Answering the Unanswered: Overcoming Language Barriers in Question Answering

Archit Bansal
ab2465@rutgers.edu

Shreesh Keskar
skk139@rutgers.edu

Saransh Sharma
ss4368@rutgers.edu

Devarata Oza
do309@rutgers.edu

Ayush Patel
ap2244@rutgers.edu

1 Introduction

Natural Language Processing (NLP) tasks, including classification and generation have taken vast strides in recent years, especially with the introduction of Large Language Models (LLMs) such as GPT-4 and applications like ChatGPT. These tasks, however, are benchmarked and perform well on languages like English and Chinese for which a large volume and variety of data is available.

For several other (regional) languages, representing communities with smaller populations and low income, there is a severe lack of data and consequently that of reliable performance benchmarks for the aformentioned NLP tasks.

In this project, we have chosen to tackle the Question Answering task. We aim to validate existing benchmarks and provide new ones where they are insufficient/absent, by leveraging and extending existing works in the Question Answering domain.

2 Related Work

2.1 mBERT

BERT [Devlin et al., 2019] is a popular language model which has been used in many transformer-based language models. mBERT (multilingual BERT) uses the same architecture as BERT. It is trained on a dataset consisting of Wikipedia pages of 104 languages and it is fine-tuned on the XLNI dataset. A variety of experiments by Pires et al. [2019] show that the efficacy of mBERT does not depend simply on vocabulary memorization but learns a deeper multilingual representation, which leads to a large gap between the performance of fine-tuned vanilla BERT and mBERT.

2.2 XLM-RoBERTa

XLM-RoBERTa [Conneau et al., 2020] is a more recent transformer-based model. It was pretrained with the Masked Language Modeling (MLM) objective. In the XLM-RoBERTa work, it is shown, similar to RoBERTa [Liu et al., 2019], that XLM and mBERT are undertuned, hence this work augments the data [Conneau et al., 2020], especially for low-resource languages, building the CommonCrawl dataset. Conneau et al. [2020] also evaluate on a wider range of tasks beyond Part-of-Speech Tagging, Named-Entity-Recognition and XNLI, namely Cross-Lingual Question Answering (of interest to this project) and GLUE Benchmark [Wang et al., 2019].

Translation Language Modeling is an extension of MLM, where it randomly masks words in both the source and target sentences. As shown in Figure 4, to predict a word masked in an English sentence, the model can either attend to surrounding English words or to the French translation. So the model

Language	Latin script[a]	White space tokens	Sentence boundaries	Word formation[b]	Gender[c]	Prodrop
English	+	+	+	+	+[d]	—
Arabic	—	+	+	++	+	+
Bengali	—	+	+	+	+	+
Finnish	+	+	+	+++	—	—
Indonesian	+	+	+	+	—	+
Japanese	—	—	+	+	—	+
Kiswahili	+	+	+	+++	—[e]	+
Korean	—	+[f]	+	+++	—	+
Russian	+	+	+	++	+	+
Telugu	—	+	+	+++	+	+
Thai	—	—	—	+	+	+

[a] '—' indicates **Latin script** is not the conventional writing system. Intermixing of Latin script should still be expected.
[b] We include inflectional and derivation phenomena in our notion of **word formation**.
[c] We limit the **gender** feature to sex-based gender systems associated with coreferential gendered personal pronouns.
[d] English has grammatical gender only in third person personal and possessive pronouns.
[e] Kiswahili has morphological noun classes (Corbett, 1991), but here we note sex-based gender systems.
[f] In Korean, tokens are often separated by whitespace, but prescriptive spacing conventions are commonly flouted.

Question Word	TyDi QA	SQuAD
What	30%	51%
How	19%	12%
When	14%	8%
Where	14%	5%
(Yes/no)	10%	<1%
Who	9%	11%
Which	3%	5%
Why	1%	2%

Figure 1: Typological features of the 11 languages in TyDi QA; Distribution of question words in the English portion of the development data. [Clark et al., 2020]

