CROSS-LINGUAL CAPABILITIES IN Q/A MODELS

Team Linguists

Kukkapalli Shravya - 2021101051
Penumalla Aditya Pavani - 2021101133
Siya Puttagunta - 2021101062

OVERVIEW

- Introduction
- Datasets
- Experiments
- Metrics
- Results
- Analysis
- Conclusion

INTRODUCTION

- Our project focuses on evaluating the cross lingual transfer capability of QA systems across multiple languages.

To evaluate cross-lingual QA capabilities, we focus on three key tasks:

- **Cross-lingual Transfer (XLT)**
- **Generalized Cross-lingual Transfer (G-XLT)**
- **Asymmetric Cross-lingual Transfer(A-XLT)**

- Our focus is on English (en), German (de) and Spanish (es) languages.

XLT

- **Cross-lingual Transfer (XLT):** We train the model in one language (i.e., English) and test it in another (e.g., Spanish).

- **NOTE:** An x-y pair refers to a context-question pair, where x represents the context and y represents the question.

G-XLT

- **Generalized Cross-lingual Transfer (G-XLT):** We train on QA data where the context and question are in the same language and test in scenarios where the context and question are in different languages.

Training pairs

en - en

Testing pairs

en - de	es - en
de - en	de - es
en - es	es - de

A-XLT

- **Asymmetric Cross-lingual Transfer(A-XLT):** We train the model on multilingual data and test it in multilingual settings.

Training pairs

en - en
de - de
es - es

Testing pairs

en - de	es - en
de - en	de - es
en - es	es - de

DATASETS

SQuAD 1.1

- SQuAD 1.1 is a benchmark dataset for training and evaluating QA models.
- It contains over 100,000 question-answer pairs derived from Wikipedia articles.
- The dataset includes 87,599 training pairs and 10,570 validation pairs.
- Each question is paired with a passage, and the task can be to identify the correct answer either as a span of text or the sequence of answer text within the passage.
- Widely used due to its high-quality annotations and diverse topics, making it a standard for assessing reading comprehension in models.

DATASETS

PaXQA

- We explored the PaXQA dataset, designed for cross-lingual QA tasks, with question-answer pairs in multiple languages like English, Chinese, Russian and Arabic.
- We fine-tuned Bert model using 50k English-English, 28k Chinese-Chinese, and 45k Arabic-Arabic pairs, achieving low scores (Exact Match: 1.60, F1: 6.82) on English-English tasks.
- Upon reviewing sample entries, we found significant ambiguities: conflicting facts, incorrect answers, and annotation errors that hindered the model's accuracy.
- For example, the context provided conflicting numbers, vague answers, or irrelevant text, making it difficult for the model to generate reliable outputs.
- Due to these inconsistencies, we concluded that the dataset’s quality is insufficient for effective training and chose not to use it for our work.

DATASETS

PaXQA Ambiguity

QA Dataset Example 1

Context:
The convict ”pulled the ladder on which the victim was standing, which brought him down. He then beat him with an iron rod on his head several times until he died. He then wrapped his body with a sheet, put it in the car, and threw it away in one of the mountains to hide his crime, following a disagreement between the two,” according to the ministry. The sentence was implemented in Aseer (south). **At least 153 people were executed in 2007** and at least 37 in 2006, and **at least 83 were executed in 2007** and 35 in 2004, while 113 people were executed in 2000. Saudi Arabia penalizes with execution the crimes of rape, murder, conversion from Islam, armed assault, and drug smuggling.
Question: How many executions were there in 2007?
Answer: { Answer start: [481], Text: [83] }

In this example regarding executions in Saudi Arabia, the context presents conflicting figures for the number of executions in 2007, citing both 153 and 83. While the answer ”83” is included, the presence of two different numbers creates confusion regarding which figure is accurate.

