A

Report

On

Cross-Lingual Question Answering with MLQA

Advanced NLP

By

Prudhvi Nadh Reddy K
2022201007

Rahul Mishra
2023201071

Md Jawed Equbal
2023201051

TEAM - NLP NEXUS

Advisors
Prof. Manish Shrivastava
Kapil

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

Project Report: Cross-Lingual Question Answering Using Multilingual Models and MLQA

---

1. Problem Statement and Motivation

The field of Natural Language Processing (NLP) has seen remarkable progress, especially in question answering (QA) systems. However, these advancements are predominantly focused on English, leaving other languages underrepresented. This project addresses the critical challenge of **cross-lingual QA**, aiming to improve **zero-shot transfer** between English and three target languages: Spanish, German, and Hindi.

The **MLQA (Multilingual Question Answering)** benchmark introduced by Lewis et al. (2020) highlights the significant performance gap in cross-lingual QA when compared to monolingual English performance. This gap underscores the pressing need for more effective cross-lingual QA systems that can generalize across languages without requiring large-scale training data for each target language. By addressing this problem, the project seeks to contribute toward developing more linguistically inclusive and accessible QA systems, ultimately helping bridge language barriers in global information retrieval.

---

2. Project Scope

The project focuses on **extractive question answering**, where the answer to a given question is a span within a provided context. The goal is to implement and compare various cross-lingual transfer approaches using the MLQA dataset as a benchmark.

Key objectives:

1. Evaluate **zero-shot** performance of multilingual models like mBERT, XLM-R, and XLM-R Large.
2. Conduct experiments in **translate-train** and **translate-test** settings to assess the impact of translation on cross-lingual QA.
3. Fine-tune models on relevant datasets and explore their performance on MLQA.
4. Establish performance baselines using **monolingual (BERT, RoBERTa)** and **multilingual models** to better understand their generalization capabilities.

This project explores these objectives using **three datasets**: MLQA, SQuAD v1.1, and XQuAD, leveraging various multilingual and monolingual pretrained models.

---

3. Datasets Overview

3.1 MLQA (Multilingual Question Answering)

- A benchmark for evaluating cross-lingual QA systems across seven languages: English, Arabic, German, Spanish, Hindi, Vietnamese, and Simplified Chinese.
- Each QA pair includes a context, question, and answer span in the same language.
- **Usage in Project**:
    - English portion used for training.
    - Spanish, German, and Hindi portions for zero-shot evaluation.
- **Dataset Size**: Over 12,000 QA pairs in English, with parallel versions in 2–4 additional languages, totaling 46,000 QA pairs.

3.2 SQuAD v1.1 (Stanford Question Answering Dataset)

- A large-scale English dataset for extractive QA tasks.
- **Usage in Project**:
    - Initial fine-tuning of monolingual and multilingual models.
- **Dataset Size**: Over 100,000 QA pairs covering 500+ articles.

3.3 XQuAD

- A multilingual QA dataset derived from SQuAD v1.1, containing 11 languages.
- **Usage in Project**:
    - Fine-tuning experiments.
    - Enables comparison with MLQA to analyze cross-dataset generalizability.
- **Dataset Size**: 1190 QA pairs per language.

---

4. Models

The project evaluates **seven models**, including four monolingual and three multilingual models.

Monolingual Models:

1. **BERT (Base and Large)**:
    - Pretrained on English data with 110M and 340M parameters, respectively.
    - Objective: Masked language modeling and next sentence prediction.
2. **RoBERTa (Base and Large)**:
    - Pretrained on English with dynamic masking and enhanced objectives.
    - Parameters: 125M (Base) and 355M (Large).

Multilingual Models:

1. **mBERT**:
    - Trained on Wikipedia in 102 languages with 110M parameters.

2. **XLM-R**:
    - Pretrained on 2.5TB of CommonCrawl data in 100 languages with 125M parameters.
3. **XLM-R Large**:
    - A larger version with 355M parameters and 24 transformer layers.

These models are fine-tuned using the datasets mentioned and evaluated in various experimental settings.

