Deep learning based question answering system in Bengali

Tasmiah Tahsin Mayeesha, Abdullah Md Sarwar and Rashedur M. Rahman

Department of Electrical and Computer Engineering, North South University, Dhaka, Bangladesh

ABSTRACT

Recent advances in the field of natural language processing has improved state-of-the-art performances on many tasks including question answering for languages like English. Bengali language is ranked seventh and is spoken by about 300 million people all over the world. But due to lack of data and active research on QA similar progress has not been achieved for Bengali. Unlike English, there is no benchmark large scale QA dataset collected for Bengali, no pretrained language model that can be modified for Bengali question answering and no human baseline score for QA has been established either. In this work we use state-of-the-art transformer models to train QA system on a synthetic reading comprehension dataset translated from one of the most popular benchmark datasets in English called SQuAD 2.0. We collect a smaller human annotated QA dataset from Bengali Wikipedia with popular topics from Bangladeshi culture for evaluating our models. Finally, we compare our models with human children to set up a benchmark score using survey experiments.

ARTICLE HISTORY

Received 17 June 2020
Accepted 3 October 2020

KEYWORDS

Natural language processing; question answering; reading comprehension

1. Introduction

Question Answering (QA) involves building systems that are able to automatically respond to questions posed by humans in a natural language. Text based Question Answering tasks can be formulated as information retrieval problems where we want to find the documents that answer a certain question, extract the potential answers from the documents and rank them, or as reading comprehension problems where the task is to find the answer (also called span) from a context passage of text. While Question Answering tasks can encompass visual (context being image), open domain, multimodal (context can be image/video/audio/text) or focus on common sense reasoning tasks, in this work we focus on text-based reading comprehension. Reading comprehension models have great practical value especially for industry purposes as a properly trained reading comprehension model can function as a chatbot for answering frequently asked questions. The driver behind progress of question answering research has been the curation of multiple high-quality large reading comprehension datasets and release of models performing well on these datasets. However, for low resource languages like Bengali similar efforts to annotate large high-quality reading comprehension datasets...
can be costly and challenging due to lack of skilled annotators, awareness and Bengali specific NLP tools.

To create a QA system, we can train models from scratch or use transfer learning. Due to lack of QA specific large dataset in Bengali training models from scratch is unfeasible, so we use transfer learning. Transfer learning refers to adapting a pretrained model for one domain or task to another domain or task. In the context of NLP unsupervised language models like BERT, ELMo, ULMFit (Devlin et al., 2018; Peters et al., 2018; Howard & Ruder, 2018) trained on large linguistic corpora have been successfully adapted to downstream tasks like classification, named entity recognition, question answering and machine translation. The experiments from (Lample & Conneau, 2019; Devlin et al., 2018; Wu & Dredze, 2019), showed that with pretrained language representation models trained on multilingual corpora such as multilingual BERT zero-shot learning is feasible for several natural language processing tasks including NER, POS, machine translation. Zero shot transfer learning refers to use a pretrained model on a new task without using any labelled examples for supervised training. The idea is to solve a task without receiving any sample of that respective task during training. The reason that this process is efficient is because the model has the ability to perform extractive question answering, but not necessarily in the desired language. Existing research (Hsu et al., 2019) on cross-lingual transfer learning on reading comprehension datasets showed successful results using multilingual BERT for zero shot reading comprehension. The authors found that multi-BERT fine-tuned on training set in source language and evaluated on a target language showed comparable performance to models trained from scratch and that m-BERT model has the ability to transfer between low lexical similarity language pairs such as English and Chinese. We found that zero shot transfer of reading comprehension on Bengali language has not been studied at all and the currently available cross-lingual benchmark reading comprehension datasets such as XQuAD (Artetxe et al., 2019) and MLQA (Lewis et al., 2019) also have not added Bengali as a benchmark language. To the best of our knowledge this is the first work systematically comparing cross-lingual transfer ability of multi-BERT on Bengali in a zero-shot setting with fine-tuned transformer models on synthetic corpus. We depicted our workflow in Figure 1, We also evaluate our

![Figure 1. Simplified diagram of our approach and curated datasets and models.](image-url)
model on a human annotated small question answering dataset for comparison. The contributions of our work are following:

- We use the multilingual BERT model for zero shot transfer learning and fine-tune it on our synthetic training dataset on the task of Bengali reading comprehension. We also use other variants of BERT model like RoBERTa (Liu et al., 2019) and DistilBERT (Sanh et al., 2019) for comparison on both zero shot and fine-tuned model setting.
- The purpose of this paper seeks to contribute a way of identifying passages within a paragraph and answering specific questions in relation to the paragraph given with relevant details provided within the paragraph. As such, this model seeks to save time, make it accessible and benefit, among other things, kids, people with attention deficit hyperactive disorders, or even adults who are seeking specific answers from any form of literature which may otherwise be very time consuming to review thoroughly.
- We have translated a large subset of SQuAD 2.0 (Rajpurkar et al., 2018) from English to Bengali and used this synthetic training set for training and testing our model. While creating this synthetic dataset, we have implemented fuzzy matching to preserve the quality of answers after translation.
- We have created a human annotated Bengali reading comprehension dataset collected from popular Bangla Wikipedia Articles for evaluation of our models.
- We conduct a human survey and compare the results with performances of BERT multilingual and other variants of BERT model, using both fine-tuning technique and zero-shot learning.

The remainder of the paper is organized as follows: Section 2 presents the related work that has been done QA domain for English, Bangla and other languages. Section 3 discusses the variants of BERT models we have used in our experiments. Section 4 discusses how we constructed our dataset. Section 5 and 6 present our evaluation metric and experiment findings, followed by the survey on children in section 7. Finally, sections 8,9 and 10 present our error analysis, impact of the project in the context of Bangladesh and conclusion.

2. Related works

In this section, we discuss the progress on reading comprehension tasks in English and attempts from other languages to replicate the progress, specifically highlighting the benchmark datasets and how translated datasets have been used in other languages for training QA models. We also discuss early attempts of creating QA systems in Bengali language and the limitations and challenges faced in Bengali.

2.1. Question answering in English

Early attempts on reading comprehension tasks include Deep Read (Hirschman et al., 1999) and MCTest (Richardson et al., 2013). From 2015–2016, there have been multiple attempts at curating large scale question answering datasets for English such as CNN/Daily Mail Corpus (Hermann et al., 2015), SQuAD 1.0 (Rajpurkar et al., 2016), RACE (Lai...
et al., 2017), SQuAD 2.0 (Rajpurkar et al., 2018), QuAC (Choi et al., 2015), CoQA (Reddy et al., 2018) and Natural Question Corpus (Kwiatkowski et al., 2019). For English language domain specific QA datasets like NewsQA (Trischler et al., 2017), TriviaQA (Joshi et al., 2017) are also available. Out of these datasets SQuAD 1.1 and SQuAD 2.0 are widely used benchmarks with human performance baselines and also have been included in NLP language understanding evaluation benchmark leaderboards like GLUE (Wang et al., 2018) and SUPERGLUE (Wang et al., 2019).

Since Question Answering literature is broad and extensive, many survey and review papers has been written recently on this topic (Zhang et al., 2019; Baradaran et al., 2020; Gupta & Rawat, 2020), however, we categorize the relevant literature similar to the taxonomy introduced by Zeng et al. (2020) where 57 MRC (machine reading comprehension) tasks have been categorized based on different characteristics, evaluation metric in Table 1. We also add how non-English languages have curated their own corpus for MRC in Table 2.

As seen in Figure 2, Question answering tasks can be classified into multiple variations like general reading comprehension, where the task is to predict an answer from a given context passage and a question about that passage. Other varieties include open domain question answering, where the input is a large collection of unstructured text, multimodal QA where the context passages include other media like video, image, audio etc. instead of only text, common sense QA tasks where the answers require the model to have common sense or world knowledge, complex reasoning tasks where reasoning skills are expected, conversational QA systems where unlike MRC tasks where the context passages are independent of each other, the input is a series of dialogues and cross-lingual QA where the questions and the context passages are of different languages.

