Multilingual Question Answering Final Task Report

Krinal Navinchandra Patel Sarvesh Gupta Gautam Sharma

21CS60R39 21CS60R53 21CS60R26

Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Sub Task-03

Giving an answer from a passage based on the question.

- Training the Tydi-QA gold passage train dataset and evaluation on validation dataset only for Telugu and bengali.
- Training the base model on English, Telugu, and Bengali of Tydi-QA gold passage and validating on dev dataset only for telugu and bengali.

Group Details

Group Name: $Trio$
Members:

- Krinal Patel (21CS60R39, M.tech CSE 1^{st} year)
- Sarvesh Gupta (21CS60R53, M.tech CSE 1^{st} year)
- Gautam sharma (21CS60R26, M.tech CSE 1^{st} year)

Individual contributions:

In this project, we all have participated in the discussion, coding part and report writing part.
However, we can broadly classify our individual contribution as below.

- *Gautam Sharma* Generated the augmented dataset with the SQuAD (translated question only) dataset and incorporated with the model
- *Sarvesh Gupta* Performed the data augmentation technique with SQuAD context translation and intgrate it with the existing SQuAD dataset.
- *Krinal Patel* Implemented the augmentation techniques with tydiQA dataset and incorporated it with the mBERT model

Metric Used

For our experiment, we are using two standard metrics

- *Exact Match(EM)*
  With this metric, we compare the golden answer with the predicted answer after normalization.
- $f_1 = 2 * \frac{Precision * recall}{Precision + Recall}$

Task Description

Implementing the base model using pretrained mBERT and XLM transformer model on Tydi-QA gold passage dataset (containing 9 languages).
Extracting the English, Telugu and Bengali language data from the Tydi-QA gold passage, training on base model and validating on dev dataset only for Telugu and Bengali datasets.

Phase 1

Tydi-QA dataset contains Question-Answer data encompassing 9 different languages with 49881 training question-answers and 5077 validation question-answers.
Number of QA examples in different languages is shown below:

Language	Training example	Validation Examples
english	3696	440
telugu	5563	669
bengali	2390	113
arabic	14805	921
finnish	6855	782
indonesian	5702	565
korean	1625	276
russian	6490	812
kiswahili	2755	499

Total number of training examples is 49881, and number of validation examples is 5077

Base Model Details: We are using BERT multilingual base model (cased) (mBERT) and base-sized XLM-RoBERTa model to train the Tydi-QA gold passage dataset. Each training examples contains 'Id', 'Context', 'Question', 'Answer(s)(starting point and answer text)'

Using PreTrainedTokenizerFast to tokenize the dataset. Hyper-parameter Details:

- Restricting the maximum length of context and question to 384
- Document stride as 160 in case of split is performed on over-sized context field.
- Batch size is 16, Number of epochs is 1, learning rate(λ) is $2 * 10^{-5}$ and weight decay parameter is 0.01

$\longrightarrow$ Both mBert and XML Models are being trained on QA of all 9 languages and validated on dev dataset consisting only 'Telugu' and 'Bengali' datasets.

$\longrightarrow$ Training base models on only ,'English', 'Telugu', and 'Bengali' QA of gold passage and evaluating on validation dataset of 'Telugu' and 'Bengali'.

Baseline results:

Training model on all languages and validating 'Telugu' and 'Bengali'

`On mBERT model`

Language	Exact Match	F1-score
Telugu	69.058	82.582
Bengali	61.946	72.063

`On XLM model`

Language	Exact Match	F1-score
Telugu	65.769	80.491
Bengali	62.831	74.207

Training model on 'English', 'Telugu', and 'Bengali' languages and validating on 'Telugu' and 'Bengali'

`On mBERT model`

Language	Exact Match	F1-score
Telugu	66.816	80.780
Bengali	60.176	72.688

`On XLM model`

Language	Exact Match	F1-score
Telugu	62.929	77.637
Bengali	49.557	66.137

Zero shot Model

Training mBert transformer model on 'Squad' dataset and validating on 'Bengali' & 'Telugu' Tydi-QA gold passage dataset.