---

5. Experiments

We performed a total of six experiments to evaluate cross-lingual QA models and approaches on the MLQA dataset. These experiments include **zero-shot transfer**, **translate-train**, **translate-test**, **fine-tuning**, and a new experiment utilizing a **data-augmented pre-trained model**. Below is a brief overview of each experiment:

1. **Zero-shot Transfer (XLT)**: Multilingual models fine-tuned on SQuAD (English) were evaluated directly on MLQA test data in other languages to assess their zero-shot transfer capabilities. We also explored Generalized Zero-shot Transfer (G-XLT), where the question and context languages differ.
2. **Translate-test (Monolingual)**: Monolingual models fine-tuned on SQuAD were evaluated on MLQA test data translated into English. This approach enables monolingual models to participate in multilingual tasks.
3. **Translate-test (Multilingual)**: Multilingual models fine-tuned on SQuAD were evaluated on MLQA test data translated into English. Results were compared with those of monolingual models in the same setting.
4. **Translate-train**: Multilingual models were fine-tuned on SQuAD training data translated into target languages (e.g., Spanish, German, Hindi) and then evaluated on MLQA test data in those languages. This setting evaluates the impact of multilingual training on cross-lingual performance.
5. **Fine-tuning**:
    - Models were fine-tuned on the XQuAD dataset and evaluated on MLQA test data to assess the impact of supervised training on cross-lingual performance.
    - Fine-tuning was also performed on the MLQA validation set to evaluate improvements in task alignment.
6. **Data Augmentation**:
    - A pre-trained model (**bert-multi-cased-finetuned-xquadv1**) from Hugging Face, fine-tuned on an **augmented XQuAD dataset**, was evaluated on MLQA test data. The augmented dataset included additional samples generated via scraping and neural machine translation (NMT), making it more robust and diverse. This experiment demonstrates the potential of data augmentation for improving cross-lingual QA performance.

6. Baselines

The MLQA dataset provides several baseline results that serve as reference points for evaluating model performance. In this section, we analyze these results to enable comparison with the outcomes of our experiments.

6.1 MLQA XLT (Cross-lingual Transfer)

The baseline results for MLQA on the cross-lingual transfer (XLT) task are presented in **Table 1.** This task evaluates how well models trained on English data can generalize to other languages without additional training in the target languages.

Baseline Results (Table 1):

Model	en	es	de	ar	hi	vi	zh	avg
BERT-Large	80.2 / 67.4	-	-	-	-	-	-	80.2 / 67.4
mBERT	77.7 / 65.2	64.3 / 46.6	57.9 / 44.3	45.7 / 29.8	43.8 / 29.7	57.1 / 38.6	57.5 / 37.3	57.7 / 41.6
XLM	74.9 / 62.4	68.0 / 49.8	62.2 / 47.6	54.8 / 36.3	48.8 / 27.3	61.4 / 41.8	61.1 / 39.6	61.1 / 43.5

6.2 MLQA G-XLT (Generalized Cross-lingual Transfer)

The generalized cross-lingual transfer (G-XLT) task evaluates model performance when the question and context are in different languages. Results for this task are shown in **Tables 2 and 3.**

Baseline Results (Tables 2 and 3):

Table 2: **XLM Performance on G-XLT**

Context/Question	en	es	de	ar	hi	vi	zh
en	74.9	65.0	58.5	50.8	43.6	55.7	53.9
es	69.5	68.0	61.7	54.0	49.5	58.1	56.5
de	70.6	67.7	62.2	57.4	49.9	60.1	57.3
ar	60.0	57.8	54.9	54.8	42.4	50.5	43.5
hi	59.6	56.3	50.5	44.4	48.8	48.9	40.2
vi	60.2	59.6	53.2	48.7	40.5	61.4	48.5
zh	52.9	55.8	50.0	40.9	35.4	46.5	61.1

Table 3: **mBERT Performance on G-XLT**

Context/Question	en	es	de	ar	hi	vi	zh
en	77.7	64.4	62.7	45.7	40.1	52.2	54.2
es	67.4	64.3	58.5	44.1	38.1	48.2	51.1
de	62.8	57.4	57.9	38.8	35.5	44.7	46.3
ar	51.2	45.3	46.4	45.6	32.1	37.3	40.0
hi	51.8	43.2	46.2	36.9	43.8	38.4	40.5
vi	61.4	52.1	51.4	34.4	35.1	57.1	47.1
zh	58.0	49.1	49.6	40.5	36.0	44.6	57.5

6.3 Summary

The baseline results provided by the MLQA dataset establish key performance benchmarks:

1. **XLM achieves the best overall performance**, especially in XLT and G-XLT tasks.
2. Multilingual BERT shows decent performance but struggles more with non-English contexts and questions.
3. **Challenges in low-resource languages (e.g., Hindi and Arabic)** highlight the need for future improvements in cross-lingual QA.

These baselines serve as critical reference points for comparing the results of our experiments in subsequent sections.

