Appendix

1 Evaluation Tool

There have been many efforts in Question Answering in multiple languages. To evaluate each model with the actual and predicted answers we have created an Evaluation Tool\(^1\). The tool provides many features to the researcher to analyze the results with user-friendly UI/UX. We have used Flask\(^2\) as a backend by synchronizing with the evaluating script. Our tool can list the best and worst-performing samples for further analysis and can display collective EM and F1 score for the input samples. Preview images of the tool are given in the Figures 1, 2, 3.

Our code and evaluation scripts are available at: https://github.com/Question-Answering-Analytic-Tool

---

\(^1\)Concerning evaluation script available for SQuAD: https://rajpurkar.github.io/SQuAD-explorer/

\(^2\)Jaccard score: https://flask.palletsprojects.com/en/2.0.x/

Figure 1: Upload training dataset and evaluated prediction json file.

Figure 2: Generating preview of: All Prediction, n-Best predictions and n-Worst predictions after evaluation predictions

Figure 3: Prediction Report with F1 score and Exact Match

2 Mathematical representation of Jaccard Score and WER:

Formula of Jaccard Score 1 and Word error rate (WER) 2 are mentioned below:

\[
J(A^A, A^P) = \frac{|A^A \cap A^P|}{|A^A \cup A^P|} \quad (1)
\]

Jaccard Score: \(A^A\) and \(A^P\) represents the actual and predicted answers respectively.

\[
WER = \frac{S + D + I}{N} \quad (2)
\]
WER: S is the number of substitutions, D is the number of deletions, I is the number of insertions, N is the number of words in the reference

3 Dataset Format

Figure 4 shows the format of dataset which we have compiled by combining MLQA and XQuAD.

```json
{
    "answers": {
        "answer_start": [378]
    },
    "text": ["उत्तर पूँछ"]
}
```

Figure 4: Json structure of dataset which we have used to fine-tune and evaluate the model.

---

3MLQA - https://github.com/facebookresearch/MLQA
4XQuAD - https://github.com/deepmind/xquad