Language	Train *(1-way)*	Dev *(3-way)*	Test *(3-way)*	Avg. Question Tokens	Avg. Article Bytes	Avg. Answer Bytes	Avg. Passage Candidates	%With Passage Answer	%With Minimal Answer
(English)	9,211	1031	1046	7.1	30K	57	47	50%	42%
Arabic	23,092	1380	1421	5.8	14K	114	34	76%	69%
Bengali	10,768	328	334	7.5	13K	210	34	38%	35%
Finnish	15,285	2082	2065	4.9	19K	74	35	49%	41%
Indonesian	14,952	1805	1809	5.6	11K	91	32	38%	34%
Japanese	16,288	1709	1706	—	14K	53	52	41%	32%
Kiswahili	17,613	2288	2278	6.8	5K	39	35	24%	22%
Korean	10,981	1698	1722	5.1	12K	67	67	26%	22%
Russian	12,803	1625	1637	6.5	27K	106	74	64%	51%
Telugu	24,558	2479	2530	5.2	7K	279	32	28%	27%
Thai	11,365	2245	2203	—	14K	171	38	54%	43%
Total	**166,916**	**18,670**	**18,751**						

Figure 2: Data statistics. [Clark et al., 2020]

can leverage the French context if the English one is not sufficient to infer the masked English words. This objective is effectively employed in the XLM-RoBERTa training process.

2.3 Datasets

These are some of the multilingual question answering datasets available:

- **SQuAD:** The Stanford Question Answering Dataset [Rajpurkar et al., 2016, 2018] which contains 100k+ answerable and 50k+ unanswerable question-answer pairs from 500+ English articles
- **chaii:** The Challenge in AI for India [cha] which contains 2k Hindi and 2k Tamil Question-Answer pairs
- **FLoRes:** The Facebook Low Resource [Team et al., 2022] which contains question-answer pairs for 101 languages, including several low-resource languages like Slovak, Ukrainian, Tajik, etc.

3 Baseline

TyDi QA [Clark et al., 2020] is a benchmark for information-seeking Question Answering in *Ty*pologically *Di*verse Languages developed by Google Research. The data consists of 200k question-answer pairs from 11 typographically diverse languages. It also contains unanswerable questions corresponding to each text entry. The data for non-English languages is obtained by translating from English. The baseline is obtained by fine-tuning mBERT on this TyDi QA data.

Figures 1, 2 & 5 show some of the details regarding the data, typographical diversity and language categories based on different typographical characteristics.

	Train Size	Passage Answer F1 (P/R)			Minimal Answer Span F1 (P/R)	
		First passage	mBERT	Lesser Human	mBERT	Lesser Human
(English)	9,211	32.9 (28.4/39.1)	62.5 (62.6/62.5)	69.4 (63.4/77.6)	44.0 (52.9/37.8)	54.4 (52.9/56.5)
Arabic	23,092	64.7 (59.2/71.3)	81.7 (85.7/78.1)	85.4 (82.1/89.0)	69.3 (74.9/64.5)	73.5 (73.6/73.5)
Bengali	10,768	21.4 (15.5/34.6)	60.3 (61.4/59.5)	85.5 (81.6/89.7)	47.7 (50.7/45.3)	79.1 (78.6/79.7)
Finnish	15,285	35.4 (28.4/47.1)	60.8 (58.7/63.0)	76.3 (69.8/84.2)	48.0 (56.7/41.8)	65.3 (61.8/69.4)
Indonesian	14,952	32.6 (23.8/51.7)	61.4 (57.2/66.7)	78.6 (72.7/85.6)	51.3 (54.5/48.8)	71.1 (68.7/73.7)
Japanese	16,288	19.4 (14.8/28.0)	40.6 (42.2/39.5)	65.1 (57.8/74.8)	30.4 (42.1/23.9)	53.3 (51.8/55.2)
Kiswahili	17,613	20.3 (13.4/42.0)	60.2 (58.4/62.3)	76.8 (70.1/85.0)	49.7 (55.2/45.4)	67.4 (63.4/72.1)
Korean	10,981	19.9 (13.1/41.5)	56.8 (58.7/55.3)	72.9 (66.3/82.4)	40.1 (45.2/36.2)	56.7 (56.3/58.6)
Russian	12,803	30.0 (25.5/36.4)	63.2 (65.3/61.2)	87.2 (84.4/90.2)	45.8 (51.7/41.2)	76.0 (82.0/70.8)
Telugu	24,558	23.3 (15.1/50.9)	81.3 (81.7/80.9)	95.0 (93.3/96.8)	74.3 (77.7/71.3)	93.3 (91.6/95.2)
Thai	11,365	34.7 (27.8/46.4)	64.7 (61.8/68.0)	76.1 (69.9/84.3)	48.3 (54.3/43.7)	65.6 (63.9/67.9)
OVERALL	166,916	30.2 (23.6/45.0)	63.1 (57.0/59.1)	79.9 (84.4/74.5)	50.5 (41.3/35.3)	70.1 (70.8/62.4)