7. Experimental Results

In this section, we present the results of our experiments. These results are compared across various language pairs and model configurations to evaluate their cross-lingual capabilities.

7.1 Zero-shot Transfer (XLT)

Zero-shot transfer evaluates models trained on English (SQuAD) without any fine-tuning on target languages. We report F1 and Exact Match (EM) scores for each language pair.

7.1.1 Zero-shot mBERT Results

Language Pair	F1	EM
de.de	59.39	43.75
de.en	62.36	46.73
de.es	56.37	40.99
de.hi	34.05	21.54
en.de	66.43	52.47
en.en	80.30	67.02
en.es	67.38	52.75
en.hi	39.27	26.27
es.de	60.59	40.20
es.en	66.93	46.39
es.es	64.88	43.61
es.hi	36.17	20.08
hi.de	47.55	33.85
hi.en	52.90	37.11
hi.es	43.73	29.08
hi.hi	46.21	30.01
Average	**55.28**	**39.49**

7.1.2 Zero-shot XLM-R Results

Language Pair	F1	EM
de.de	62.20	46.73
de.en	59.95	44.30
de.es	44.00	29.73
de.hi	29.41	17.62
en.de	60.75	47.13
en.en	80.78	68.02
en.es	57.84	43.92
en.hi	44.97	32.03
es.de	50.50	32.60
es.en	65.96	45.08
es.es	66.53	46.09
es.hi	31.81	17.06
hi.de	42.80	27.83
hi.en	60.59	43.41
hi.es	37.39	23.04
hi.hi	61.35	44.23
Average	**53.55**	**38.05**

7.1.3 Zero-shot XLM-R Large Results

Language Pair	F1	EM
de.de	68.47	52.38
de.en	67.73	51.69
de.es	65.32	49.55
de.hi	36.15	21.89
en.de	77.67	65.09
en.en	83.97	71.18
en.es	77.24	64.21
en.hi	43.64	30.68
es.de	69.96	48.65
es.en	72.09	50.31
es.es	72.11	50.16
es.hi	42.10	26.12
hi.de	65.07	49.93
hi.en	70.51	52.56
hi.es	63.25	45.91
hi.hi	69.83	51.30
Average	**65.32**	**48.85**

7.2 Translate-train

Translate-train experiments involve fine-tuning multilingual models on training data translated into the target languages.

7.2.1 Translate-train (German - XLM-R)

Language Pair	F1	EM
de.de	59.85	41.89
de.en	56.49	39.43
de.es	49.70	34.18
de.hi	39.00	25.45
en.de	66.34	53.62
en.en	76.42	62.29
en.es	60.83	45.80
en.hi	51.58	38.76
es.de	59.69	39.41
es.en	64.21	42.34
es.es	65.17	42.89
es.hi	39.91	23.22
hi.de	50.79	34.69
hi.en	56.47	38.25
hi.es	42.60	26.64
hi.hi	59.19	40.83
Average	**56.14**	**39.35**

7.2.2 Translate-train (Hindi - XLM-R)

Language Pair	F1	EM
de.de	57.22	41.02
de.en	54.03	38.03
de.es	41.02	28.32
de.hi	47.24	33.36
en.de	57.68	45.34
en.en	72.45	57.83
en.es	50.74	37.77
en.hi	57.63	44.18
es.de	48.60	30.74
es.en	58.68	37.77
es.es	61.74	40.53
es.hi	49.44	31.80
hi.de	52.76	36.57
hi.en	57.20	39.00
hi.es	46.52	30.64
hi.hi	64.00	45.04
Average	**54.81**	**38.62**

7.2.3 Translate-train (Spanish - XLM-R)

Language Pair	F1	EM
de.de	62.65	45.56
de.en	60.75	44.19
de.es	57.44	42.57
de.hi	40.79	27.13
en.de	65.94	52.09
en.en	79.05	64.10
en.es	70.96	56.92
en.hi	55.73	42.17
es.de	60.05	38.23
es.en	66.49	43.37
es.es	69.35	46.11
es.hi	48.25	28.67
hi.de	49.34	33.64
hi.en	60.08	41.99
hi.es	54.02	37.78
hi.hi	62.61	43.84
Average	**60.22**	**43.02**

7.3 Translate-test

Translate-test experiments involve evaluating monolingual and multilingual models trained on English (SQuAD) on MLQA test data translated into English. Both monolingual and multilingual models are tested to compare their capabilities in this setting.

