q_{EN} can be pre-embedded offline. At inference time, the low resource query q_{LRL} is embedded, then maximal inner product search (MIPS) finds the approximate closest query q_{EN} among all Q_{EN} by cosine similarity. By comparison, Cross Encoders leverage cross-attention between q_{LRL} and candidate match q_{EN} at inference time, thus requiring $O(|Q_{EN}|)$ forward passes at inference time to find the best paraphrase. While usually more accurate this is computationally infeasible for a large set of candidates.

We propose a method that combines both Single Encoders and Cross Encoders, which we refer to as Reranked Mulilingual Maximal Inner Product Search (RM-MIPS). The process, shown in Figure 2, first uses a multilingual sentence embedder with MIPS to isolate the top-k candidate similar queries, then uses the cross encoder to rerank the candidate paraphrases. This approach reflects the Retrieve and Read paradigm common in OR-QA, but applies it to a multilingual setting for semantic similarity search.

The model first queries the English database using the Multilingual Single Encoder $\text{SE}(q_i) = z_i$ to obtain the k-nearest English query neighbors $\mathcal{N}_{q_{LRL}} \subseteq Q_{EN}$ to the given query q_{LRL} by cosine similarity.

$$\mathcal{N}_{q_{LRL}} = \underset{\{q_1,\ldots,q_k\} \subseteq Q_{EN}}{\arg\max} \sum_{i=1}^{k} \text{sim}(z_{LRL}, z_i)$$

Then, it uses the Multilingual Cross Encoder $\text{CE}(q_1, q_2)$ to score the remaining set of queries $\mathcal{N}_{q_{LRL}}$ to obtain the final prediction.

$$\textbf{RM-MIPS}(q_{LRL}) = \underset{q_{EN} \in \mathcal{N}_{q_{LRL}}}{\arg\max} \text{CE}(q_{EN}, q_{LRL})$$

RM-MIPS(q_{LRL}) proposes an equivalent English query q_{EN}, whose English answer can be pulled directly from the database.

	XQuAD	MKQA
High	es, de, ru, zh	de, es, fr, it, ja, pl, pt, ru, zh_cn
Medium	ar, tr, vi	ar, da, fi, he, hu, ko, nl, no, sv, tr, vi
Low	el, hi, th	km, ms, th, zh_hk, zh_tw

Table 1: We evaluate cross-lingual pivot methods by language groups, divided into high, medium, and low resource according to Wikipedia coverage Wu and Dredze (2020). Note that due to greater language diversity, MKQA contains lower resource languages than XQuAD.

4 Experiments

We compare systems that leverage an English QA database to answer questions in lower resource languages. Figure 1 illustrates a cross-lingual pivot (XLP), where the task is to map an incoming query from a low resource language to a query in the high resource language database (LRL → HRL, discussed in 4.2), and then a high resource language answer to a low resource language answer (HRL → LRL, discussed in 4.3).

4.1 Datasets

We provide an overview of the question answering and paraphrase datasets relevant to our study.

4.1.1 Question Answering

To assess cross-lingual pivots, we consider multilingual OR-QA evaluation sets that (a) contain a diverse set of language families, and (b) have "parallel" questions across all of these languages. The latter property affords us the opportunity to change the distributional overlap and analyze its effect (5.3).

XQuAD Artetxe et al. (2019) human translate 1.2k SQuAD examples (Rajpurkar et al., 2016) into 10 other languages. We use all of SQuAD 1.1 (100k+) as the associated English database, such that only 1% of database queries are represented in the LRL evaluation set.

MKQA Longpre et al. (2020) human translate 10k examples from the Natural Questions (Kwiatkowski et al., 2019) dataset to 25 other languages. We use the rest of the Open Natural Questions training set (84k) as the associated English database, such that only 10.6% of the database queries are represented in the LRL evaluation set[2].

[2]Open Natural Questions train set found here: https://github.com/google-research-datasets/natural-questions/tree/master/nq_open

4.1.2 Paraphrase Detection

To detect paraphrases between LRL queries and HRL queries we train multilingual sentence embedding models with a mix of the following paraphrase datasets.

PAWS-X Yang et al. (2019b) machine translate 49k examples from the PAWS (Zhang et al., 2019) dataset to six other languages. This dataset provides both positive and negative paraphrase examples.

Quora Question Pairs (QQP) Sharma et al. (2019) provide English question pair examples from Quora; we use the 384k examples from the training split of Wang et al. (2017). This dataset provides both positive and negative examples of English paraphrases.

4.2 Query Matching Baselines: LRL Query → HRL Query

We consider a combination of translation techniques and cross-lingual sentence encoders to find semantically equivalent queries across languages. We select from pretrained models which report strong results on similar multilingual tasks, or finetune representations for our task using publicly available paraphrase datasets (4.1.2).[3]. Each finetuned model receives basic hyperparameter tuning over the learning rate and the ratio of training data from PAWS-X and QQP.[4]

NMT + MIPS We use a many-to-many, Transformer-based (Vaswani et al., 2017), encoder-decoder neural machine translation system, trained on the OPUS multilingual corpus covering 100 languages (Zhang et al., 2020). To match the translation to an English query, we use the Universal Sentence Encoder (USE) (Cer et al., 2018) to perform maximal inner product search (MIPS).