Figure 3: Quality on the TyDi QA MinSpan using mBERT model. F1, precision, and recall measurements are averaged over four fine-tuning replicas for mBERT. [Clark et al., 2020]

There are three tasks that can be performed with the TyDi QA baseline:

- **Passage selection task (SelectP):** Given a list of the passages in the article, returns either (a) the index of the passage that answers the question or (b) NULL if no such passage exists.
- **Minimal answer span task (MinSpan):** Given the full text of an article, returns one of (a) the start and end byte indices of the minimal span that completely answers the question; (b) YES or NO if the question requires a yes/no answer and a conclusion can be drawn from the passage; (c) NULL if it is not possible to produce a minimal answer for that question.
- **Gold passage task (GoldP):** Given a passage that is guaranteed to contain the answer, predicts the single contiguous span of characters that answers the question.

We found that the MinSpan task is best suited for our project. Let's say we input a passage and a question to the mBERT model. The mBERT model tokenizes the words and converts them into numerical vectors. It generates a set of probability scores for each word in the passage, indicating how likely it is that each word is part of the answer span. The probability scores are then used to identify the start and end points of the answer span in the passage. The answer is then extracted from the passage and returned as the predicted answer to the question.

The baseline evaluation results for MinSpan task are shown in Figure 3. To evaluate the performance of the mBERT model, the F1 scores are calculated separately for each language. To compute the F1 score, the model's predictions are compared to the gold standard answers provided in the dataset. If the predicted answer span overlaps with the gold standard, it is considered a true positive, and a partial score is assigned in case of some overlap above a pre-defined threshold. We have used the Jaccard Index (Section A.1) to calculate the extent of overlap of our prediction with the ground truth in the evaluation to determine what score should be assigned to the answer in the MinSpan task where the answer is not a YES/NO or NULL; we have used the empirically determined threshold 0.71 for the same.

If the predicted answer span does not overlap with the gold standard answer span, it is considered a false positive. If the gold standard answer span is not included in the predicted answer span, it is considered a false negative. To reduce the training bias, the mBERT model is fine-tined four times for every language, and the scores are averaged across these iterations.

4 Methodology

Our approach follows the following three steps:

Table 1: Minimal Answer Span

	F1	Precision	Recall
(English)	53.9	64.8	46.0
Arabic	76.2	80.5	72.3
Bengali	56.7	60.3	53.7
Finnish	57.7	66.8	51.0
Hindi*	58.1 (50.9)	54.0 (37.4)	54.1 (39.5)
Indonesian	60.5	65.8	57.3
Japanese	39.1	52.9	32.4
Kiswahili	58.2	66.4	54.7
Korean	49.2	54.2	43.1
Persian*	50.5 (42.8)	48.3 (31.8)	43.4 (32.3)
Russian	54.7	62.2	50.8
Tamil*	52.3 (44.5)	53.5 (36.3)	44.1 (32.3)
Telugu	78.7	73.3	69.3
Thai	57.3	65.7	52.3
OVERALL	**57.7**	**61.9**	**53.4**

*Baselines in parenthesis, calculated separately using mBERT

1. **Augment TyDi QA Data:** We increase the number of languages the model(s) is (are) trained on by augmenting the original TyDi QA data with other datasets. These include chaii [cha] (2k Hindi and Tamil question-answer pairs, with a 700/100/200 train/test/val split each), PQuAD [Darvishi et al., 2023] (12k Persian question-answer pairs with a 8k/1k/3k train/test/val split). We also augment the English question-answer pairs using the SQuAD dataset [Rajpurkar et al., 2016, 2018] as we found that this leads to the model(s) more effectively encapsulating the question-answer paradigm.
2. **Fine-tune Ensemble of XLM-RoBERTa models:** We choose to fine-tune the XLM-RoBERTa model since it outperforms the baseline mBERT model on all benchmarks, additionally it is evaluated on the cross-lingual question answering task which is relevant to this project. We also found that fine-tuning XLM-RoBERTa separately on each of the language categories (as defined in the TyDi QA work) led to better performance on each language). Therefore, we fine-tune the XLM-Roberta model for each language category in the TydiQA baseline as described in Figure 5. Thus we train an ensemble of XLM-RoBERTa models; the fine-tuned model to be used is decided during inference time based on the input language.
3. **Evaluate on TyDi QA validation set along with other untranslated question-answer datasets:** We evaluate the models on a validation set withheld from the TyDi QA augmented dataset (from Step 1). We also use additional samples from datasets with question-answer pairs generated from untranslated text in the original language (as opposed to text translated from English) to benchmark the performance of our approach more accurately.