7.3.1 Translate-test Monolingual Results

The performance of monolingual models, including BERT and RoBERTa (base and large versions), on translated test data is summarized below:

BERT Results

Language Pair	F1	EM
translate-test.de	54.39	35.73
translate-test.es	64.96	43.16
translate-test.hi	52.84	32.05
Average	**57.40**	**36.98**

BERT Large Results

Language Pair	F1	EM
translate-test.de	56.68	37.15
translate-test.es	67.15	45.19
translate-test.hi	55.20	33.81
Average	**59.68**	**38.72**

RoBERTa Results

Language Pair	F1	EM
translate-test.de	54.13	34.09
translate-test.es	66.02	43.44
translate-test.hi	52.33	30.99
Average	**57.49**	**36.17**

RoBERTa Large Results

Language Pair	F1	EM
translate-test.de	57.41	37.90
translate-test.es	67.92	45.92
translate-test.hi	55.58	33.92
Average	**60.30**	**39.24**

7.3.2 Translate-test Multilingual Results

The performance of multilingual models (mBERT, XLM-R, and XLM-R Large) on translated test data is summarized below:

mBERT Results

Language Pair	F1	EM
translate-test.de	53.61	34.82
translate-test.es	64.28	43.02
translate-test.hi	51.89	31.23
Average	**56.59**	**36.36**

XLM-R Results

Language Pair	F1	EM
translate-test.de	53.57	34.89
translate-test.es	64.77	43.02
translate-test.hi	52.76	32.05
Average	**57.03**	**36.65**

XLM-R Large Results

Language Pair	F1	EM
translate-test.de	56.56	37.44
translate-test.es	68.57	46.54
translate-test.hi	55.57	34.51
Average	**60.23**	**39.50**

7.4 Fine-tuning Experiments

In the fine-tuning experiments, multilingual models are trained on the **XQuAD** dataset and the **MLQA validation dataset** to evaluate their supervised learning capabilities. We report F1 and Exact Match (EM) scores for multiple languages to assess model performance.

7.4.1 Fine-tuning on XQuAD

The models were fine-tuned on the **XQuAD dataset** and evaluated across four languages: German (de), English (en), Spanish (es), and Hindi (hi). Results are summarized below:

mBERT Results

Language Pair	F1	EM
de.de	53.26	36.91
en.en	66.42	51.84
es.es	56.94	36.04
hi.hi	47.61	31.50
Average	**56.06**	**39.07**

XLM-R Results

Language Pair	F1	EM
de.de	51.73	36.04
en.en	66.80	53.01
es.es	57.77	37.03
hi.hi	52.86	36.11
Average	**57.29**	**40.55**

XLM-R Large Results

Language Pair	F1	EM
de.de	61.41	44.72
en.en	75.57	61.44
es.es	65.13	43.78
hi.hi	63.10	45.24
Average	**66.30**	**48.80**

7.4.2 Fine-tuning on MLQA Validation Data

The models were fine-tuned on the **MLQA validation set** and evaluated across the same four languages. Results are presented below:

mBERT Results

Language Pair	F1	EM
de.de	55.54	38.52
en.en	69.06	54.14
es.es	58.47	35.88
hi.hi	50.54	32.49
Average	**58.40**	**40.26**

XLM-R Results

Language Pair	F1	EM
de.de	54.79	38.50
en.en	70.38	56.12
es.es	60.21	38.38
hi.hi	57.01	39.65
Average	**60.60**	**43.16**

XLM-R Large Results

Language Pair	F1	EM
de.de	64.32	46.11
en.en	78.60	64.58
es.es	67.59	44.47
hi.hi	66.73	48.17
Average	**69.31**	**50.83**

7.5 Data Augmentation

In this experiment, we evaluate the pre-trained model **bert-multi-cased-finetuned-xquadv1**, available on Hugging Face, on the MLQA test dataset**.** This model was fine-tuned on an augmented version of the **XQuAD dataset** using data augmentation techniques to improve its robustness and multilingual generalizability**.**

Model Details

The **bert-multi-cased-finetuned-xquadv1** model was fine-tuned on an enhanced XQuAD dataset. Since XQuAD is primarily an evaluation dataset, the following **data augmentation techniques** were employed to create a larger, balanced training set:

1. **Data Scraping**: Additional samples were sourced to increase dataset size.
2. **Neural Machine Translation (NMT)**: Translation techniques were applied to existing samples, enabling multilingual fine-tuning.
3. **Balanced Test Set**: The test set was curated to ensure an equal number of samples for each language, ensuring fair evaluation.