Corpus, question and answers can also have different subtypes (Zeng et al., 2020). Unit of corpus can be varied like single/multiple passages, documents, URL, paragraphs with images, only images, video clips and more. Questions can be natural text, cloze style (a sentence with a placeholder for both text or images like fill in the blanks questions for children), or synthetic (instead of a grammatically correct question the question is a list of words, e.g. ‘US president lives in’). Answers can be a span extracted from the context passage (constraint is the span is a part of the context passage), free-form (the answer does not have to be a part of the context passage) or multiple-choice answers including yes/no. The datasets also vary in their complexity, size, source and the chosen evaluation metrics.

QA systems are classified into 9 different types of evaluation metrics according to (Zeng et al., 2020), but commonly used metrics are: Precision, Recall, Accuracy (in general custom defined for datasets), F1, Exact Match, ROUGE, BLEU, METEOR and HEQ. For multiple choice/cloze datasets accuracy is generally used, for multimodal tasks accuracy is defined customly (e.g RecipeQA), for span prediction tasks such as ours Exact Match and F1 score has been used in both our work for Bengali and in other languages in section 2.2. Many of the metrics like ROUGE, BLEU and METEOR are derived from other areas of NLP like neural machine translation and text summarization. Organized overview of the English MRC tasks has been added in Table 1.
| Task | Dataset | Data Source | Context Type | Question Type | Answer Type | Number of Samples | Metric |
|------|---------|-------------|--------------|---------------|-------------|------------------|--------|
| **Reading Comprehension with/without Unanswerable Question** | SQuAD 1.1 (Rajpurkar et al., 2016) | Wikipedia | Text | Natural | Span | 100k | EM,F1 |
| | SQuAD 2.0 (Rajpurkar et al., 2018) | Wikipedia | Text | Natural | Span | 150k | EM,F1 |
| | Natural Questions (Kwiatkowski et al., 2019) | Wikipedia | Text | Natural | Span | 323,045 | Precision, Recall, F1 |
| | CNN/Daily Mail (Hermann et al., 2015) | News | Text | Cloze | Span | 387,420/997,467 | Accuracy |
| | BookTest (Bajgar et al., 2016) | Stories | Text | Cloze | Free-form, multichoice | 14,160,825 | Accuracy |
| | MS MARCO (Nguyen et al., 2016) | Bing | Text | Natural | Free-form | 1,010,916 | ROUGE BLEU |
| **Multihop RC** | HotpotQA (+reasoning) (Yang et al., 2018) | Wikipedia | Text | Natural | Span | 113k | EM,F1 |
| | NarrativeQA (Kočiský et al., 2018) | Movies | Text | Natural | Free form | 46,765 | ROUGE, BLEU, |
| | Qangaroo (Welbl et al., 2018) | Wikipedia, Scientific paper News | Text | Synthesis | Span, Multichoice | 51k+ | Accuracy |
| | MultiRC (Khashabi et al., 2018) | Text | Natural | Free-form, Multichoice | 6k | EM,F1 |
| **Common Sense** | ARC (Clark et al., 2018) | School Science Curriculum Narrative Text | Text | Natural | Free-form, Multichoice | 5k+ | Accuracy |
| | MCScript (Ostermann et al., 2018) | Text | Natural | Free-form, Multichoice | 13,939 | Accuracy |
| | OpenBookQA (Mihaylov et al., 2018) | School Science Curriculum | Text | Natural | Free-form, Multichoice | 5957 | Accuracy |
| | ReCoRD (Zhang et al., 2018) | News | Text | Cloze | Free-form, Multichoice | 120,730 | EM,F1 |
| | CommonSenseQA (Talmor et al., 2019) | Narrative Text | Text | Natural | Free-form, Multichoice | 12,247 | Accuracy |
| | WikiMovies (Hewlett et al., 2016) | Movies | Text | Natural | Free-form | 100k | Accuracy |
| | WikiReading (Miller et al., 2016) | Wikipedia | Text | Synthesis | Free-form | 18.87m | F1 |
| **Complex Reasoning** | DROP (Dua et al., 2019) | Wikipedia | Text | Natural | Free-form | 96,567 | EM,F1 |
| | Facebook CBT (Hill et al., 2016) | Factoid Story | Text | Cloze | 687k | Accuracy |

(Continued)
| Task                           | Dataset                                      | Context Type | Question Type | Answer Type          | Number of Samples | Metric       |
|-------------------------------|----------------------------------------------|--------------|---------------|----------------------|-------------------|--------------|
|                               | Google MC-AFP (Soricut & Ding, 2016)         | Gigaword     | Synthesis     | Free-form, Multichoice | 1,742,618         | Accuracy     |
|                               | LAMBADA (Boleda et al., 2016)                | Book Corpus  | Text          | Cloze                | 10,022            | Accuracy     |
|                               | NewsQA (Trischler et al., 2017)              | News         | Text          | Natural             | 119,633           | EM, F1       |
|                               | RACE (Lai et al., 2017)                      | English Exam | Text          | Natural             | 97,687            | Accuracy     |
|                               | TriviaQA (Joshi et al., 2017)                | Bing         | Text          | Natural             | 70k+              | EM, F1       |
|                               | CLOTH (Xie et al., 2018)                     | English Exam | Text          | Cloze                | 99,433            | Accuracy     |
|                               | ProPara (Dalvi et al., 2018)                 | Process defining paragraph | Text          | Natural             | 488               | Accuracy     |
|                               | DREAM (Sun et al., 2019)                     | English exam | Text          | Natural             | 10,197            | Accuracy     |
|                               | CoQA (Reddy et al., 2018)                    | Jeopardy     | Text          | Natural             | 127k              | F1           |
|                               | QuAC (Choi et al., 2015)                     | Wikipedia    | Text          | Natural             | 98,407            | F1, HEQ      |
|                               | ShARC (Saeidi et al., 2018)                  | Government   | Text          | Natural             | 948               | Accuracy, BLEU |
| Domain Specific               | SciQ (Welbl et al., 2017)                    | Science Exam | Text          | Natural             | 13,679            | Accuracy     |
|                               | CiICR (Suster & Daelemans, 2018)             | Medical      | Text          | Cloze                | 104,919           | EM, F1, BLEU |
|                               | PaperQA (Park et al., 2019)                  | Scientific Paper | Text          | Cloze                | 80,118            | Accuracy, F1 |
|                               | ReviewQA (Grail & Perez, 2018)               | Hotel Comment | Text          | Natural             | 587,492           | Accuracy     |
|                               | SciTail (Khot et al., 2018)                  | School Science Curriculum | Text          | Natural             | 1834              | Accuracy     |
| Visual QA                    | VQA (Antol et al., 2015)                     | Coco Images  | Image         | Natural             | 1,105,904         | Accuracy     |
|                               | MovieQA (Tapaswi et al., 2016)               | Movies       | Multimodal    | Natural             | 21,406            | Accuracy     |
|                               | COMICS (Iyyer et al., 2017)                  | Comics       | Multimodal    | Cloze, Natural      | N/A               | Accuracy     |
|                               | TQA (Kembhavi et al., 2017)                  | School Science Curriculum | Multimodal    | Natural             | 26,260            | Accuracy     |
| Multi Modal QA               | RecipeQA (Yagcioglu et al., 2018)            | Recipes      | Multimodal    | Cloze                | 36k               | Accuracy     |
| Dataset | Domain | Type | Text | Form | Size | Measure |
|---------|--------|------|------|------|------|---------|
| Open Domain QA | MCTest (Richardson et al., 2013) | Factoid | Text | Natural | Free-form, Multichoice | 2000+ | Accuracy |
| | CuratedTREC (Baudiš & Šedivý, 2015) | Factoid | Text | Natural | Free-form | 2180 | EM |
| | Quasar (Dhingra et al., 2017) | Stackoverflow | Text | Cloze, Natural | Span | 37,362(S), 43,013(T) | Accuracy, EM, F1 |
| | SearchQA (Dunn et al., 2017) | Jarchive and Google | Text | Natural | Free-form | 140,461 | Accuracy, F1 |
| | WikiQA (Yang et al., 2015) | Bing Query log, Wikipedia Movies | Text | Natural | Free-form | 3047 | Precision, Recall, F1 |
| MRC with paraphrased paragraph | DuoRC (Saha et al., 2018) | Movies | Text | Natural | Free-form | 100,316 (paraphrase), 85,773(self) | Accuracy |
| | Who-did-What (Onishi et al., 2016) | News | Text | Cloze | Free form | 200k | Accuracy |
2.2. Question answering in non-English languages

For low resource languages curating large reading comprehension datasets are often impossible due to lack of native language speakers, large monolingual corpora, skilled annotators and cost of curating large datasets with correct labels. However there has been work on translating English QA datasets to Arabic (Mozannar et al., 2019), Korean (Lee et al., 2018), Hindi (Gupta et al., 2019) and Spanish (Carrino et al., 2019) question answering systems where they translated English data like SQuAD 1.1 (Rajpurkar et al., 2018) to their respective languages and used transformer models for transfer learning in training their own QA systems; we also closely follow these works in training our Bengali systems. Translated datasets have also been used in making cross lingual benchmark datasets like XQuAD (Artetxe et al., 2019) and MLQA (Lewis et al., 2019).