Pretrained Single Encoders We consider pretrained multilingual sentence encoders for sentence retrieval. We explore mUSE[5] (Yang et al., 2019a), LASER (Artetxe and Schwenk, 2019), and m-SentenceBERT as the Single Encoder (Reimers and Gurevych, 2019).

[3]Retriever-Reader models do not fit in the Cross-Lingual Pivots task due to requiring document retrieval, but assuming perfect cross-lingual retrieval/reading, these systems would perform as well as *Perfect LRL → HRL* in Tables 2 and 3

[4]We used an optimal learning rate of 1e-5, and training data ratio of 75% PAWS-X and 25% QQP.

[5]mUSE was only trained on the following 16 languages: ar, ch_cn, ch_tw, en, fr, de, it, ja, ko da, pl, pt, es, th, tr ru

MKQA + Natural Questions	LRL → HRL (Acc.)				LRL → HRL → LRL (F1)			
Language Groups	All	High	Medium	Low	All	High	Medium	Low
NMT + MIPS	74.4 ± 15.8	78.8 ± 13.3	78.3 ± 10.0	57.7 ± 19.0	65.8 ± 16.3	70.7 ± 14.5	69.9 ± 11.0	47.8 ± 17.0
mUSE	71.8 ± 21.2	**88.2** ± 4.4	57.8 ± 20.4	73.2 ± 19.6	62.8 ± 18.3	**77.8** ± 8.9	52.6 ± 16.9	58.2 ± 15.8
LASER	74.2 ± 15.0	70.0 ± 14.6	82.6 ± 8.5	63.3 ± 16.8	65.4 ± 15.4	62.8 ± 14.3	73.6 ± 9.4	52.0 ± 16.6
Single Encoder (XLM-R)	73.0 ± 6.8	72.6 ± 3.7	73.4 ± 8.3	72.6 ± 7.3	63.2 ± 8.1	63.9 ± 4.9	65.4 ± 8.9	57.1 ± 8.0
RM-MIPS (mUSE)	78.2 ± 12.5	86.9 ± 3.1	71.9 ± 12.5	76.7 ± 14.0	68.1 ± 12.4	76.3 ± 8.0	64.9 ± 11.3	60.4 ± 12.7
RM-MIPS (LASER)	80.1 ± 9.4	79.5 ± 7.8	**83.7** ± 5.6	73.1 ± 13.6	69.4 ± 11.2	70.0 ± 9.3	74.1 ± 7.3	57.8 ± 13.2
RM-MIPS (XLM-R)	**83.5** ± 5.2	84.9 ± 2.7	**83.7** ± 5.7	**80.7** ± 6.1	**72.0** ± 9.3	74.7 ± 7.6	**74.2** ± 7.7	**62.7** ± 9.5
Perfect LRL → HRL	-	-	-	-	90.1 ± 7.3	91.8 ± 7.1	92.4 ± 4.2	81.9 ± 7.5

Table 2: **MKQA results by language group with MKQA + Natural Questions as the HRL Database**: (left) the accuracy for the LRL → HRL Query Matching stage; (right) the F1 scores for the End-to-End XLP task, using WikiData translation for Answer Translation; and (bottom) the F1 score only for Wikidata translation, assuming Query Matching (LRL → HRL) was perfect. Macro standard deviation are computed for language groups (±). The difference between all method pairs are significant.

XQuAD + SQuAD	LRL → HRL (Acc.)				LRL → HRL → LRL (F1)			
Language Group	All	High	Medium	Low	All	High	Medium	Low
NMT + MIPS	77.7 ± 14.4	78.4 ± 21.4	76.5 ± 4.7	78.0 ± 8.0	24.5 ± 12.0	28.8 ± 17.3	24.5 ± 3.3	18.7 ± 3.8
mUSE	68.0 ± 38.5	**94.5** ± 3.0	66.4 ± 34.5	34.2 ± 40.7	21.1 ± 15.8	**31.9** ± 15.6	20.3 ± 9.8	7.3 ± 7.8
LASER	46.7 ± 24.9	54.7 ± 24.3	63.9 ± 1.6	18.8 ± 10.9	15.2 ± 11.6	20.1 ± 14.1	19.9 ± 2.3	4.1 ± 2.3
Single Encoder (XLM-R)	81.4 ± 6.2	85.1 ± 1.9	79.4 ± 9.4	78.6 ± 2.2	24.3 ± 10.8	29.1 ± 14.4	24.5 ± 5.3	17.7 ± 3.0
RM-MIPS (mUSE)	72.0 ± 34.0	**94.4** ± 2.5	75.1 ± 25.4	39.1 ± 37.8	22.4 ± 14.7	**31.8** ± 15.4	23.7 ± 6.0	8.5 ± 6.9
RM-MIPS (LASER)	69.2 ± 23.7	77.5 ± 14.8	85.4 ± 3.0	41.9 ± 21.8	21.2 ± 12.3	26.7 ± 14.3	26.0 ± 3.1	9.2 ± 4.0
RM-MIPS (XLM-R)	**92.2** ± 2.4	93.4 ± 1.7	**90.4** ± 2.7	**92.3** ± 1.4	**27.2** ± 10.8	31.5 ± 15.2	**27.4** ± 3.1	**21.2** ± 2.8
Perfect LRL → HRL	-	-	-	-	46.6 ± 13.1	51.0 ± 15.5	51.2 ± 5.0	36.3 ± 8.4

Table 3: **XQuAD results by language group with XQuAD + SQuAD as the HRL Database**: (left) the accuracy for the LRL → HRL Query Matching stage; (right) the F1 scores for the End-to-End XLP task, using machine translation to translate answers from HRL → LRL; and (bottom) the F1 score only for Wikidata translation, assuming Query Matching (LRL → HRL) was perfect. Macro standard deviations are computed for language groups (±). The difference between all method pairs are significant.