Language	Exact Match	F1-score
Bengali	46.017	56.297
Telugu	40.657	50.195

Phase 2

From our experiment, It is apparent that the model's performance can be improved by adding more telugu and bengali QA pairs. However, there was scarity of properly annotated bengali and telugu dataset.

But from our previous zero-shot experiment, we can find out that model's performance can be improved by training the model on the TyDiQA dataset appended with the english

SQuAD dataset. And we got the result shown below:

Language	Exact Match	F1-score
Telugu	71.748	83.984
Bengali	65.486	77.528
Overall	70.844	83.052

Phase 3

Although the model has shown positive feedback, it is still not good and can be improved. But due to sacrity of correctly annotated dataset, we can't improve on that. so we have tried to experiment with the dataset using the data augmentation methods and train our mBERT model on the augmented dataset obtained.

By our previous experiment on mBERT and XLM-roberta model on filtered and non-filtered datasets, we observed that the model was performing well with non-filtered datasets and mBERT model and so we try to continue our experiment with mBERT model without filtering the datasets.

Translated questions of TydiQA dataset

We have used the google translation APIs provided to translate the TydiQA dataset. But due to the character limitation on the translation by the API, we have been able to translate $\sim$ 20,000 Questions into bengali and telugu of the tydiQA dataset. And then we've concatenated the augmented dataset with the TydiQA itself and train the mBERT model on it.

Language	Exact Match	F1-score
Telugu	69.10	82.77
Bengali	64.60	74.32
Overall	68.41	81.52

Translated questions of SQuAD dataset

We have performed the same experiment with the SQuAD dataset $i.e.$, translated questions of SQuAD dataset into bengali and telugu, concatenated it with TydiQA and English SQuAD dataset and train the model.

Language	Exact Match	F1-score
Telugu	70.70	83.81
Bengali	69.03	79.34
Overall	70.46	83.15

Translated questions, context and answers of SQuAD dataset

In this technique, we've translated all the fields $i.e.$, questions, context and answer into bengali and telugu of SQuAD dataset concatenated it with the TydiQA and english SQuAD dataset and train the model on it.

Language	Exact Match	F1-score
Telugu	70.25	83.55
Bengali	60.18	73.6
Overall	68.79	82.11

Combined 'Translated all fields' and 'Translated only questions'

To keep our experiment diverse, we tried to build a dataset containing both forms of translation. so, we concatenated both SQuAD with only question translated, SQuAD with all field translated, English SQuAD and tydiQA dataset and train our mBERT model on it.

Language	Exact Match	F1-score
Telugu	70.85	83.18
Bengali	64.60	78.31
Overall	69.95	82.47

Translated using another Translation model

To overcome the character translation limitation by google translation API, we try to use a different translation model 'mBART-50 one to many multilingual machine translation'(5) to translate questions of SQuAD dataset to bengali and telugu. But with the freely available GPU resources, we're able to translate 8193 English questions. Along with this translated dataset, we concatenated english SQuAD dataset and TydiQA dataset and train our mBERT model on it.

Language	Exact Match	F1-score
Telugu	71.30	84.07
Bengali	71.68	79.76
Overall	71.35	83.45

Results and Discussion

We've tried various data augmentation techniques to generate multiple synthesised datasets upon which mBERT model is trained and tested on their respective validation splits. Basically, adding more data has helped us to improve the performance of the model. However, model has performed better when we have appended only question translated dataset than the entire translated dataset.