Aside from using translated dataset, there has also been attempts of curating large question answering datasets in multiple other languages including French (d’Hoffschmidt et al., 2020), Korean (Lim et al., 2019), Russian (Efimov et al., 2020), Chinese (Cui et al., 2018; Shao et al., 2018) and benchmark models like QANet (Yu et al., 2018), BiDAF (Seo et al., 2016), BERT (Devlin et al., 2018) have been trained on them. In contrast to gathering translated or human annotated dataset for model training, zero shot transfer learning where pretrained models were evaluated directly on a new language after task specific training on question answering has also been attempted on reading comprehension tasks (Artetxe et al., 2019; Hsu et al., 2019; Siblini et al., 2019). To the best of our knowledge none of similar work have ever been attempted on Bengali so far. Table 2 presents an overview of QA in different datasets.

Table 2. Overview of SQuAD like QA datasets in different languages.

| Dataset | Language | Number of Questions | Data Collection | Data Source | Creator |
|---------|----------|---------------------|-----------------|-------------|---------|
| ARCD    | Arabic   | 1,395               | CrowdSourced    | Arabic Wikipedia | Mozannar et al. (2019) |
| Arabic-SQuAD | Arabic | 48,344            | Machine Translation | SQuAD 1.1 | Lee et al. (2018) |
| K-QuAD  | Korean   | 77,000+            | Machine Translation | SQuAD 1.1 | Lim et al. (2019) |
| Kor-QuAD | Korean   | 60,407             | Crowd Sourced   | Korean Wikipedia | Gupta et al. (2019) |
| Hindi SQuAD | Hindi | 18,454              | Machine Translation | SQuAD 1.1 | Carrino et al. (2019) |
| SQuAD-es | Spanish | 87,595              | Machine Translation | SQuAD 1.1 | Carrino et al. (2019) |
| FQuAD   | French   | 25,000+            | CrowdSourced    | French Wikipedia | d’Hoffschmidt et al. (2020) |
| SberQuAD | Russian  | 50,364             | CrowdSourced    | Russian Wikipedia | Efimov et al. (2020) |
| DRCd    | Chinese  | 30,000+            | CrowdSourced    | Chinese Wikipedia | Shao et al. (2018) |
| CMRC    | Chinese  | 20,000+            | CrowdSourced    | Chinese Wikipedia | Cui et al. (2018) |
| XQuAD   | Multilingual | 1190(For each 10 language) | Human Translated | SQuAD 1.1 | Artetxe et al. (2019) |
| MLQA    | Multilingual | 12k+ English,5k+ other language | CrowdSourced | Wikipedia | Lewis et al. (2019) |
| TyDi QA | Multilingual | 200k+             | CrowdSourced    | Wikipedia | Clark et al. (2020) (Google AI) |
2.3. Question answering in Bengali

Compared to the maturity of question answering research in English, due to lack of fundamental natural language processing resources and scarcity of data Bengali question answering research is at its infancy. Majority of the QA research in Bengali can be grouped into attempts of creating question classification systems to be used as a component of a full-fledged question answering system or building factoid-like question answering systems using information retrieval centric techniques.

Banerjee and Bandyopadhyay (2012) first attempted to build a question classification system for Bengali by extracting lexical, syntactic and semantic features like wh-word or interrogative words existence and position, question length, end marker, Parts of Speech (POS) tags etc to classify questions into nine categories. The issues they faced were the existence of numerous interrogative words in Bengali compared to English, the fact that Bengali interrogatives can appear in any part of the sentence and lack of high-quality NLP tools like POS tagger, NER system and benchmark corpora. Nirob et al. (2017) also built a question classification system with similar feature extraction methods along with using uni-grams as features with support vector machines.

Banerjee et al. (2014) attempted to build the first Bengali factoid question answering system called BFQA, a complex IR system that classifies questions, retrieves relevant sentences, ranks them and extracts correct answers. BQAS (Hoque et al., 2015), a bilingual question answering system that can generate and answer fact-based questions from Bengali and English documents was also built. Islam and Nurul Huda (2019) also implemented a similar QA system that extracts relevant answers from multiple documents and gives specific answers for time related questions. However, to best of our knowledge no attempt with deep learning techniques on SQuAD like reading comprehension datasets so far has been made in Bengali previously.

3. Models

In a typical NLP pipeline a text is initially tokenized, encoded to numerical vectors and then fed as input to the models. To encode the words as numerical vectors many different representations so far has been consider like Bag of Words, tf-idf, Word2Vec, GLOVE, FastText, each with their own limitations. In BOW representation each word is assigned to a one-hot-encoded vector, however as the vocabulary
Table 4. Results from the zero shot setting experiments where we evaluate the pretrained models directly on the validation set.

| Model       | Pretraining Language               | Exact Match | F1 score |
|-------------|------------------------------------|-------------|----------|
| BERT        | multilingual (includes Bengali)    | 47.18       | 48.52    |
| DistilBERT  | Multilingual (contains Bengali)    | 50.05       | 51.18    |
| RoBERTA     | English                            | 54.41       | 54.41    |

Table 5. DistilBERT and RoBERTa results broken down by answerable vs unanswerable questions.

| Type   | Number of Examples | DistilBERT | RoBERTa |
|--------|--------------------|------------|---------|
|        |                    | EM | F1 | EM | F1 |
| Total  | 5865               | 64.14 | 64.13 | 66.34 | 66.34 |
| HasAns | 1971               | 15.47 | 16.25 | 10.45 | 10.45 |
| NoAns  | 3272               | 93.45 | 93.41 | 100  | 100  |

Table 6. Results from training the models on the translated training dataset for 5 epochs.

| Model   | EM(test) | F1(test) |
|---------|----------|----------|
| BERT    | 33.80    | 36.38    |
| DistilBERT | 35.61    | 38.75    |
| RoBERTA | 19.16    | 19.16    |

Table 7. Evaluation on a human annotated dataset of 134 questions from 53 paragraphs.

| Model       | Setting       | Exact Match | F1 score |
|-------------|---------------|-------------|----------|
| DistilBERT  | zero-shot     | 25.76       | 32.89    |
| BERT        | zero-shot     | 37.88       | 53.15    |
| RoBERTA     | zero-shot     | 17.43       | 17.43    |
| DistilBERT  | fine-tuned    | 21.21       | 30.61    |
| BERT        | fine-tuned    | 31.06       | 43.63    |
| RoBERTA     | fine-tuned    | 16.67       | 17.05    |

Table 8. Exact Match and F1 scores from all models in zero shot and fine-tuned setting.

| Model       | Setting       | Exact Match | F1 score |
|-------------|---------------|-------------|----------|
| BERT        | zero-shot     | 14.16       | 24.77    |
| DistilBERT  | zero-shot     | 29.16       | 33.80    |
| RoBERTA     | zero-shot     | 19.16       | 19.16    |
| DistilBERT  | fine-tuned    | 32.5        | 42.47    |
| BERT        | fine-tuned    | 11.66       | 23.40    |
| RoBERTA     | fine-tuned    | 19.16       | 19.16    |

Table 9. Survey Results for zero-shot and fine-tuned models according to grades.