DATASETS

PaXQA Ambiguity

QA Dataset Example 2

Context:
Wayne Whittaker, Benś father, told the BBC that the physicians said that Ben would be in severe pain if he were a young teenager and had suffered this injury. ”Things Get More Difficult.” After PRO conducting a series of tests on Ben, it appeared that he was afflicted * with a congenital defect that * prevented him from PRO feeling pain, which * is an extremely rare condition. The experts hypothesize that the condition is attributable to a deficiency in a substance in the body called * PRO beta-endorphin, which * is a substance that * is spontaneously secreted * in the body PRO to make humans feel pain.

Question: What is Ben’s congenital defect?
Answer: { Answer start: [160], Text: [Things Get More Difficult] }

In this instance involving Ben’s congenital defect, the provided answer, ”Things Get More Difficult,” does not correctly represent the condition itself. The accurate answer should pertain to his inability to feel pain, highlighting a significant gap in the dataset’s quality.

DATASETS

PaXQA Ambiguity

QA Dataset Example 3

{'context': 'و في هذا الاطار, قال مسؤولون و ديبلوماسيون في الدوحة ان قطر تعارض * اي هجوم على العراق, و لكن ها تخشى * ان يضر رفض ها طلباً اميركياً محتملاً ل استخدام * اراضي ها * قاعدة ل هجوم ب علاقات ها مع واشنطن. و أكد مسؤول قطري ب ان الدوحة لم تتلق * بعد اي طلب من الولايات المتحدة ل استخدام اراضي ها, "و لكن اذا ما حدث * ذلك, ف سوف تكون لحظة حرجة جدا ل لقيادة ان تقبل * الطلب او ترفض * ه', 'question': 'ماذا يقول مسؤول قطري لو طلبت الولايات المتحدة استخدام اراضي ها؟', 'answers': {'answer_start': [None], 'text': ['سوف تكون لحظة حرجة جدا ل لقيادة ان تقبل الطلب او ترفض ه']}}

This example illustrates an annotation error, where the answer start is marked as **None** but text is not None. This indicates a failure in the dataset's annotation process, leading to ambiguity about where the answer can be found within the context.

DATASETS

PaXQA Ambiguity

QA Dataset Example 4

Context:
Shanghai, which * has entered into the post-industrialization mid-development stage, has been providing wide investment scope for foreign businesses from various countries. Shanghai is the largest modernized industrial city in China and has complete industry categories, developed infrastructure and perfect municipal functionality. The financial center, business center and shipping center of the inland of China are also here. The Pudong development and the Yangtze River Valley development strategy spearheaded * by Shanghai implemented * by the central government have redoubled the strength of this city. The good investment environment makes the success rate of investment by foreign businesses in Shanghai reach 80

Question: What is the name of China’s business center?
Answer: { Answer start: [337], Text: [financial center] }

In this example about Shanghai, the question asks for the name of China’s business center, but the answer provided is just ”financial center,” which doesn’t fully answer the question.

DATASETS

MT-SQuAD

- We utilized MT-SQuAD, a machine-translated version of the SQuAD training set.
- The translations were generated using Facebook’s production models to create multilingual QA pairs.
- This dataset enables multi-lingual training by providing question-answer pairs in multiple target languages.

```
Size of train_data_de:  80069
Size of val_data_de:  9927
Size of train_data_es:  81810
Size of val_data_es:  10123
Size of train_data_en:  87599
Size of val_data_en:  10570
```

DATASETS

MLQA

- We used MLQA, a multi-way aligned benchmark for testing extractive QA models.
- The dataset includes QA instances in 7 languages: English, Arabic, German, Spanish, Hindi, Vietnamese, and Simplified Chinese.
- Each instance is parallel across an average of 4 languages, enabling cross-lingual evaluation.
- MLQA contains over 12K English instances and 5K instances per other language, totaling over 46,000 QA annotations.