Dataset Statistics

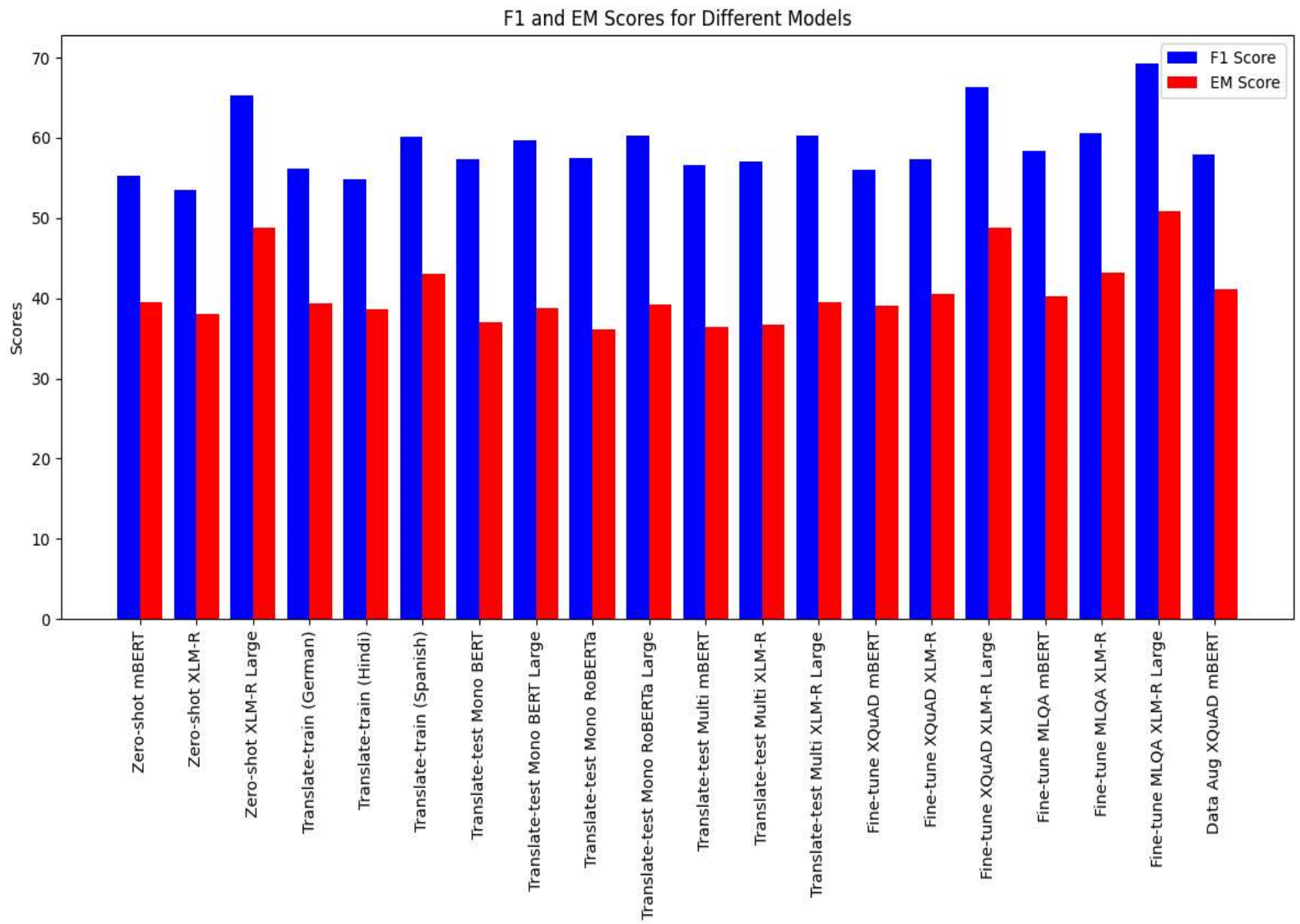
Dataset	Number of Samples
XQuAD Train	50,000
XQuAD Test	8,000

Evaluation on MLQA Test Data

The model was evaluated on the MLQA test dataset across four languages: German (de), English (en), Spanish (es), and Hindi (hi). The results are presented below:

Results

Language Pair	F1	EM
de.de	54.31	38.17
en.en	67.59	53.33
es.es	58.99	38.38
hi.hi	50.97	34.89
Average	**57.96**	**41.19**