| Model       | Setting       | Grade 3 | Grade 4 |
|-------------|---------------|---------|---------|
|             |               | EM | F1 | EM | F1 |
| BERT        | zero-shot     | 3.33 | 17.02 | 25.0 | 32.53 |
| DistilBERT  | zero-shot     | 13.33 | 20.06 | 45.0 | 47.55 |
| RoBERTA     | zero-shot     | 19.16 | 19.16 | 33.33 | 33.33 |
| BERT        | fine-tuned    | 3.33 | 19.967 | 3.33 | 19.967 |
| DistilBERT  | fine-tuned    | 16.66 | 30.67 | 48.33 | 54.26 |
| RoBERTA     | fine-tuned    | 5.0 | 5.0 | 33.33 | 33.33 |
grows this approach has its limitations. Tf-idf representations have been used in early age information retrieval systems, however it relies heavily on term overlaps between query and the document and unable to learn more complex representations. Word2Vec, GLOVE and FastText are considered distributed representations where the semantic relationship between different words are learnt, out of these FastText is able to handle out of vocabulary words since it relies on learning character level n-gram representations, however FastText can not handle Polysemy(a word can have different possible meaning based on its context). For a complex task like question answering having contextualized word representations (where the word embeddings take context into account by looking at the entire sentence) is very important. While models like BiDAF (Seo et al., 2016), DocQA (Clark & Gardner, 2017), DrQA (Chen et al., 2017) showed good performance on SquAD 1.1 leaderboard, all of the current high performing models in SquAD 2.0 use variations of BERT based models which can produce contextualized representations of words. Since we wanted to create a benchmark dataset and models for deep learning based Bengali Question Answering work, we followed the approaches from Russian (SberQuAD), Spanish (SquAD-es), Arabic and French language (details in section 2.2) and focused on training multilingual BERT models with contextualized representations.

While the ideal would have been to train BERT models from scratch on Bengali and then fine-tuning them for QA in Bengali, as a low resource language Bengali does not have large unsupervised text datasets like English. Transfer learning in NLP context generally refers to learning embedding or language representation in an unsupervised setting over a large corpus and then modifying the architecture for task specific enhancements. Also, since transfer learning requires significantly less data and training compared to training from scratch, due to lack of computational resources and data it is more advantageous for us to use transfer learning models pretrained on large multilingual corpus like Wikipedia of different language. From that inspiration we have chosen BERT (Devlin et al., 2018), DistilBERT (Sanh et al., 2019), a compressed version of the original BERT and RoBERTA (Liu et al., 2019) as our chosen models for prototyping a Bengali question answering solution. The components of the models and our hyperparameters are described in the following sections.

### 3.1 BERT

BERT (Devlin et al., 2018) refers to a neural network pretraining technique called Bidirectional Encoder Representation from Transformers. Due to its excellent performance in several NLP tasks on GLUE (Wang et al., 2019) including SQuAD 1.1 and SQuAD 2.0. There have been many other architectures that are variants of BERT. It has recently been included in Google Search too. BERT is based on transformers which processes words in relation to all the other words in a sentence instead of looking at them one by one. This allows BERT to look at contexts both before and after a particular word and helps it to pick up features of a language.

#### 3.1.1. BERT architecture

BERT makes use of the transformer architecture (Vaswani et al., 2017) which uses the self-attention mechanism to learn contextual relationships between words and sub-words.
general, the transformer architecture uses encoder-decoder method, an encoder for reading the sequence of inputs and the decoder that creates predictions for the actual task. Sequence to sequence models based on RNN or LSTM architectures reads the sequence left-to-right or right-to-left, but transformer encoder reads the entire input sequence at once. BERT however only uses the encoder half of the transformer. BERT base architecture has 12 transformer blocks, 768 hidden size and 12 attention heads, total around 110M parameters. BERT large architecture has 24 transformer blocks, 1024 hidden size and 16 attention heads with a staggering number of 340M parameters. Input for BERT is a sequence of tokens which are mapped to embedding vectors and then processed in the transformer blocks. The output of the encoder is a sequence of vectors indexed to tokens each of hidden size H.

### 3.1.2. Input representation

Input representation for BERT can handle both a single sentence and a pair of sentences representation e.g (Question, Answer) in a single token sequence. A sentence is considered to be an arbitrary span of contiguous text rather than a linguistic sentence. A sequence can be one sentence or two sentences concatenated together.

Word piece tokenization is used with a 30k vocabulary. First token of every sequence is always a special token called the [CLS] token. For classification tasks the final hidden state assigned to the CLS token is used as input. If pair of sentences are given as input then the sentences are separated using a separator token [SEP]. Segment embedding are also used to indicate whether a sentence is from sentence 1 or 2. The final input to the model is the summation of the token embedding, segment embedding and positional embedding.

### 3.1.3. Pretraining BERT

Traditionally language models are trained to predict the next word from a sequence of word, either from left-to-right or right-to-left. For example, ‘The cat is sitting’ would be the context and the next target word would be ‘on’. However, BERT is pretrained using two unsupervised tasks called masked language modelling and next sentence prediction.

**Masked Language Modelling:** To train a deep bidirectional representation of the input BERT masks some percentage of the input tokens at random and then predict those masked tokens for training. The final output vectors from the encoder are fed into a softmax layer over the vocabulary to get the word predictions.

**Next Sentence Prediction:** Many downstream NLP tasks like question answering is based on understanding the relationship between two sentences which is not directly captured by language modelling. In order to capture sentence level relationships, the unsupervised task of next sentence prediction is added. For generating examples of this task in each example 50% of the time the next sentence is the real next sentence and other half of the time it is a random sentence from the corpus.

### 3.1.4. BERT for question answering

With the utilization of pre-trained BERT models, we can use pre-trained memory data of sentence structure, language, and grammar related memory of huge corpus of millions, or billions, of commented on preparing models, that it has been trained on. The pre-trained model would then be able to be adjusted on little information NLP errands like question
answering and sentiment analysis, bringing about notable performance enhancements compared to training on these datasets from scratch.

BERT model is slightly modified for the question answering task. Given the question and the context the total sequence is considered to be a single packed sequence. The question is assigned the segment A embedding and the passage is given the sentence B embedding. Two new vectors of same dimensionality as the output vectors is added, the start vector and the end vector. The probability of a word being the start of answer span is computed as the dot product between the output representation of the word and the start vector followed by a softmax over the set of words in the passage. The probability of a word being the end of span is also calculated by Equation (1).

$$P_i = \frac{e^{s_i}T_i}{\sum e^{s_i}T_i}$$

(1)

The score of a candidate span from ith word to jth word is defined as $s_iT_i + e^jT_j$. The maximum scoring span where $j \geq i$ is used as a prediction. The training objective is the sum of the log-likelihoods of the correct start and end positions. For SQuAD 2.0 for questions that do not have any answer start and end spans are considered to be at the CLS token.

### 3.2. DistilBERT

DistilBERT (Sanh et al., 2019) is a compressed version of BERT where knowledge distillation technique was used to compress the BERT architecture. Using knowledge distillation in the pretraining phase of BERT it was possible to reduce the size of a BERT model by 40% while retaining 97% of its language understanding capabilities and being 60% faster. For training it uses a triplet loss combining language modelling, distillation and cosine embedding loss. Model compression techniques like knowledge distillation is motivated by the desire to deploy massive architectures like transformers to low computing resource environment as well as reducing training time and computational cost.

#### 3.2.1. Knowledge distillation

Knowledge Distillation (Hinton et al., 2015) is a compression technique where a smaller compact model is trained to reproduce the behaviour of a larger model or an ensemble of models. The larger model is often called the teacher and the smaller model is called the student. In knowledge distillation the student network minimizes a loss function where the target is the distribution of class probabilities predicted by the teacher. This probability distribution generally has the correct class at a high probability while the other classes have near zero probability. Knowledge Distillation can be thought of the teacher network teaching the student how to produce outputs like itself. Both of the networks are fed same input. While the target of the teacher network are the actual labels, the student network is rewarded for mimicking the behaviour of the teacher network.