Split	en	de	es	ar	zh	vi	hi
dev	1148	512	500	517	504	511	507
test	11590	4517	5254	5335	5137	5495	4918

EXPERIMENTS

Approaches

- We evaluated the performance of QA systems in cross-lingual settings using three model architectures: encoder-only, encoder-decoder, and decoder-only models. All models were fine-tuned and tested under consistent setups to compare their capabilities across the MLQA tasks.
- **Encoder-Only Models:** Multilingual BERT and XLMRoberta were fine-tuned to predict the start and end token positions of answers within the context.
- **Encoder-Decoder Models:** Multilingual T5 and Multilingual Bart were trained to generate answers as text sequences conditioned on the input question and context.
- **Decoder-Only Models:** GPT-2 was fine-tuned as a generative model, focusing on completing the input with the correct answer.

EXPERIMENTS

Training

- We fine-tuned the models using two different datasets to evaluate their performance in cross-lingual QA tasks:

Task 1: Training on SQuAD

- Models were trained on English-only context-question-answer triples from the SQuAD dataset. This setup tests the models' ability to generalize to cross-lingual settings despite being trained monolingual.

Task 2: Training on MT-SQuAD

- Models were fine-tuned on the MT-SQuAD dataset, which combines context-question-answer triples in English, Spanish, and German. This approach evaluates whether multilingual training improves performance compared to English-only training.

EXPERIMENTS

Testing

- We evaluated the trained models on the MLQA dataset using two strategies to assess their performance:

Same-Language Testing (XLT):

- Models were tested on monolingual pairs (en-en, de-de, es-es) to evaluate their ability to perform when context, question, and answer are all in the same language.

Cross-Language Testing (G-XLT):

- Models were tested on cross-lingual pairs (e.g., en-de, en-es, de-en) where the context and question are in different languages. This setup assesses the models' capability to handle cross-lingual understanding and provide answers in the context language.

METRICS

Encoder-only Models

- We evaluated the performance of encoder-only models using the following metrics, suitable for span-based predictions:

  Exact Match (EM):
- Measures the percentage of predictions that exactly match the ground truth answers, ensuring both content and format are identical.

  F1 Score:
- Combines precision and recall to reward partial matches, balancing the trade-off between precision and recall for a more nuanced evaluation of accuracy.

METRICS

Decoder-only and Encoder-Decoder Models

- We evaluated the performance of decoder-only and encoder-decoder models using the following metrics, suitable for generative predictions:

ROUGE-L :

- Measures the overlap of longest common subsequence between the generated and reference texts, focusing on recall to evaluate content similarity in generative tasks.

BERTScore :

- Utilizes pre-trained language models to compute semantic similarity between token embeddings of generated and reference texts

RESULTS

Encoder-only (EM/F1)

Language → Model ↓	en-en	en-de	en-es	de-en	es-en	de-de	de-es	es-de	es-es
mBERT V1	**63.71** **76.95**	53.71 66.14	51.44 64.76	**45.85** **59.85**	**48.18** **65.50**	**45.89** **59.96**	**40.32** 53.87	41.84 59.51	**46.64** **64.32**
mBERT V2	61.53 75.23	**55.01** **67.26**	**52.66** **67.00**	38.88 55.72	42.98 62.49	39.43 56.48	37.56 **54.13**	**42.00** **60.32**	44.13 64.00
XLMRoberta V1	61.88 74.70	37.55 48.18	31.43 42.63	**42.40** **55.56**	**45.36** 61.76	**42.24** **56.39**	19.93 29.82	26.07 39.56	43.58 60.41
XLMRoberta V2	**61.98** **74.51**	**48.64** **59.63**	**49.84** **62.44**	40.12 55.18	44.93 **62.60**	41.33 56.87	**36.2** **50.33**	**39.02** **56.42**	**46.13** **64.07**

- v1 - Models trained only on English
- v2- Models trained on combined data (English, German,Spanish)

ANALYSIS

Encoder-only (EM/F1)

Pair Type	v1 Performance (Limited Transfer)	v2 Performance (Improved Transfer)
x-x (e.g., de-de)	Limited transfer for non-English x-x pairs (e.g., de-de, es-es). Performance is better on en-en due to training data.	Slight improvement in de-de and es-es, with better multilingual adaptation.
x-y (en-de)	Moderate transfer for en-x or x-en pairs, with English acting as a pivot.	Noticeable improvement for en-x, and non-English pairs like de-es and es-de.
x-y (de-es)	Very weak transfer for non-English x-y pairs (e.g., de-es, es-de).	Significant improvement for non-English x-y pairs.