F1 and EM Scores for Different Models
F1 Score
EM Score
Scores
0
10
20
30
40
50
60
70
Zero-shot mBERT
Zero-shot XLM-R
Zero-shot XLM-R Large
Translate-train (German)
Translate-train (Hindi)
Translate-train (Spanish)
Translate-test Mono BERT
Translate-test Mono BERT Large
Translate-test Mono RoBERTa
Translate-test Mono RoBERTa Large
Translate-test Multi mBERT
Translate-test Multi XLM-R
Translate-test Multi XLM-R Large
Fine-tune XQuAD mBERT
Fine-tune XQuAD XLM-R
Fine-tune XQuAD XLM-R Large
Fine-tune MLQA mBERT
Fine-tune MLQA XLM-R
Fine-tune MLQA XLM-R Large
Data Aug XQuAD mBERT

8. Final Summary of Experiments

In this project, six experiments were conducted to evaluate various cross-lingual QA models and approaches using the MLQA dataset. Below is a comprehensive summary of the experiments and their respective results:

8.1 Zero-shot Transfer (XLT)

In this experiment, multilingual models fine-tuned on SQuAD (English) were directly evaluated on the MLQA test dataset in other languages without any additional target-language training. We also tested the **Generalized Zero-shot Transfer (G-XLT)** setting, where the question and context languages differ.

- **mBERT** achieved an average F1 score of **55.28** and EM score of **39.49**. Performance was highest for English (80.3/67.02) and lowest for Hindi (46.21/30.01).
- **XLM-R** showed slightly lower performance with an average F1 of **53.55** and EM of **38.05**. It struggled with low-resource languages like Hindi (61.35/44.23 for Hindi-to-Hindi pairs).
- **XLM-R Large** outperformed the other models, achieving an average F1 of **65.32** and EM of **48.85**. It performed best on English (83.97/71.18), while Hindi showed significant improvement compared to smaller models.

Key Insight: Larger models like XLM-R Large exhibit better zero-shot transfer capabilities, but all models show a performance gap in low-resource languages like Hindi and morphologically complex languages like German.

8.2 Translate-train

In this setting, multilingual models were fine-tuned on SQuAD training data translated into the target languages (Spanish, German, and Hindi) and evaluated on MLQA test data.

- **XLM-R (German)** achieved an average F1 of **56.14** and EM of **39.35**, with the best performance on English (76.42/62.29) and the lowest on Hindi (59.19/40.83).
- **XLM-R (Hindi)** performed similarly, with an average F1 of **54.81** and EM of **38.62**, performing well on Hindi-to-Hindi pairs (64.0/45.04).
- **XLM-R (Spanish)** achieved the highest scores among translate-train experiments, with an average F1 of **60.22** and EM of **43.02**. It performed well on Spanish (69.35/46.11) and English (79.05/64.1).

Key Insight: Translate-train improves performance in specific languages, especially when training data is abundant and translations are of high quality. XLM-R's performance highlights its robustness in multilingual settings.

8.3 Translate-test (Monolingual)

Monolingual models fine-tuned on SQuAD (English) were evaluated on MLQA test data translated into English.

- **BERT** achieved an average F1 of **57.40** and EM of **36.98**, with its highest scores for Spanish-to-English (64.96/43.16).
- **BERT Large** showed improved performance with an average F1 of **59.68** and EM of **38.72**, highlighting the benefit of larger models.
- **RoBERTa** and **RoBERTa Large** achieved average F1 scores of **57.49** and **60.30**, respectively, with RoBERTa Large performing best on Spanish (67.92/45.92).

Key Insight: Monolingual models, particularly larger variants, perform well in translate-test settings but are limited in handling cross-lingual tasks directly.

8.4 Translate-test (Multilingual)

Multilingual models fine-tuned on SQuAD were evaluated on MLQA test data translated into English.

- **mBERT** achieved an average F1 of **56.59** and EM of **36.36**.
- **XLM-R** slightly outperformed mBERT with an average F1 of **57.03** and EM of **36.65**.
- **XLM-R Large** performed the best, with an average F1 of **60.23** and EM of **39.50**, excelling in Spanish-to-English translation (68.57/46.54).

Key Insight: Multilingual models consistently outperform monolingual models in translate-test settings, with XLM-R Large demonstrating the highest robustness.

8.5 Fine-tuning

This experiment involved fine-tuning multilingual models on **XQuAD** and the **MLQA validation dataset** and evaluating their performance on the MLQA test dataset.

Fine-tuning on XQuAD:

- **mBERT**: Average F1: **56.06**, EM: **39.07**.
- **XLM-R**: Average F1: **57.29**, EM: **40.55**.
- **XLM-R Large**: Average F1: **66.30**, EM: **48.80**, outperforming other models significantly, particularly in Hindi (63.10/45.24).