4.1 Training Details

We ran the training, fine-tuning and evaluation experiments on the Rutgers iLab GPU NVIDIA GeForce GTX 1080 Ti in a multi-GPU setting (since required vRAM exceeded single GPU capacity). 12.8 GB vRAM is required to train and fine-tune a single XLM-RoBERTa model with a batch size of 16 examples (consisting of context and answerable/unanswerable questions).

5 Results

Table 1 depicts the results obtained from our approach. We see that these significantly outperform the baseline. We have calculated the baselines for Hindi, Persian and Tamil languages using mBERT, similar to the TyDi QA process. The F1 scores for each language have been calculated using precision and recall. Similar to TyDi QA, each score is calculated as an average over four fine-tuned replicas of XLM-RoBERTa.

6 Future Work

Fine-tuning large language models can be seen as a viable future task for this project. Due to the demonstrably huge capacity and embedded knowledge of large language models such as GPT-3/GPT-4 and others, it may be possible to fine-tune such models for generalized multilingual question answering with only few-shot prompt engineering. This is something we plan to explore in the future.

References

Chaii - hindi and tamil question answering. URL `https://www.kaggle.com/competitions/chaii-hindi-and-tamil-question-answering`.

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

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale, 2020.

Kasra Darvishi, Newsha Shahbodaghkhan, Zahra Abbasiantaeb, and Saeedeh Momtazi. PQuAD: A persian question answering dataset. *Computer Speech &amp Language*, 80:101486, may 2023. doi: 10.1016/j.csl.2023.101486.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual bert?, 2019.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text, 2016.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad, 2018.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. No language left behind: Scaling human-centered machine translation. 2022.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding, 2019.

A Appendix

A.1 Jaccard Index

Jaccard Index is a statistic used to analyze the similarities between sample sets, often known as the Jaccard similarity coefficient. Formally, the measurement is the size of the intersection divided by the size of the union of the sample sets and emphasizes similarity between finite sample sets. It can be mathematically written as:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

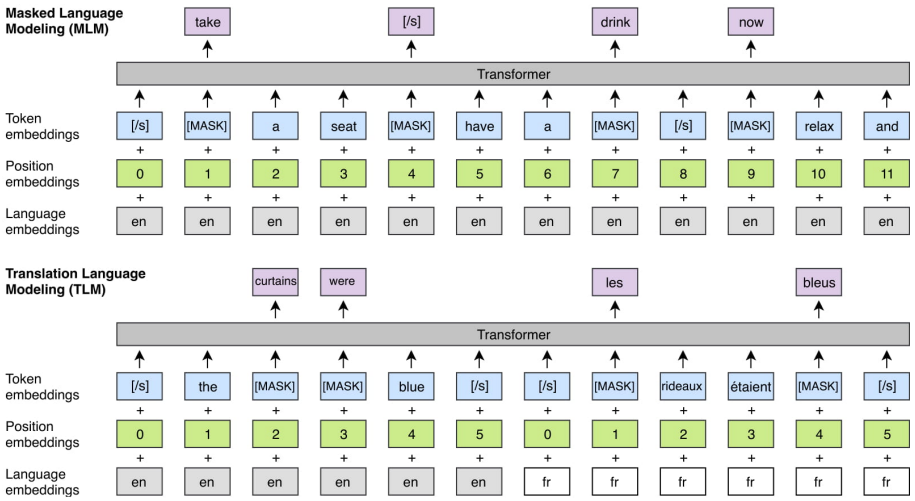

Figure 4: XLM-R uses representation learning to learn language agnostic representations which display cross-lingual effectiveness, especially for low-resource languages.

Figure 5: Typologically diverse languages.