Finetuned Single Encoders We finetune transformer encoders to embed sentences, per Reimers and Gurevych (2019). We use the softmax loss over the combination of $[x; y; |x - y|]$ from Conneau et al. (2017a) and mean pool over the final encoder representations to obtain the final sentence representation. We use XLM-R Large as the base encoder (Conneau et al., 2019).

Cross Encoders We finetune XLM-R Large (Conneau et al., 2019) which is pretrained using the multilingual masked language modelling (MLM) objective.[6] For classification, a pair of sentences are given as input for classification, taking advantage of cross-attention between sentences.

4.3 Answer Translation: HRL Answer → LRL Answer

Once we've found an English (HRL) query using RM-MIPS, or one of our "Query Matching" baselines, we can use the English database to lookup the English answer. Our final step is to generate an equivalent answer in the target (LRL) language.

We explore straightforward methods of answer generation, including basic neural machine translation (NMT), and WikiData entity translation.

Machine Translation For NMT we use our many-to-many neural machine translation as described in Section 4.2.

WikiData Entity Translation We propose our WikiData entity translation method for QA datasets with primarily entity type answers that would likely appear in the WikiData knowledge graph (Vrandečić and Krötzsch, 2014).[7] This method uses a named entity recognizer (NER) with a WikiData entity linker to find an entity (Honnibal and Montani, 2017). [8] We train our own entity linker on the public WikiData entity dump according to spaCy's instructions. If a WikiData entity is found, its structured metadata often contains the equivalent term in the target language, localized to the relevant script/alphabet. For our implementation, when a WikiData entity is not found, or its translation is not available in the target language, we simply return

[6] We use the pretrained Transformer encoder implementations in the Huggingface library (Wolf et al., 2019).

[7] https://www.wikidata.org

[8] https://github.com/explosion/spaCy

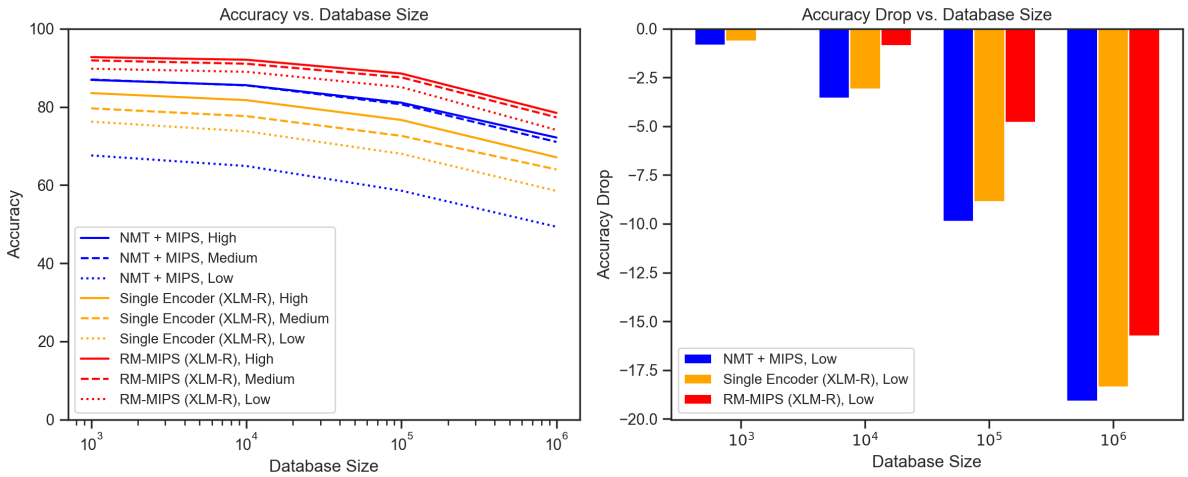

Figure 3: **Effect of Database Size on LRL → HRL.** Left: Query Matching accuracy of the strongest methods on different language groups as the amount of "unaligned" queries in the English database increases. Right: The accuracy drop of the different methods on low resource languages as the amount of queries in the English database increases beyond the original parallel count.

the English answer.

For XQuAD end-to-end experiments we find straightforward machine translation works best, whereas for MKQA, which contains more short, entity-type answers, we find WikiData Entity Translation works best. We report results using these simple methods and leave more sophisticated combinations or improvements to future work.

5 Results

5.1 End-To-End (E2E) Results

We benchmark the performance of the cross-lingual pivot methods on XQuAD and MKQA. To simulate a realistic setting, we add all the English questions from SQuAD to the English database used in the XQuAD experiments. Similarly we add all of Natural Questions queries (not just those aligned across languages) in the MKQA experiments. For each experiment we group the languages into high, medium, and low resource, as shown in Table 1, according to Wu and Dredze (2020). Tables 2 and 3 present the mean performance by language group, for query matching (LRL → HRL), and end-to-end results (LRL → HRL → LRL), query matching and answer translation in sequence.