Fine-tuning on MLQA Validation Data:

- **mBERT**: Average F1: **58.40**, EM: **40.26.**
- **XLM-R**: Average F1: **60.60**, EM: **43.16.**
- **XLM-R Large**: Average F1: **69.31**, EM: **50.83**, achieving the best results across all languages, with the highest performance in English (78.60/64.58).

Key Insight: Fine-tuning on language-aligned datasets (e.g., MLQA) yields substantial performance improvements, with XLM-R Large achieving the best results.

8.6 Data Augmentation

In this experiment, we evaluated the pre-trained model **bert-multi-cased-finetuned-xquadv1**, which was fine-tuned on an augmented version of XQuAD using data augmentation techniques (e.g., scraping and neural machine translation). The model was directly tested on MLQA test data.

- **Results**:
  - German: **54.31/38.17**
  - English: **67.59/53.33**
  - Spanish: **58.99/38.38**
  - Hindi: **50.97/34.89**
  - **Average**: F1: **57.96**, EM: **41.19.**

Key Insight: Data augmentation techniques, combined with pre-trained multilingual models, significantly enhance robustness and generalizability for cross-lingual QA tasks.

Sample model outputs (Fine-tuning on MLQA XLR_Large)

'context': "कोरोनावायरस पश्चिम में आतंक बो रहा है क्योंकि यह इतनी तेजी से फैलता है।",
'question': "कोरोनावायरस घबराहट कहां है?"
Model output :{'score': 0.9748640656471252, 'start': 12, 'end': 18, 'answer': 'पश्चिम'}

'context': 'उसी "एरिया XX " नामकरण प्रणाली का प्रयोग नेवादा परीक्षण स्थल के अन्य भागों के लिए किया गया है।मूल रूप में 6 बटे 10 मील का यह आयताकार अड्डा अब तथाकथित \'ग्रूम बॉक्स " का एक भाग है, जो कि 23 बटे 25.3 मील का एक प्रतिबंधित हवाई क्षेत्र है। यह क्षेत्र NTS के आंतरिक सड़क प्रबंधन से जुड़ा है, जिसकी पक्की सड़कें दक्षिण में मरकरी की ओर और पश्चिम में युक्का फ्लैट की ओर जाती हैं। झील से उत्तर पूर्व की ओर बढ़ते हुए व्यापक और और सुव्यवस्थित ग्रूम झील की सड़कें एक दर्रे के जरिये पेचीदा पहाड़ियों से होकर गुजरती हैं। पहले सड़कें ग्रूम घाटी',
'question': 'Where does the Groom Lake Road head relative to the lake?'
Model output : {'score': 0.6938328742980957,'start': 469,'end': 485,'answer': 'पेचीदा पहाड़ियों'}

'context':'In 1994, five unnamed civilian contractors and the widows of contractors Walter Kasza and Robert Frost sued the USAF and the United States Environmental Protection Agency. Their suit, in which they were represented by George Washington University law professor Jonathan Turley, alleged they had been present when large quantities of unknown chemicals had been burned in open pits and trenches at Groom. Biopsies taken from the complainants were analyzed by Rutgers University biochemists, who found high levels of dioxin, dibenzofuran, and trichloroethylene in their body fat. The complainants alleged they had sustained skin, liver, and respiratory injuries due to their work at Groom, and that this had contributed to the deaths of Frost and Kasza. The suit sought compensation for the injuries they had sustained, claiming the USAF had illegally handled toxic materials, and that the EPA had failed in its duty to enforce the Resource Conservation and Recovery Act (which governs handling of dangerous materials). They also sought detailed information about the chemicals to which they were allegedly exposed, hoping this would facilitate the medical treatment of survivors. Congressman Lee H. Hamilton, former chairman of the House Intelligence Committee, told 60 Minutes reporter Lesley Stahl, "The Air Force is classifying all information about Area 51 in order to protect themselves from a lawsuit."',
'question': 'Who analyzed the biopsies?',
Model output:{'score': 0.3615085780620575,'start': 457,'end': 488, 'answer': 'Rutgers University biochemists,'}