The student is trained with a distillation loss over the soft target probabilities of the teacher. Following softmax temperature is also used where $t$ controls the smoothness
of the output distribution. It is shown in Equation (2).

\[ L_{ce} = \sum t_i + \log(s) \]

\[ p_i = \frac{\exp \left( \frac{z_i}{T} \right)}{\sum \exp \left( \frac{z_i}{T} \right)} \]  

(2)

3.2.2. DistilBERT architecture

The student network is a small transformer architecture which is trained with the supervision of a larger transformer architecture (BERT-base) and initialized from the layers of the teacher architecture. The number of layers is reduced by a factor 2 and other minor modifications like removing token type embedding and poolers are also done. DistilBERT was also trained on the same corpus like the original BERT model, a concatenation of English Wikipedia with other large datasets. The training loss is a linear combination of distillation loss and supervised training loss (in this case masked language modelling loss) as the next sentence prediction objective was dropped. A cosine embedding loss was also added to align the embedding of the student and the teacher transformer network.

3.4. RoBERTa

Introduced by Facebook, RoBERTa (Robustly Optimized BERT Pretraining Approach) (Liu et al., 2019), is a retraining of BERT with various adjustments to improve performance. These include: expanding the batch size, training the model for a more extended timeframe, and over larger data. This helps in refining the perplexity for the masked language model goal, and the end-task accuracy. The next sentence prediction (NSP) goal of BERT was eliminated. RoBERTa also introduced dynamic masking so that the masked tokens were changed during training epochs and used 160 GB of text for pretraining.

4. Dataset collection

Since Bangla is a low resource language and so far, there has been no crowdsourced human annotated reading comprehension dataset collected in Bangla so far, we translated a large benchmark dataset of reading comprehension from English to Bangla to create a synthetic training set. However, for evaluation purposes and in hope of creating a human annotated QA dataset for Bengali for future work, a second dataset from Bengali Wikipedia was also curated.

We used the SQuAD 2.0 (Stanford Question Answering Dataset, Rajpurkar et al., 2018), a reading comprehension dataset, composed of questions annotated by crowd workers on a selection of Wikipedia articles for the training dataset that was translated to Bengali. The answer to each question is a section of the text (a span) of the corresponding reading passage. Our experiments were performed based on two datasets: Translated subset of SQuAD 2.0 training dataset from English to Bengali and a human annotated benchmark evaluation dataset gathered from Bengali Wikipedia passages.
4.1. Bengali-SQuAD (Translated)

While the crowd sourced annotated dataset has been built for model evaluation, training data obtained by translation had enough noise that can be sufficient for model training. Similar synthetic data obtained by translation of SQuAD 1.1 (Rajpurkar et al., 2016) has been used in training Arabic (Mozannar et al., 2019), Korean (Lee et al., 2018), Hindi (Gupta et al., 2019) and Spanish (Carrino et al., 2019) Question Answering systems. Recently launched MLQA (Lewis et al., 2019) and XQuAD (Artetxe et al., 2019) cross lingual parallel QA datasets also use translated SQuAD to multiple popular languages like Spanish, Chinese, German, Arabic etc. for cross lingual question answering.

Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2016), is a large-scale reading comprehension dataset collected in English with annotations from crowd workers. The dataset contains around 100k QA pairs from 442 topics. For each topic there is a set of passages and for each passage QA pairs are annotated by marking the span or the part of text that answers the question. SQuAD 2.0 (Rajpurkar et al., 2018) poses a more challenging task by adding 50k more adversarial questions to the dataset where the questions are impossible to answer. The adversarial questions are introduced so that QA systems built with the dataset are capable of understanding when a question cannot be answered as well as answer the possible questions properly. Since SQuAD 2.0 dataset contains a leaderboard for model comparison along with a human baseline score, SQuAD 2.0 was chosen for translation to Bengali to create synthetic training data.

Google cloud translation API was used to translate the context, question and answers of SQuAD 2.0 samples for 294 articles. We then randomly split the data into 80:20 split for training and validation sets. The 235 training topics consisted of 11,588 paragraphs and 73,812 question answer pairs, while 59 validation topics consisted of 17,607 questions and 2714 paragraphs. Out of 73,812 question answer pairs in the training set 36,067 questions are answerable and 37,745 questions are unanswerable. Out of 17,607 question answer pairs in the test set 8166 questions are answerable and 9441 questions are impossible to answer from the context.

The format of the dataset mimics that of SQuAD 2.0. Each sample in the dataset is a triplet of context paragraph, question and answer structured like the following:

- context: The paragraph or text from which the question is asked
- qa: A list of questions and answers. The list contains the following:
  - id: A unique id for the question
  - question: A question.
  - is_impossible: A Boolean value indicating if a question can be answered correctly from the provided context.
  - answers: A list that contains following:
    * answer: The answer to the question
    * answer_start: The starting index of the answer in the context

4.1.1. Examples and basic statistics

We summarize the dataset statistics in Table 3 with general statistics. Figure 3 shows a sample translated SQuAD 2.0 context paragraph with a question and marked span in the context colour coded to show the similarity between context span and answer. Figure 4 shows question, answer and context length histograms for the translated dataset.
After translation we found by observation that the context, questions and answers were translated of high quality in general and the meanings of the sentences were preserved. However, neural machine translation tends to be context aware. Identical words and phrases in the context passages and the questions/answers can be translated differently based on the context words which violates the constraint that the answer of a question should be in the span of the passage for the reading comprehension task. We found that in many instances data was corrupted (Figure 5) due to translation errors such as minor spelling mistakes, insertion of special characters and inconsistent translations of words and phrases in the answer and the paragraph where the answer could no longer be found from the context. For named entities specifically minor mistakes like insertion of vowels at the end of the name can change the meaning.

These issues also occurred in the Arabic (Mozannar et al., 2019) QA system where they found the machine translation was unable to recognize named entities without context and thus transliterated them instead of translating them and for Arabic specifically minor typographic errors like adding glyphs in wrong places can vastly change the meaning of the sentences and Korean (Lee et al., 2018) model where they found filtering noisy question-answer pairs out they could enhance the performance of the system.

An alternative approach for translating using Google Cloud Translation – Basic, is to use HTML tags like in Figure 6. The API does not translate the HTML tags, but it translates the

![Question Answering Taxonomy](image)

**Figure 2.** Classification of QA literature based on task characteristics, corpus, evaluation metric, question and answer type.

![Sample Bangla Translated SQuAD context, question and answer.](image)

**Figure 3.** Sample Bangla Translated SQuAD context, question and answer.
4.1.2. Fuzzy answer matching
To handle the data corruption issue, we used fuzzy string matching to find the approximate answer from the context by comparing to the span. We compute the Levenshtein Distance between the translated Bengali answer and all possible spans of the translated context paragraph to find near matches in an approximate substring search. If we can find an approximate match in the given edit distance of min (10, answer_length/2), we replace the substring in the context to match with the answer. If the minimum edit distance is larger than the constraint, we drop the answer and treat it as an impossible to answer question from the context. Following this approach (Figure 7), we have been able to recover the majority of corrupted answers.

4.2. Bengali question answering evaluation dataset (CrowdSourced)
To compare our models for evaluation, a human annotated small test set with questions written in Standard Bengali was also collected using Bengali Wikipedia articles. We collected 330 Bengali Wikipedia articles using mean number of views over a week as a measure of popularity of the topic in November, 2019. We used Wikipedia-API, a python wrapper for Wikipedia’s official API for scraping which provides texts, sections, links, categories and other information from articles. The articles covered a huge range of topics including religious, historical and political figures, pages of specific countries and organizations and more.

Then we split the text of each article into context paragraphs and filtered paragraphs with less than 600 characters. This resulted in 2306 unique context paragraphs. These
Figure 5. Context, Question and Answer of corrupted samples. Answer and the context has been highlighted by blue underline to show the mismatch.

Figure 6. Using HTML tags to preserve the quality of translation.

Figure 7. Actual Translation vs Fuzzy Translation.
paragraphs were annotated using CDQA (Closed Domain Question Answering) Suite Annotation tool and 300 question answer pairs were gathered for test purposes. We will use the paragraph sets to create a larger human annotated dataset for further experiments.