Among the models, RM-MIPS typically outperforms baselines, particularly on lower resource languages. We find the reranking component in particular offers significant improvements over the non-reranked sentence encoding approaches in low resource settings, where we believe sentence embeddings are most inconsistent in their performance. For instance, RM-MIPS (LASER) outperforms LASER by 5.7% on the Lowest resource E2E MKQA task, and 4.0% across all languages. The margins are even larger between RM-MIPS (mUSE) and mUSE as well as RM-MIPS (XLM-R) and XLM-R.

For certain high resource languages, mUSE performs particularly strongly, and for XQuAD languages, LASER performs poorly. Accordingly, the choice of sentence encoder (and its language proportions in pretraining) is important in optimizing for the cross-lingual pivot task. The modularity of RM-MIPS offers this flexibility, as the first stage multiligual encoder can be swapped out: we present results for LASER, mUSE, and XLM-R.

Comparing query matching accuracy (left) and end-to-end F1 (right) in Tables 2 and 3 measures the performance drop due to answer translation (HRL → LRL, see section 4.3 for details). We see this drop is quite small for MKQA as compared to XQuAD. Similarly, the "Perfect LRL → HRL" measures the Answer Translation stage on all queries, showing XQuAD's machine translation for answers is much lower than MKQA's Wikidata translation for answers. This observation indicates that (a) Wikidata translation is particularly strong, and (b) cross-lingual pivot techniques are particularly useful for datasets with frequent entity, date, or numeric-style answers, that can be translated with Wikidata, as seen in MKQA. Another potential factor in the performance difference between MKQA and XQuAD is that MKQA contains naturally occurring questions, whereas XQuAD does not. Despite the lower mean end-to-end perfor-

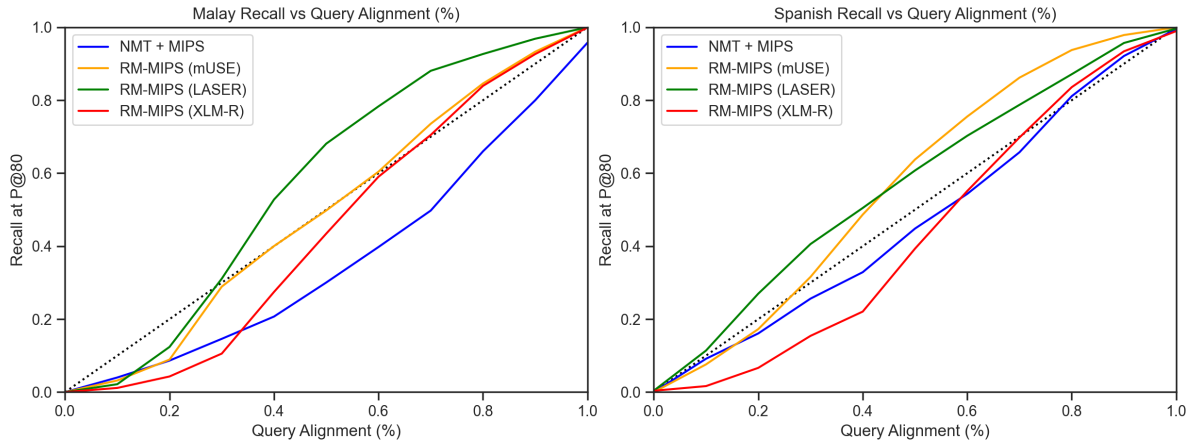

Figure 4: **Effects of Query Alignment on MKQA end-to-end Performance:** At a target precision of 80%, the end-to-end Malay (left) and Spanish (right) recall are plotted for each degree of query alignment. The query alignment axis indicates the percentage of 10k queries with parallel matches retained in the English database.

mance for XQuAD, this cross-lingual pivot can still be used alongside traditional methods, and can be calibrated for high precision/low coverage by abstaining from answering questions that are Wikidata translatable.

One other noteable advantage of paraphrase-based pivot approaches, is that no LRL-specific annotated training data is required. A question answering system in the target language requires in-language annotated data, or an NMT system from English. Traditional NMT "translate test" or "MT-in-the-middle" (Asai et al., 2018; Hajič et al., 2000; Schneider et al., 2013) approaches also require annotated parallel data to train. RM-MIPS and our other paraphrase baselines observe monolingual corpora at pre-training time, and only select language pairs during fine-tuning (those present in PAWS-X), and yet these models still perform well on XLP even for non-PAWS-X languages.

5.2 Database Size

To understand the impact of database size on the query matching process, we assemble a larger database with MSMARCO (800k), SQuAD (100k), and Open-NaturalQuestions (90k). Note that none of the models are explicitly tuned to MKQA, and since MSMARCO and Open-NQ comprise natural user queries (from the same or similar distribution), we believe these are challenging "distractors". In Figure 3 we plot accuracy of the most performant models from Tables 2 and 3 on each of the high, medium, and low resource language groups over different sizes of database on MKQA. We report the initial stage query matching (LRL → HRL) to isolate individual model matching performance. We observe that RM-MIPS degrades less quickly with database size than competing methods, and that it degrades less with the resourcefulness of the language group.

