Multilingual Question Answering (MQA)

Group Name – BitsNBytes Subtask ID – 03
Sayan Mahapatra (21CS60R12), Arkapravo Ghosh (21CS60R64), Anima Prasad (21CS60R66)

Abstract

In this project we have experimented with multilingual models (mBERT and XLM-R) and tried to solve the Multilingual Question Answering NLP Task. In this task we are given a passage and a question and we are required to find an answer from the given passage based on the question. For this task we have used multilingual pretrained models from Hugging Face and then finetuned the models for our specific task. The models are trained with 9 languages in total and then we used Bengali and Telugu as the validation set to evaluate the models. We also experimented with training the model only on the English, Telugu and Bengali language and then evaluating on the same Bengali and Telugu evaluation dataset as used earlier.

1 Subtask ID + Group Details (Names, Roll Numbers, Group Name)

- Subtask ID – 03
- Group Name – BitsNBytes
- Members – Sayan Mahapatra, Arkapravo Ghosh, Anima Prasad

2 Individual Contributions of Students

- Sayan Mahapatra – worked on task1 mBERT model and task2 XLM model and data augmentation. Explored the IndicBERT experiment to find if promising results could be obtained. Explored the AI4Bharat dataset and formatted the dataset to match with the existing SQUAD format for data augmentation purposes. Also worked on the report for the final project.
- Arkapravo Ghosh – worked on task1 XLM model and task2 (mBERT + Dutch) model and data augmentation. Explored the AI4Bharat dataset and also tried to choose the subset of data with promising f1-score. Also worked on the report for the final project.
- Anima Prasad – worked on task2 mBERT model and task1 (mBERT + Dutch) model and data augmentation. Explored the AI4Bharat dataset and augmented the chosen subset of data with the squad dataset for self-training the model. Also worked on the report for the final project.

3 Task Description

Multilingual Question Answering is considered one of the more challenging NLP Tasks. Given a context (passage), and a question the task is to extract out the answer to the question from the context.

The Stanford Question Answering Dataset (SQuAD) is benchmarks dataset. In this project we experiment on a multilingual version of the dataset, TyDi QA [1] dataset. mBERT and XLM were used to obtain baseline performance on this dataset and then data augmentation was considered as the next step for improving over the baseline performance. We also tried out two other model – IndicBERT [2] without the use of translation (unlike MLQA and XQuAD)

4 Approach / Model Architectures

The following two models were used

- BERT multilingual base model (cased) (referred to as mBERT) [3]
- XLM-RoBERTa [4]

We tried various approaches. Firstly, we used the whole Tydi-QA gold passage for training and evaluated Tydi-QA dataset dev data (only Telugu

and Bengali) on the models. This is done as part of Task 1.

Secondly, as part of task 2 we extract English, Telugu, and Bengali from the Tydi-QA gold passage train dataset to train the base models, and evaluate the Tydi-QA dataset dev data (only Telugu and Bengali).

These were used to set the baseline performance. After this we tried Data Augmentation approach to improve the performance over the baseline

As part of Data Augmentation, we used the AI4Bharat Indic Question Answering dataset [5]. This dataset was not in SQuAD format hence data preprocessing was done to convert it to SQuAD format. Bengali & Telegu Language data was used. We also used the SQuAD v1 dataset English data for data augmentation.

Once the all datasets were in SQuAD format, they were merged with the TyDi-QA dataset and the best performing model from baseline was run

5 Metrics used

The metric used for the performance measurement us Validation Set F1-score and Exact Match.

6 Experiments

We wanted to investigate another multilingual model IndicBERT. Preliminary experiments showed that the model was not performing well for Question Answering task, hence this model was not explored further.

7 Results / Discussions

The figure below shows our baseline results. mBERT (trained for 2 epochs) was the best performing model

Results of baselines:

Parts	mBERT		XLM		mBERT Multilingual + Dutch Model	
	Epoch 1	Epoch 2	Epoch 1	Epoch 2	Epoch 1	Epoch 2
Part 1	80.9664	82.2277	81.5198	NA	79.4113	81.0626
Part 2	78.7635	80.8313	77.8194	81.3484	78.8825	80.579

Across all runs our F1 scores improved Epoch over Epoch. NA entries in the table above were for runs which failed due to hardware limitations (we used Kaggle Notebooks)

After data augmentation Validation Accuracy for task Part 2 mBERT (trained for 1 epoch) improved from 78.76 to 80.3715. We expect that the accuracy would improve further if training is done for more epochs.

8 Difficulty Faced

Data preprocessing, finding good data splits, and hardware limitations were the chief difficulties we faced.

Acknowledgments

We would like to thank our course instructor Prof. Pawan Goyal for giving us this project from where we could learn a lot about the multilingual question answering task. We would also like to thank our TA Aniruddha Roy for his immense help and guidance in this project.

References

1. https://github.com/google-research-datasets/tydiqa
2. https://huggingface.co/ai4bharat/indic-bert
3. https://huggingface.co/bert-base-multilingual-cased
4. https://huggingface.co/xlm-roberta-base
5. https://huggingface.co/datasets/ai4bharat/IndicQuestionGeneration/tree/main/data