RESULTS

Encoder-Decoder (Rouge/Bert)

Language → Model ↓	en-en	en-de	en-es	de-en	es-en	de-de	de-es	es-de	es-es
BART v1	0.59 0.87	0.31 0.78	0.24 0.74	0.29 0.78	0.32 0.78	0.28 0.76	0.17 0.68	0.23 0.74	0.25 0.71
BART v2	**0.61** **0.88**	**0.51** **0.85**	**0.48** **0.84**	**0.44** **0.82**	**0.53** **0.85**	**0.48** **0.83**	**0.41** **0.81**	**0.50** **0.84**	**0.53** **0.84**
T5 v1	0.61 0.88	0.53 0.85	0.54 0.85	0.45 0.82	0.49 0.84	0.46 0.82	0.42 0.81	0.43 0.81	0.51 0.84
T5 v2	**0.71** **0.91**	**0.69** **0.90**	**0.67** **0.89**	**0.59** **0.86**	**0.65** **0.88**	**0.60** **0.86**	**0.59** **0.86**	**0.64** **0.88**	**0.65** **0.88**

ANALYSIS

Encoder-Decoder (Rouge/Bert)

Pair Type	BART-v1 (Limited Transfer)	BART-v2 (Improved Transfer)	T5-v1 (Strong Transfer)	T5-v2 (Superior Transfer)
x-x (e.g., de-de)	Weak for non-English pairs (de-de: 0.28).	Moderate improvement (de-de: 0.48).	Strong x-x performance (de-de: 0.46).	Exceptional x-x performance (de-de: 0.60).
x-y (e.g., en-de)	Weak cross-lingual performance (en-de: 0.31, de-es: 0.17).	Moderate improvement (en-de: 0.51, de-es: 0.41).	Strong transfer in English-centric and non-English pairs.	Best cross-lingual performance (en-de: 0.69, de-es: 0.59).

RESULTS

Decoder only (Rouge/Bert)

Language → Model ↓	en-en	en-de	en-es	de-en	de-de	de-es	es-en	es-de	es-es
GPT2 v1	**0.12** 0.62	0.07 0.60	0.07 0.60	0.09 0.56	0.09 0.56	0.08 0.56	0.10 0.59	0.10 0.62	0.11 0.62
GPT2 v2	0.11 **0.63**	**0.10** **0.63**	**0.10** **0.63**	**0.12** **0.60**	**0.12** **0.60**	**0.11** **0.60**	**0.15** **0.65**	**0.14** **0.65**	**0.16** **0.66**

ANALYSIS

Decoder only (Rouge/Bert)

Pair Type	GPT2-v1 Performance (Limited Transfer)	GPT2-v2 Performance (Improved Transfer)
x-x (e.g., de-de)	Weak for non-English pairs. Scores like de-de: 0.09, es-es: 0.11 highlight limited transfer from English-only training.	Significant improvement for non-English x-x pairs (de-de: 0.12, es-es: 0.16), reflecting better multilingual capacity.
x-y (e.g., en-de)	Poor transfer for English-to-non-English pairs (e.g., en-de: 0.07, en-es: 0.07). Non-English pairs like de-es also perform poorly.	Improved transfer across all x-y pairs, particularly for non-English directions (e.g., de-es: 0.11, es-de: 0.14).

CONCLUSION

- In the context of cross-lingual transfer, encoder-decoder architectures excel, offering superior transfer across language pairs and strong multilingual alignment, especially in advanced models like T5-v2, making them ideal for generation tasks requiring cross-lingual reasoning.
- Encoder-only models perform efficiently on retrieval and monolingual QA but show moderate capabilities for cross-lingual tasks, relying heavily on multilingual fine-tuning for alignment.
- Decoder-only models, while excelling in single-language generation, struggle with cross-lingual tasks and exhibit limited multilingual alignment, even with specialized training.

THANK YOU