5.3 Query Alignment

In some cases, incoming LRL queries may not have a corresponding semantic match in the HRL database. To assess the impact of this, we vary the percentage of queries that have a corresponding match by dropping out their parallel example in the English database (in increments of 10%). In Figures 4 we report the median end-to-end recall scores over five different random seeds, at each level of query alignment (x-axis). At each level of answer query alignment we recompute a No Answer confidence threshold for a target precision of 80%. Due to computational restraints, we select one low resource (Malay) and one high resource language (Spanish) to report results on. We find that even calibrated for high precision (a target of 80%) the cross-lingual pivot methods can maintain proportional, and occasionally higher, coverage to the degree of query misalignment. RM-MIPS methods in particular can *outperform* proportional coverage to alignment (the dotted black line on the diagonal) by sourcing answers from similar queries in the database to those dropped out. Consequently, a practitioner can maintain high precision and respectable recall by selecting a threshold for any degree of query misalignment observed in their test distribution.

The primary limitation of RM-MIPS, or other pivot-oriented approaches, is that their performance is bounded by the degree of query alignment. How-

ever, QA systems still fail to replicate their English answer coverage in LRLs (Longpre et al., 2020), and so we expect pivot techniques to remain essential until this gap narrows completely.

6 Related Work

Cross-Lingual Modeling Multilingual BERT (Devlin et al., 2019), XLM (Lample and Conneau, 2019), and XLM-R (Conneau et al., 2019) use masked language modeling (MLM) to share embeddings across languages. Artetxe and Schwenk (2019) introduce LASER, a language-agnostic sentence embedder trained using many-to-many machine translation. Yang et al. (2019a) extend Cer et al. (2018) in a multilingual setting by following Chidambaram et al. (2019) to train a multi-task dual-encoder model (mUSE). These multilingual encoders are often used for semantic similarity tasks. Reimers and Gurevych (2019) propose finetuning pooled BERT token representations (Sentence-BERT), and Reimers and Gurevych (2020) extend with knowledge distillation to encourage vector similarity among translations. Other methods improve multilingual transfer via language alignment (Roy et al., 2020; Mulcaire et al., 2019; Schuster et al., 2019) or combining machine translation with multilingual encoders (Fang et al., 2020; Cui et al., 2019; Mallinson et al., 2018).

Multilingual Question Answering Efforts to explore multilingual question answering include MLQA (Lewis et al., 2019), XQuAD (Artetxe et al., 2019), MKQA (Longpre et al., 2020), TyDi (Clark et al., 2020), XORQA (Asai et al., 2020) and MFAQ (De Bruyn et al., 2021).

Prior work in multilingual QA achieves strong results combining neural machine translation and multilingual representations via **Translate-Test**, **Translate-Train**, or **Zero Shot** approaches (Asai et al., 2018; Cui et al., 2019; Charlet et al., 2020; Stepanov et al., 2013; He et al., 2013; Dong et al., 2017). This work focuses on *extracting* the answer from a multilingual passage (Cui et al., 2019; Asai et al., 2018), assuming passages are provided.

Improving Low Resource With High Resource Efforts to improve performance on low-resource languages usually explore language alignment or transfer learning. Chung et al. (2017) find supervised and unsupervised improvements in transfer learning when finetuning from a language specific model, and Lee and Lee (2019) leverage a GAN-inspired discriminator (Goodfellow et al., 2014) to enforce language-agnostic representations. Aligning vector spaces of text representations in existing models (Conneau et al., 2017b; Schuster et al., 2019; Mikolov et al., 2013) remains a promising direction. Leveraging high resource data has also been studied in sequence labeling (Xie et al., 2018; Plank and Agić, 2018; Schuster et al., 2019) and machine translation (Johnson et al., 2017; Zhang et al., 2020).

Paraphrase Detection The paraphrase detection task determines whether two sentences are semantically equivalent. Popular paraphrase datasets include Quora Question Pairs (Sharma et al., 2019), MRPC (Dolan and Brockett, 2005), and STS-B (Cer et al., 2017). The adversarially constructed PAWS dataset Zhang et al. (2019) was translated to 6 languages, offering a multilingual option, PAWS-X Yang et al. (2019b). In a multilingual setting, an auxiliary paraphrase detection (or nearest neighbour) component, over a datastore of training examples, has been shown to greatly improve performance for neural machine translation (Khandelwal et al., 2020).

7 Conclusion

In conclusion, we formulate a task to cross-lingual open-retrieval question answering more realistic to the constraints and challenges faced by practitioners expanding their systems' capabilities beyond English. Leveraging access to a large English training set our method of query retrieval followed by reranking greatly outperforms strong baseline methods. Our analysis compares multiple methods of leveraging this English expertise and concludes our two-stage approach transfers better to lower resource languages, and is more robust in the presence of extensive distractor data and query distribution misalignment. Circumventing retrieval, this approach offers fast online or offline answer generation to many languages straight off-the-shelf, without necessitating additional training data in the target language.

We hope this analysis will promote creative methods in multilingual knowledge transfer, and the cross-lingual pivots task will encourage researchers to pursue problem formulations better informed by the needs of existing systems. In particular, leveraging many location and culturally-specific query knowledge bases, with cross-lingual pivots across many languages is an exciting extension of this work.

References

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.

Akari Asai, Akiko Eriguchi, Kazuma Hashimoto, and Yoshimasa Tsuruoka. 2018. Multilingual extractive reading comprehension by runtime machine translation. *arXiv preprint arXiv:1809.03275*.