During annotation the annotator is shown a context paragraph and can add questions using the tool (Figure 8). For each question the annotator selects the smallest span of the
5. Evaluation metrics

We follow the metrics for comparison in the English SQuAD leaderboard, exact match (EM) score and F1 score.

context that answers the question. For questions that do not have any answer the annotator keeps the answer field empty. For each paragraph we collected at least three question-answer pairs.
Figure 13. Samples of correctly predicted answers.

Figure 14. Samples of incorrectly predicted answers.

Figure 15. Attention Visualization for the BERT model on Bangla data.
5.1. Exact match

This is a binary (true/false) metric that measures the percentage of predictions that match any of the ground truth answers exactly. For a question if the predicted answer and the ground truth answers are exactly the same then the score is 1, otherwise 0. For example, if the predicted answer is ‘Einstein’ and ground truth is ‘Albert Einstein’, EM or exact match will be 0.

\[
EM = \frac{\sum_{i=1}^{N} F(x_i)}{N}
\]

where \( F(x_i) = \begin{cases} 1, & \text{if predicted answer} = \text{correct answer} \\ 0, & \text{otherwise} \end{cases} \).

5.2. F1 score

F1 score is a less strict metric than exact match calculated as the harmonic mean of precision and recall. Precision and recall are calculated while treating the prediction and the ground truth answers as bag of tokens. Precision is the fraction of predicted text tokens which are present in the ground truth text. Recall is the fraction of ground truth tokens that are present in the prediction.

\[
F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]

For the ‘Einstein’ example, the precision is 100% as the answer is a subset of the ground truth, but the recall is 50% as out of two ground truth tokens [‘Albert’, ‘Einstein’] only one is predicted, so resulting F1 score is 66.67%.

If a question has multiple answer then the answer that gives the maximum F1 score is taken as ground truth. F1 score is also averaged over the entire dataset to get the total score.

6. Experiments and results

6.1. Zero-Shot transfer with BERT and variants

BERT based models are first trained on a language modelling task such as masked language modelling and next sentence prediction in case of BERT, then the last hidden layer is modified in a task specific manner for fine-tuning the pretrained model to perform well on that specific task like question answering. Zero shot cross lingual transfer learning (Conneau et al., 2018; Pires et al., 2019) in the NLP context refers to transferring a model which is trained to solve a specific task in a source language to solve that specific task in a different language. Initially we get our baseline results using the zero-shot setting, where we use pretrained transformer models fine-tuned on English question answering task and check their performance on the Bengali evaluation dataset following similar research on Chinese (Hsu et al., 2019), French (d’Hoffschmidt et al., 2020) and Japanese (Siblini et al., 2019) language. In the next section we fine-tune these models further with our translated Bengali SQuAD dataset and compare the baselines with the fine-tuned models.
In our experiments we leverage the pretrained models provided by Transformers library of Huggingface (Wolf et al., 2019) which provides ready-to-use implementations and pretrained weights for multiple transformer-based models such as BERT, RoBERTa, GPT-2 or DistilBERT, which achieved state of the art performance on multiple NLP tasks like text classification, data retrieving, question answering, and text generation. The pretrained m-BERT model has 110 million parameters, 12 layers, $H=768$ and 12 self-attention heads and it is very computationally expensive to fine-tune directly on BERT. However, we have been able to evaluate our translated test set with the m-BERT model. The pretrained multilingual BERT from Google was pretrained on a corpora of 104 languages including Bengali. Data in different languages were concatenated to one another without any special pre-processing while training m-BERT.

We fine-tuned multilingual BERT (Figure 9), a compressed version of BERT called DistilBERT and RoBERTa, another variant of BERT pretrained on only English corpora as our selected models on our translated synthetic Bengali dataset and compared our results with the zero-shot setting. For the zero shot results we fine-tuned the models for 2 epochs on the English SQuAD dataset to teach them about the task of question answering and evaluated directly on the Bangla translated dataset. Following hyperparameters were used for all models during training on both English and Bengali corpus in the next section, however for further experiments we trained on Bengali translated set and increased the number of epochs to 5.

- **learning_rate**: 5e-5. The initial learning rate for Adam.
- **max_seq_length**: 384. The maximum total input sequence length after WordPiece tokenization. Sequences longer than this will be truncated, and sequences shorter than this are padded.
- **docs_stride**: 128. When splitting up a long document into chunks, how much stride to take between chunks.
- **max_query_length**: 64. The maximum number of tokens for the question. Questions longer than this will be truncated to this length.
- **max_answer_length**: 30. The maximum length of an answer that can be generated.
- **train_batch_size**: 8. Batch size per GPU/CPU for training.
- **gradient_accumulation_steps**: 8. Number of backward steps to accumulate before making an update.
- **activation**: GELU. Gaussian Error Linear Unit.

Table 4 contains the results we obtained from the experiments in the zero-shot setting.

### 6.2. Fine-Tuning on translated train data with BERT variants

After performing the zero shot experiments, we wanted to compare the models finetuned on the translated train Bengali-SQuAD data. For finetuning we trained the DistilBERT, ALBERT and RoBERTa models on our translated synthetic training set for 5 epochs with the same hyperparameters. Note that the data corruption issue mentioned in section 4 was resolved by the fuzzy answer matching for preprocessing and all experiments followed are done on this preprocessed dataset instead of the raw translations. As the models already gained knowledge of language specific structure and semantics from
pretraining, adapting them to Bengali question answering task requires a small dataset only to get overall good results. We have shown the results we obtained from DistilBERT and RoBERTa in Table 5.

To see the impact of fine-tuning on a larger dataset we augmented our translated dataset with more topics and collected the Bangla-SQuAD dataset described earlier. After fine-tuning on this synthetic training set for 5 epochs with the same hyperparameters we get the results shown in Table 6. We found on a large dataset to get similar results we would have to train longer. Our next set of experiments will focus on training the models longer and adding more human annotated dataset to the training set. As running the experiments is computationally expensive and each model takes around 24 h to train only for 5 epochs in our low-cost computation environment, we really require better hardware for improved results.

6.3. Evaluation on human annotated data

We evaluate our zero shot and fine-tuned models on the Bengali Question Answering Evaluation dataset from section 4. Results are given in Table 7.

6.4. Computational cost

According to BERT official implementations from Google Research pretraining BERT from scratch is computationally expensive and to pre-train a BERT-Base on a single preemptible Cloud TPU v2, may take about 2 weeks at a cost of about $500 USD. DistilBERT training requires 4x less time than BERT and RoBERTa is more computationally extensive as its trained on 160 GB data compared to BERT(16GB data). In our lab 16 GB NVIDIA gpu was used to fine tune the models. While the zero shot experiments only require inference and can be done in an hour, the fine-tuning experiments for BERT and DistilBERT took more than 24 h and RoBERTa took around 3 days. We believe with better hardware and GPU our models will be able to reach significantly better performance even with synthetic data.

7. Survey on Bangladeshi children

Our purpose on surveying on Bangladeshi children was to experiment with data from a real-life application. Wikipedia is an enormous database, but it is also formal and does not reflect the practical use of Bengali language. In Bangladesh’s primary education sector, books such as Bangla Literature and Social Science provide a much more realistic use of the language. The primary education sector in Bangladesh is primarily focused on reading comprehension style learning. The children are tested mainly on the basis of multiple-choice questions or brief questions from short passages. With the rate of students and teachers being poor relative to the population in primary schools, we decided to see if in the primary education sector, it was possible to use machine reading comprehension techniques. Many other reading comprehension datasets in English have also used school science curriculum previously (ARC, OpenBookQA,SciTail,TQA) (details in section 2.1), so our studies are also grounded in historical work. Also, in the SQuAD 2.0 (Rajpurkar et al., 2018) and previous versions, a human baseline score of (EM 86.831 and F1 89.452)
was provided for the dataset by analyzing answers from human annotators. Since there has been no dataset collected and no study on human baseline score for reading comprehension tasks in Bangladesh historically, we decided to compare our model performance with humans by experimenting on children.

7.1. Questionnaire design and experiment

Annotated survey questions and passages (Figure 10) were given to children for answering and same passages was given to the model for prediction. We compare the performance of Bangladeshi children for Bengali QA reading comprehension tasks against our model in this section.