'context': 'In 1994, five unnamed civilian contractors and the widows of contractors Walter Kasza and Robert Frost sued the USAF and the United States Environmental Protection Agency. Their suit, in which they were represented by George Washington University law professor Jonathan Turley, alleged they had been present when large quantities of unknown chemicals had been burned in open pits and trenches at Groom. Biopsies taken from the complainants were analyzed by Rutgers University biochemists, who found high levels of dioxin, dibenzofuran, and trichloroethylene in their body fat. The complainants alleged they had sustained skin, liver, and respiratory injuries due to their work at Groom, and that this had contributed to the deaths of Frost and Kasza. The suit sought compensation for the injuries they had sustained, claiming the USAF had illegally handled toxic materials, and that the EPA had failed in its duty to enforce the Resource Conservation and Recovery Act (which governs handling of dangerous materials). They also sought detailed information about the chemicals to which they were allegedly exposed, hoping this would facilitate the medical treatment of survivors. Congressman Lee H. Hamilton, former chairman of the House Intelligence Committee, told 60 Minutes reporter Lesley Stahl, "The Air Force is classifying all information about Area 51 in order to protect themselves from a lawsuit.'",
'question': 'बायोप्सी का विश्लेषण किसने किया?',
{'score': 0.7921487092971802,'start': 457,'end': 488,'answer': 'Rutgers University biochemists,'}

9. Conclusion

This project comprehensively evaluated various cross-lingual question answering (QA) models and strategies on the MLQA dataset, leveraging multilingual and monolingual pretrained models. The following key conclusions were drawn from the experiments:

1. **Performance of Multilingual Models**:
    - **XLM-R Large** consistently outperformed other models across all experimental settings, showcasing its superior ability to handle multilingual QA tasks effectively. Its performance highlights the importance of larger model architectures in cross-lingual scenarios.
2. **Cross-lingual Challenges**:
    - Languages with fewer resources and higher morphological complexity, such as Hindi, consistently demonstrated lower performance. This highlights the need for better pretraining strategies and dataset alignment to address these challenges.
3. **Effectiveness of Data Augmentation**:
    - The data augmentation experiment demonstrated that techniques like data scraping and neural machine translation significantly enhance model robustness and cross-lingual generalizability, even for pre-existing datasets like XQuAD.
4. **Impact of Fine-tuning**:
    - Fine-tuning models on language-aligned datasets, such as the MLQA validation set, resulted in significant performance improvements across all languages,

particularly for XLM-R Large. This underscores the value of task-specific and multilingual fine-tuning for improving QA systems.

5. **Approaches Comparison**:
    - Among the approaches evaluated, **fine-tuning** yielded the best overall improvements, followed by **translate-train**, which performed well when high-quality translated data was available. **Zero-shot transfer** showed promise for cross-lingual generalization, but performance gaps remain, especially for low-resource languages.

This project provides a strong foundation for improving cross-lingual QA systems, demonstrating the efficacy of data augmentation, robust multilingual models, and targeted fine-tuning strategies. Future work could explore integrating advanced pretraining techniques, expanding datasets for low-resource languages, and experimenting with few-shot or zero-shot learning to further enhance performance and linguistic inclusivity in QA systems

10: References:

1.https://huggingface.co/mrm8488/bert-multi-cased-finetuned-xquadv1

2.https://huggingface.co/alon-albalak/xlm-roberta-large-xquad

3.https://huggingface.co/alon-albalak/xlm-roberta-base-xquad

4.https://huggingface.co/alon-albalak/bert-base-multilingual-xquad

5. MLQA: Evaluating Cross-lingual Extractive Question Answering (Lewis et al., 2020)

- Introduces the MLQA benchmark dataset.
- Establishes baselines for cross-lingual transfer in QA.
- Provides a framework for evaluating cross-lingual QA performance.

6. XNLI: Evaluating Cross-lingual Sentence Representations (Conneau et al., 2018)

- Introduces the XNLI dataset for cross-lingual natural language inference.
- Explores techniques for cross-lingual sentence embedding.
- Relevant for our work on improving cross-lingual representations.

7. Cross-lingual Language Model Pretraining (Lample and Conneau, 2019)
- Introduces XLM, a powerful cross-lingual language model.
- Based on the Transformer architecture.
- Informs our approach to fine-tuning cross-lingual language models.

8. Zero-shot Reading Comprehension by Cross-lingual Transfer Learning with Multi-lingual Language Representation Model (Hsu et al., 2019)
- Explores zero-shot cross-lingual transfer for reading comprehension.
- Uses multilingual language models.
- Directly relevant to our zero-shot transfer goals.

9. XLM-RoBERTa: Unsupervised Cross-lingual Representation Learning at Scale (Conneau et al., 2020)
- Introduces XLM-RoBERTa, an advanced multilingual language model.
- Demonstrates state-of-the-art performance on cross-lingual benchmarks.
- Will be a key component in our fine-tuning experiments.