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

Ewa S. Callahan and Susan C. Herring. 2011. Cultural bias in wikipedia content on famous persons. *Journal of the American Society for Information Science and Technology*, 62(10):1899–1915.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder for english. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174.

Delphine Charlet, Geraldine Damnati, Frederic Bechet, Gabriel Marzinotto, and Johannes Heinecke. 2020. Cross-lingual and cross-domain evaluation of machine reading comprehension with squad and CALOR-quest corpora. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5491–5497, Marseille, France. European Language Resources Association.

Swapnil Chaudhari. 2014. Cross lingual information retrieval. *Center for Indian Language Technology*.

Muthu Chidambaram, Yinfei Yang, Daniel Cer, Steve Yuan, Yunhsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Learning cross-lingual sentence representations via a multi-task dual-encoder model. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 250–259.

Yu-An Chung, Hung-Yi Lee, and James Glass. 2017. Supervised and unsupervised transfer learning for question answering. *arXiv preprint arXiv:1711.05345*.

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. *arXiv preprint arXiv:2003.05002*.

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, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017a. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017b. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2019. 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.

Maxime De Bruyn, Ehsan Lotfi, Jeska Buhmann, and Walter Daelemans. 2021. Mfaq: a multilingual faq dataset. *arXiv preprint arXiv:2109.12870*.

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.

Bill Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Third International Workshop on Paraphrasing (IWP2005)*.

Li Dong, Jonathan Mallinson, Siva Reddy, and Mirella Lapata. 2017. Learning to paraphrase for question answering. *arXiv preprint arXiv:1708.06022*.

Yuwei Fang, Shuohang Wang, Zhe Gan, Siqi Sun, and Jingjing Liu. 2020. Filter: An enhanced fusion method for cross-lingual language understanding. *arXiv preprint arXiv:2009.05166*.

Christian Fluhr, Robert E Frederking, Doug Oard, Akitoshi Okumura, Kai Ishikawa, and Kenji Satoh. 1999. Multilingual (or cross-lingual) information retrieval. *Proceedings of the Multilingual Information Management: Current Levels and Future Abilities*.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Miniwatts Marketing Group. 2011. Internet world stats: Usage and population statistics. *Miniwatts Marketing Group*.

Jan Hajič, Jan Hric, and Vladislav Kuboň. 2000. Machine translation of very close languages. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*, ANLC '00, page 7–12, USA. Association for Computational Linguistics.

Xiaodong He, Li Deng, Dilek Hakkani-Tur, and Gokhan Tur. 2013. Multi-style adaptive training for robust cross-lingual spoken language understanding. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8342–8346. IEEE.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

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

Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Nearest neighbor machine translation. *arXiv preprint arXiv:2010.00710*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

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

Manpreet Lehal. 2018. Challenges in cross language information retrieval.

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

Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. 2020. Question and answer test-train overlap in open-domain question answering datasets. *arXiv preprint arXiv:2008.02637*.

Shayne Longpre, Yi Lu, and Joachim Daiber. 2020. Mkqa: A linguistically diverse benchmark for multilingual open domain question answering.

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2017. Paraphrasing revisited with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 881–893.

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2018. Sentence compression for arbitrary languages via multilingual pivoting. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2453–2464.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Phoebe Mulcaire, Jungo Kasai, and Noah A Smith. 2019. Polyglot contextual representations improve crosslingual transfer. 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 3912–3918.

Barbara Plank and Željko Agić. 2018. Distant supervision from disparate sources for low-resource part-of-speech tagging. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 614–620, Brussels, Belgium. Association for Computational Linguistics.

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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. 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 3973–3983.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. *arXiv preprint arXiv:2004.09813*.

Uma Roy, Noah Constant, Rami Al-Rfou, Aditya Barua, Aaron Phillips, and Yinfei Yang. 2020. Lareqa: Language-agnostic answer retrieval from a multilingual pool. *arXiv preprint arXiv:2004.05484*.

Nathan Schneider, Behrang Mohit, Chris Dyer, Kemal Oflazer, and Noah A. Smith. 2013. Supersense tagging for Arabic: the MT-in-the-middle attack. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 661–667, Atlanta, Georgia. Association for Computational Linguistics.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. 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 1599–1613.

Lakshay Sharma, Laura Graesser, Nikita Nangia, and Utku Evci. 2019. Natural language understanding with the quora question pairs dataset. *arXiv preprint arXiv:1907.01041*.

Evgeny A Stepanov, Ilya Kashkarev, Ali Orkan Bayer, Giuseppe Riccardi, and Arindam Ghosh. 2013. Language style and domain adaptation for cross-language slu porting. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 144–149. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

Zhiguo Wang, Wael Hamza, and Radu Florian. 2017. Bilateral multi-perspective matching for natural language sentences. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 4144–4150.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A. Smith, and Jaime Carbonell. 2018. Neural cross-lingual named entity recognition with minimal resources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 369–379, Brussels, Belgium. Association for Computational Linguistics.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, et al. 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.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. *arXiv preprint arXiv:2004.11867*.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. Paws: Paraphrase adversaries from word scrambling. 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 1298–1308.

Imed Zitouni and Radu Florian. 2008. Mention detection crossing the language barrier. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 600–609, Honolulu, Hawaii. Association for Computational Linguistics.