To compare the results of our model with the question answer ability of real human beings we surveyed children of grade three and four from Cambrian College, Dhaka. We designed our questionnaire from the National Curriculum and Textbook Board (NCTB) of grade three and grade four Bengali textbooks. From the textbooks randomly three to four-line length passages were first selected. Then we annotated questions and answers from those passages. Each passage had two questions with answers and one question which is impossible to answer from the passages. A sample of these from grade 3 questionnaire is given below:

We made the following design for experiments. In SQuAD 1.1 and 2.0 (Rajpurkar et al., 2016, 2018) to establish human performance each question was annotated by three separate human annotators. Then the middle answer was considered as ground truth and other answers were considered as predictions. Since in the child survey the children are not annotating the answers, but only answering questions, we ensured each passage was given to at least three children. If we had given each question to only one child and for any reason the child would be unable to answer the question, then we wouldn’t be able to compare performance for those questions. From the answers we gather for each question, we pick the most frequent answer as a human prediction.

We collected 40 passages from grade three and four textbooks. Grade three and grade four students were surveyed separately. 20 passages were from grade three textbook and 20 passages were from grade four textbook. Each passage contained three questions and we had in total $40 \times 3 = 120$ question answers. Out of these 40 questions were impossible to answer. We collected around 200 student responses for these questions. We found that children often were unable to answer the questions and, in some cases, declined to do the survey as they found the survey out of their reading skill level.

To compare with model performances as we had annotated the ground truth answers by ourselves previously, we took the annotated dataset and used SQuAD 2.0 evaluation script to compare our results.

7.2. Survey results

We found that grade three students achieved an EM and F1 score of 71.66 and 80.52 respectively. For grade four questionnaire children baseline results were 30 and F1 score of 54 respectively. We also merged both of the results and the total EM and F1 was 50.83 and 67.32 for 120 questions. For the questions students did not make any
answer we treated the ‘’ empty string as an answer. Average answer length for grade 3 questions was 17.61 and grade 4 questions were 27.56.

For each question, we maintain question id, text and 3 answers (Figure 11). Final answer for a question is collected by majority vote. Compared to that results from our zero shot and fine-tuned models are given in Table 8 for all survey questions.

In Table 9, the scores are shown for grade 3 and grade 4 students separately.

We found that the DistilBERT fine-tuned model performed best on this survey dataset. The reason is possibly that the BERT model is computationally costly and we should have trained the BERT model further to gain significant results and RoBERTa model is not pre-trained in Bangla. But DistilBERT is both fast to train (compared to the other two), and has fewer parameters. We found grade three questions performed worse than grade four questions in general, perhaps related to question difficulty. Contrary to that all models underperformed in the grade 3 dataset, while children were able to answer the questions quite well. The teachers mentioned one of the grade 3 classrooms we surveyed were quite gifted, perhaps this has impacted the results.

8. Results and Discussion

We found zero-shot model results were consistent with previous research on other languages but the fine-tuned models on larger translated datasets underperformed. We believe the reason is two: Our dataset has a 50:50 distribution of answerable and unanswerable questions while the original SQuAD 2.0 dataset has 75:25 distribution of answerable and unanswerable questions, and the second reason is we have not trained the models long enough for the large dataset. Since transformer models are very resource consuming perhaps, we should train our models for longer, 3–5 days long for 15–20 epochs instead of only 5 epochs.

Also, our data is translated with Google Translate while the original SQuAD 2.0 is using human annotated high-quality data. If we collect more data like the original dataset and annotate them by hand it’s possible the performance will increase much more. Figure 12 contains which Bengali interrogatives showed up in our dataset frequently.

Samples from Question-Answer pairs with predicted answers and actual answers are given below (Figure 13). Since the data quality is not as good as human annotated data it is possible even humans would have trouble comprehending answers in some cases.

In the case of incorrect answers (Figure 14) there are two cases. There can be predicted answers that have some portion of the correct answers, but not an exact match and there can also be predicted answers that are far away from the correct answer. We show examples of both cases in the following samples.

We have also tried to peek inside what the model sees by using a visualization library made for BERT called BERTviz (Vig, 2019) which is a tool for visualizing transformer models. In Figure 15 we check word piece tokenization for a sample Bengali question and how it considers all the other word pieces in the sentence when forming the representation for each word to be used in downstream question answering tasks.

A demo of our work can be seen in the following prototype deployed model represented in Figure 16. We also want to Curate more bangla QA datasets for reasoning/
synthesis, multiple choice, visual question answering datasets, experimenting with different models of retriever and reader models and creating question generation models.

9. Future work

Information processing by machines from reading comprehension is a task of great potential. We have only begun the Bengali language studies, experimenting with a few language models. To run our experiments and evaluate the results, we plan to explore newer and more advanced language models. With good performance, we can transform this into a deployable web interface ready for development. Demand for this form of product is growing in the education sector, in the healthcare industry, telecom industry,
and many more. Call centres that rely solely on data workers will profit primarily from such an application, which can respond to their requests much more precisely than anything like a web search. We can also make our reader model to expand to an open domain question answering system. The reading comprehension task assumes there is a passage given already and we simply have to retrieve the answer from the passage, but the more realistic open domain QA systems retrieve information from a set of large unstructured collection of documents where we first have to retrieve the relevant passages/documents to be able to find the answer. Figure 17 demonstrates one such pipeline.

In future, we would also like to scale up our models and datasets, try different models like ALBERT, XLM-ROBERTA, expand our work to cross lingual question answering where the context passage and the question can be of different languages. We expect after improving the quality of these baseline models more we would be able to publish our models and dataset as open source benchmark models and datasets that other industry users or research projects would be able to finetune further on their specific domain. For example, a healthcare company may choose to create a healthcare question answering chatbot based on their data and answer frequently asked questions. Government users can refine the system for creating legal chatbots and telecom companies can also create similar bots. Also, we can release the finetuned models embedding vectors as word vectors because they can be reused as components of other models. Word2Vec, FastText and many other language models are released as open source word vectors to be reused as multipurpose components of other models. HuggingFace has a recently developed model publishing platform, and we expect to publish our QA models there as first benchmark Bangla QA models soon.

10. Conclusion

Question Answering, formulated as a problem of information retrieval, is an important section of NLP with the potential to be utilized in numerous tasks of high value. Given a context passage, a model is trained to be able to answer questions from the given passage. The recent rise in powerful language models like BERT and its variants has made it possible for all kinds of language processing tasks to make great progress. Bengali is still an unexplored language in NLP, so much so that our current work is the first work to successfully apply the transformer model to Bengali Question Answering and setting up a benchmark. We demonstrate that in the initial stages of the prototyping of a data-heavy deep learning model such as transformers, even a translation-based data collection approach can also be used. This opens up a wider use of the Bengali language
in the NLP since the lack of accessible Bengali data has been one of the key concerns for researchers. The BERT embeddings trained in the context of Question Answering with translated SQuAD can be used elsewhere and our model can also be finetuned for newly collected corpus for specific domains such as telecoms, healthcare, and e-commerce industry chatbots. In both fine-tuned and zero-shot environments, we also address a comprehensive study of BERT and its variant models and how they work with a low-resource language such as Bengali. Based on common topics related to Bengali culture, we also curated a human annotated QA dataset on Bengali and finally compared our model output with that of Bengali children to set a benchmark for future work.

**Disclosure statement**

No potential conflict of interest was reported by the author(s).

**Notes on contributors**

*Tasmiah Tahsin Mayeesha* will be completing BSc. in Computer Science and Engineering from Electrical and Computer Engineering Department, North South University. Previously she has worked in Tensorflow, Google and Berkman Klein Center of Internet and Society in Harvard University as a Google Summer of Code participant, and also in Cramstack as a junior data scientist. Her research interests are in NLP and deep learning. She has previously published in CSCW workshop, solidarity across borders in 2018.

*Abdullah Md. Sarwar* is currently working as a software engineer at Early Advantage. He will be completing his BSc. in Computer Science and Engineering from North South University. His research interests include Natural Language Processing, Data Mining and Data Science.

*Rashedur M. Rahman* is working as a Professor in Electrical and Computer Engineering Department, North South University, Dhaka, Bangladesh. He received his PhD in Computer Science from University of Calgary, Canada. He has authored more than 150 peer-reviewed research papers. His current research interests are in data science, cloud load characterization, optimization of cloud resource placements, deep learning, etc.