Dataset	Bengali		Telugu		Overall	
	Exact Match	F1 Score	Exact Match	F1 Score	Exact Match	F1 Score
Phase - 1						
Tydiqa – [eng, ben, tel] (mbert Model)	60.18	72.69	66.82	80.78	65.86	79.61
Tydiqa – [eng, ben, tel] (xlm Model)	49.56	66.14	62.93	77.64	60.99	75.96
Tydiqa – [all languages] (mbert Model)	61.95	72.06	69.06	82.58	68.03	81.06
Tydiqa – [all languages] (xlm Model)	62.83	74.21	65.77	80.49	65.35	79.58
Squad –zero shot approach (mbert Model)	46.02	56.29	40.66	50.19	41.43	51.08
Phase - 2						
Tydiqa + squad	65.49	77.52	71.75	83.98	70.84	83.05
Phase - 3						
Tydiqa + Bengali & Telugu translation of Tydiqa (20k pairs)	64.60	74.32	69.10	82.77	68.41	81.52
Tydiqa + squad + squad only question translation	69.03	79.34	70.70	83.81	70.46	83.15
Tydiqa + squad + squad context translation	60.18	73.6	70.25	83.55	68.79	82.11
Tydiqa + squad + squad question translation + squad context translation	64.60	78.31	70.85	83.18	69.95	82.47
Tydiqa + squad + squad question translation using mbart (approx. 8k pairs)	71.68	79.76	71.30	84.07	71.35	83.45

Figure 1: Combined Analysis

We got the the best performance of 71.68% EM on bengali (which is approx 11% improvement over base model trained in phase 1) and 71.30% EM on Telugu (which is approx $2\% \sim 3\%$ improvement over base model trained in phase 1). We have also observed that we are not able to improve performance of Telugu much. But, we have found an interesting pattern in wrong predictions as shown below. There is only one character difference between golden answer and predicted answer. Additionally, the majority of the wrong predictions(125 out of 192) are facing this type of issue. So, basically meaning of the predicted answer and original answer is same though it is not being captured in exact match.

Wrong Predictions (Telugu):

{'id': 'telugu-2319056933804294016-13', 'prediction': 'మైఖేల్ క్రింక్టన్', 'original:': 'స్టీవెన్ స్పీల్బర్గ్'}
{'id': 'telugu--7211034341454150780-1', 'prediction': 'విశాఖపట్నం', 'original:': 'విశాఖపట్నంలో'}
{'id': 'telugu-4527046517267694248-0', 'prediction': '1179 హెక్టార్లలో', 'original:': '1179 హెక్టార్ల'}

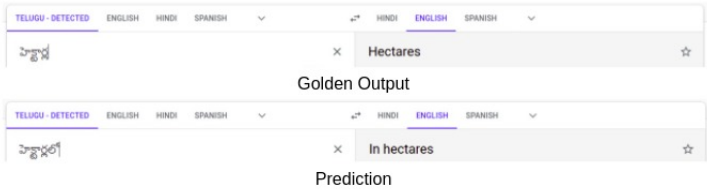

Figure 2: Telugu Dataset analysis

```
c=0
for i in wrong_prediction['telugu']:
    if i['original:']+'లో'==i['prediction']:
        c+=1
print('total wrong prediction:',len(wrong_prediction['telugu']))
print('Minor mismatches:',c)
```

```
total wrong prediction: 192
Minor mismatches: 125
```

Figure 3: Exact Match Issue

Difficulty Faced/Path not taken One of the major challenge we faced was to generate different kinds of good quality translated datasets due to the scarity of resources. The models available have character limitation, take a lot of time for translation, and doesn't provide ideal translated text. Another Issue with the translation is the context alignment problem, when the context and answers are translated, the start and end-token of the answer also needs to be adjusted accordingly for the translated context.

We hope to translate different combinations of Context, Question, and answer to continue our experiment on different models. And to overcome the context- alignment problem, we can try the technique to wrap the answers in the context with a unique tag as specified by Mihaela Bornea(4)

References

[1] Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, Holger Schwenk (2019) *MLQA: Evaluating Cross-lingual Extractive Question Answering.*,

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova (2018) *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*

[3] Guillaume Lample, Alexis Conneau (2019) *Cross-lingual Language Model Pretraining*

[4] Multilingual Transfer Learning for QA Using Translation as Data Augmentation *Mihaela Bornea, Lin Pan, Sara Rosenthal, Radu Florian, Avirup Sil*

[5] mBART-50 one to many multilingual machine translation Model by Facebook *https://huggingface.co/facebook/mbart-large-50-one-to-many-mmt*