A Reproducibility

A.1 Experimental Setup

Computing Infrastructure. For all of our experiments, we used a computation cluster with 4 NVIDIA Tesla V100 GPUs, 32GB GPU memory and 256GB RAM.

Implementation We used Python 3.7, PyTorch 1.4.0, and Transformers 2.8.0 for all our experiments. We obtain our datasets from the citations specified in the main paper, and link to the repositories of all libraries we use.

Hyperparameter Search For our hyper parameter searches, we perform a uniformly random search over learning rate and batch size, with ranges specified in Table 4, optimizing for the development accuracy. We find the optimal learning rate and batch size pair to be $1e-5$ and 80 respectively.

Evaluation For query matching, we use scikit-learn [9] to calculate the accuracy. For end-to-end performance, we use the MLQA evaluation script to obtain the F1 score of the results[10].

Datasets We use the sentences in each dataset as-is, and rely on the pretrained tokenizer for each model to perform preprocessing.

A.2 Model Training

Query Paraphrase Dataset We found the optimal training combination of the PAWS-X and QQP datasets by training XLM-R classifiers on training dataset percentages of $(100\%, 0\%)$, $(75\%, 25\%)$, and $(50\%, 50\%)$ of (PAWS-X, QQP) – with the PAWS-X percentage entailing the entirety of the PAWS-X dataset – and observe the performance on matching multilingual XQuAD queries. We shuffle the examples in the training set, and restrict the input examples to being (English, LRL) pairs. We perform a hyperparameter search as specified in Table 5 for each dataset composition, and report the test results in Table 4.

A.3 Cross Encoder

We start with the pretrained `xlm-roberta-large` checkpoint in Huggingface's transformers[11] library and perform a hyperparameter search with the parameters specified in Table 1 by using a modified version of Huggingface's text classification training pipeline for GLUE.

The cross encoder was used in all the RM-MIPS methods. In particular, it was used in the RM-MIPS (mUSE), RM-MIPS (LASER), and RM-MIPS (XLM-R) rows of tables in the main paper.

[9] https://scikit-learn.org/stable/

[10] https://github.com/facebookresearch/MLQA

[11] https://github.com/huggingface/transformers

(PAWS-X, QQP)	XQuAD
$(100\%, 0\%)$	0.847
$(75\%, 25\%)$	**0.985**
$(50\%, 50\%)$	0.979

Table 4: **XLM-R Query Paraphrase Performance On Different Query Compositions.** The performance of XLM-Roberta on matching XQuAD test queries when finetuned on different training set compositions of PAWS-X and QQP.

Model Parameters	Value/Range
Fixed Parameters	
Model	XLM-Roberta Large
Num Epochs	3
Dropout	0.1
Optimizer	Adam
Learning Rate Schedule	Linear Decay
Max Sequence Length	128
Tuned Parameters	
Batch Size	[8, 120]
Learning Rate	$[9e-4, 1e-6]$
Extra Info	
Model Size (# params)	$550M$
Vocab Size	250,002
Trials	30

Table 5: **Cross Encoder Hyperparameter Selection And Tuning Ranges** The hyper parameters we chose and searched over for XLM-Roberta large on the query paraphrase detection datasets.

B Full Results Breakdowns

B.1 LRL→HRL Results

See Table 6 and 7 for the non-aggregated LRL→HRL language performances of each method on MKQA and XQuAD respectively.

B.2 LRL→HRL→LRL Results

See Table 8 and 9 for the non-aggregated LRL→HRL→LRL language performances of each method on MKQA and XQuAD respectively.

	ar	zh_{cn}	da	de	es	fi	fr	he	zh_{hk}	hu	it	ja	km
NMT + MIPS	69.2	48.0	89.8	87.5	86.5	76.0	87.6	74.3	42.5	79.1	86.6	62.0	45.4
mUSE	80.0	83.2	51.7	90.9	91.7	37.6	91.5	33.5	80.8	40.7	91.6	80.0	35.6
LASER	81.5	62.8	88.6	52.0	79.9	81.6	78.5	85.5	64.0	69.1	80.4	39.3	40.2
Single Encoder (XLM-R)	58.0	76.3	84.8	74.6	73.3	65.5	74.1	67.8	77.0	66.9	69.0	71.4	59.0
RM-MIPS (mUSE)	77.6	81.2	77.2	88.8	88.9	59.9	88.8	44.1	81.2	64.1	88.4	81.2	50.6
RM-MIPS (LASER)	77.2	77.7	89.2	66.9	84.7	84.8	84.4	83.3	78.1	77.5	84.7	64.2	48.2
RM-MIPS (Ours)	72.6	80.7	90.1	86.8	87.0	82.7	87.2	80.6	81.0	81.4	85.5	79.5	72.4

	ko	ms	nl	no	pl	pt	ru	sv	th	tr	zh_{tw}	vi
NMT + MIPS	54.2	86.0	88.8	87.2	81.9	87.4	81.9	87.2	75.0	79.6	39.7	76.0
mUSE	73.7	87.6	92.0	50.3	84.9	93.3	87.2	50.3	88.6	87.0	73.2	38.6
LASER	68.6	92.5	93.1	92.4	73.7	85.2	78.1	92.8	62.1	75.2	57.9	79.9
Single Encoder (XLM-R)	72.3	76.4	79.1	81.3	70.6	65.7	78.8	83.8	79.4	68.7	71.2	78.6
RM-MIPS (mUSE)	74.7	89.9	90.9	75.6	87.3	89.8	87.1	76.0	86.8	86.6	75.0	64.4
RM-MIPS (LASER)	73.1	89.5	90.2	89.7	81.9	86.7	84.3	89.8	77.0	82.3	72.5	84.0
RM-MIPS (XLM-R)	75.2	89.0	89.8	88.8	85.6	85.5	86.1	90.0	85.4	83.6	75.5	85.9

Table 6: **MKQA + Natural Questions Per-Language LRL→HRL Results.** The accuracy scores for each method on query matching.

	ar	de	el	es	hi	ru	th	tr	vi	zh
NMT + MIPS	71.7	90.8	86.7	95.2	79.9	85.7	67.4	82.9	74.8	41.8
mUSE	87.4	96.4	7.5	98.1	3.4	93.2	91.6	94.1	17.8	90.3
LASER	61.7	33.1	3.7	86.2	28.6	70.4	24.2	65.3	64.7	29.2
Single Encoder (XLM-R)	66.8	85.1	81.7	87.8	77.6	85.0	76.6	81.9	89.4	82.3
RM-MIPS (mUSE)	90.4	96.3	14.8	97.3	10.1	93.2	92.6	95.7	39.3	91.0
RM-MIPS (LASER)	81.6	59.9	11.1	95.5	59.1	88.3	55.5	89.0	85.7	66.2
RM-MIPS (XLM-R)	86.6	94.2	94.1	95.5	92.0	93.0	90.7	92.5	92.1	90.8

Table 7: **XQuAD + SQuAD Per-Language LRL→HRL Results.** The accuracy scores for each method on query matching.

	ar	zh_{cn}	da	de	es	fi	fr	he	zh_{hk}	hu	it	ja	km
NMT + MIPS	60.0	41.7	85.8	83.8	82.4	72.0	83.7	63.3	41.2	74.5	82.5	60.1	44.8
mUSE	68.6	62.7	50.1	87.2	87.4	37.2	87.5	31.9	68.7	40.0	87.2	74.9	35.0
LASER	70.1	49.5	84.6	50.8	76.3	77.3	75.3	72.8	56.2	65.0	76.8	39.1	38.1
Single Encoder (XLM-R)	50.9	57.5	81.0	71.7	70.2	62.0	70.9	58.6	65.8	63.1	66.0	68.0	54.9
RM-MIPS (mUSE)	66.9	61.3	74.4	85.2	84.8	58.0	84.9	39.9	68.8	61.5	84.1	75.8	46.0
RM-MIPS (LASER)	66.7	59.0	85.0	64.6	80.7	80.3	80.6	71.0	66.3	72.7	80.6	61.7	45.3
RM-MIPS (Ours)	62.8	60.8	85.9	83.3	83.1	78.4	83.3	68.7	68.6	76.6	81.5	74.4	64.4

	ko	ms	nl	no	pl	pt	ru	sv	th	tr	zh_{tw}	vi
NMT + MIPS	47.5	81.1	85.3	80.2	77.6	83.3	72.6	84.1	62.9	74.7	35.2	70.6
mUSE	63.0	82.7	88.5	48.4	80.4	88.9	77.2	49.4	72.2	81.7	55.6	37.7
LASER	59.1	87.4	89.7	85.1	70.0	81.2	69.4	89.5	53.7	70.7	45.7	74.4
Single Encoder (XLM-R)	62.5	72.2	76.0	75.2	67.0	62.5	70.1	80.8	66.3	64.6	54.1	73.1
RM-MIPS (mUSE)	64.2	84.8	87.3	70.6	82.5	85.3	77.1	73.9	70.7	81.2	56.6	61.3
RM-MIPS (LASER)	63.1	84.4	86.7	81.8	77.3	82.4	74.7	86.6	64.3	77.1	55.1	78.2
RM-MIPS (XLM-R)	64.7	84.0	86.3	81.6	81.0	81.3	76.3	86.9	69.9	78.6	56.8	79.9

Table 8: **MKQA + Natural Questions Per-Language LRL→HRL→LRL WikiData Results.** The F1 scores for end-to-end performance of each method on every language when using WikiData translation

	ar	de	el	es	hi	ru	th	tr	vi	zh
NMT + MIPS	35.3	55.5	39.2	68.2	32.9	30.7	17.8	42.1	45.6	19.0
mUSE	40.8	58.2	4.4	70.0	1.6	33.4	23.4	47.0	11.8	33.6
LASER	29.9	22.7	1.5	61.8	10.8	24.2	6.4	33.0	38.6	12.7
Single Encoder (XLM-R)	31.3	52.9	37.3	63.9	30.9	30.1	18.6	42.0	52.7	30.6
RM-MIPS (mUSE)	42.6	58.1	7.8	69.6	4.2	33.4	23.2	47.5	26.1	33.8
RM-MIPS (LASER)	38.3	38.2	5.7	68.3	22.9	31.1	13.7	44.5	50.7	26.3
RM-MIPS (XLM-R)	40.9	57.3	42.1	68.7	36.7	33.0	22.9	45.7	54.5	33.6

Table 9: **XQuAD + SQuAD Per-Language LRL→HRL→LRL NMT Results.** The F1 scores for end-to-end performance of each method on every language when using NMT translation