**References**

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., & , Parikh, D. (2015). VQA: Visual question answering. In Proceedings of the IEEE International Journal of Computer Vision, 123, 4–31. https://doi.org/10.1109/ICCV.2015.279

Artetxe, M., Ruder, S., & Yogatama, D. (2019). On the cross-lingual transferability of monolingual representations. arXiv preprint arXiv:1910.11856.

Bajgar, O., Kadlec, R., & Kleindienst, J. (2016). Embracing data abundance: Booktest dataset for reading comprehension. arXiv preprint arXiv:1610.00956, 2016.

Banerjee, S., & Bandyopadhyay, S. (Dec. 2012). Bengali Question Classification: Towards Developing QA System.

Banerjee, S., Naskar, S., Bandyopadhyay, S. (2014). BFQA: A Bengali Factoid question answering system. In P. Sojka (Ed.), Text, Speech and Dialogue (pp. 217–224). Springer International.

Baradaran, R., & Ghiasi, R., & Amirkhani, H. (2020). A survey on machine reading comprehension systems. arXiv preprint arXiv:2001.01582.

Baudiš, P., & Šedivý, J. (September, 2015). Modeling of the question answering task in the yodaqa system. International Conference of the cross-language evaluation Forum for European languages (pp. 222–228). Springer.
Boleda, G., Paperno, D., Kruszewski, G., Lazoridou, A., Pham, Q. N., Bernardi, R., Pezzelle, S., Baroni, M., & Fernandez, R. (2016). The lambada dataset: Word prediction requiring a broad discourse context. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long papers, pp. 1525–1534). Association for Computational Linguistics (ACL).

Carrino, C. P., Costa-jussà, M. R., & Fonollosa, J. A. (2019). Automatic Spanish translation of the SQuAD dataset for multilingual question answering. arXiv preprint arXiv:1912.05200.

Chen, D., Fisch, A., Weston, J., & Bordes, A. (2017). Reading wikipedia to answer open-domain questions. arXiv preprint arXiv:1704.00051.

Choi, E., He, H., Iyyer, M., Yatskar, M., Yih, W.-t., Choi, Y., Liang, P., & Zettlemoyer, L. (2015). QuAC: Question answering in context. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 2174–2184). Brussels, Belgium, October–November 2018. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1241

Clark, J. H., Choi, E., Collins, M., Garrette, D., Kwiatkowski, T., Nikolaev, V., & Palomaki, J. (2020). TyDiQA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages, TACL 2020.

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Bowman, S. R., Schwenk, H., & Stoyanov, V. (2018). Think you have solved question answering? try arc, the ai2 reasoning challenge. ArXiv, abs/1803.05457.

Clark, C., & Gardner, M. (2017). Simple and effective multi-paragraph reading comprehension. arXiv preprint arXiv:1710.10723.

Conneau, A., Lample, G., Rinott, R., Williams, A., Bowman, S. R., Schwenk, H., & Stoyanov, V. (2018). XNLI: Evaluating cross-lingual sentence representations. arXiv preprint arXiv:1809.05053.

Cui, Y., Liu, T., Che, W., Xiao, L., Chen, Z., Ma, W., & Wang, S. (2018). A span-extraction dataset for chinese machine reading comprehension. arXiv preprint arXiv:1810.07366.

Dalvi, B., Huang, L., Tandon, N., Yih, W.-t., & Clark, P. (2018). Tracking state changes in procedural text: A challenge dataset and models for process paragraph comprehension. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (long papers). pp. 1595–1604, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1144

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Dhingra, B., Mazaitis, K., & Cohen, W. W. (2017). Quasar: Datasets for question answering by search and reading. ArXiv, abs/1707.03904.

d’Hoffschmidt, M., Vidal, M., Belblidia, W., & Brendlé, T. (2020). FQuAD: French question answering dataset. arXiv preprint arXiv:2002.06071.

Dua, D., Wang, Y., Dasigi, P., Stanovsky, G., Singh, S., & Gardner, M. (2019). DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In Proc. of NAACL.

Dunn, M., Sagun, L., Higgins, M., Guney, V. U., Cirik, V., & Cho, K. (2017). Searchqa: A new q&a dataset augmented with context from a search engine. arXiv preprint arXiv:1704.05179.

Efimov, P., Chertok, A., Boytsov, L., & Braslavski, P. (2020). SberQuAD - Russian reading comprehension dataset: Description and analysis. Experimental IR Meets multilinguality, multimodality, and interaction. CLEF 2020 (Vol. 12260). Springer. https://doi.org/10.1007/978-3-030-58219-7_1

Grail, Q., & Perez, J. (2018). Reviewqa: a relational aspect-based opinion reading dataset. arXiv preprint arXiv:1810.12196.

Gupta, D., Ekbal, A., & Bhattacharyya, P. (2019). A deep neural network Framework for English Hindi question answering. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 19(2), Article 25 (November 2019), pp.1–22. https://doi.org/10.1145/3359988

Gupta, S., & Rawat, B. P. S. (2020). Conversational machine comprehension: a literature review. arXiv preprint arXiv:2006.00671.

Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., & Blunsom, P. (2015). Teaching machines to read and comprehend. In Advances in neural information processing systems (pp. 1693–1701).
Hewlett, D., Lacoste, A., Jones, L., Polosukhin, I., Fandrianto, A., Han, J., Kelcey, M., & Berthelot, D. (2016). WikiReading: A novel large-scale language understanding task over Wikipedia. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long papers), pp. 1535–1545, Berlin, Germany, August 2016. Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1145

Hill, F., Bordes, A., Chopra, S., & Weston, J. (2016). The goldilocks principle: Reading children’s books with explicit memory representations. In Y. Bengio, & Y. LeCun (Eds.), 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2–4, 2016, Conference Track Proceedings.

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

Hirschman, L., Light, M., Breck, E., & Burger, J. D. (1999, June). Deep read: A reading comprehension system. In Proceedings of the 37th annual meeting of the Association for Computational Linguistics (pp. 325–332).

Hoque, S., Arefin, M. S., & Hoque, M. M. (2015). BQAS: A bilingual question answering system. 2015 2nd international conference on electrical information and communication technologies (EICT) (pp. 586–591). Khulna.

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. arXiv preprint arXiv:1801.06146.

Hsu, T. Y., Liu, C. L., & Lee, H. Y. (2019). Zero-shot reading comprehension by cross-lingual transfer learning with multi-lingual language representation model. arXiv preprint arXiv:1909.09587.

Islam, S. T., & Nurul Huda, M. (2019). BQAS: A bilingual question answering system. Transactions of the Association for Computational Linguistics, 6, 317–328.

Kembhavi, A., Seo, M., Schwenk, D., Choi, J., Farhadi, A., & Hajishirzi, H. (2017). Are you smarter than a sixth grader? Textbook question answering for multimodal machine comprehension. Proceedings of the IEEE Conference on Computer Vision and Pattern recognition (pp. 4999–5007).

Khashabi, D., Chaturvedi, S., Roth, M., Upadhyay, S., & Roth, D. (2018, June). Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies. Vol. 1 (long papers, pp. 252–262).

Khot, T., Sabharwal, A., & Clark, P. (2018). SciTail: A textual entailment dataset from science question answering. Thirty-Second AAAI Conference on Artificial Intelligence.

Kočiský, T., Schwarz, J., Blunsom, P., Dyer, C., Hermann, K. M., Melis, G., & Grefenstette, E. (2018). The narrativeqa reading comprehension challenge. Transactions of the Association for Computational Linguistics, 6, 317–328.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Kelcey, M., Chang, M., Dai, A., Uszkoreit, J., Le, Q., & Petrov, S. (2019). Natural questions: A benchmark for question answering research. Transactions of the Association for Computational Linguistics, 7, 453–466. https://doi.org/10.1162/tacl_a_00276

Lai, G., Xie, Q., Liu, H., Yang, Y., & Hovy, E. (2017). RACE: Large-scale ReAding comprehension dataset from examinations. Proceedings of the 2017 Conference on Empirical methods in Natural language processing, pp. 785–794, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1082

Lample, G., & Conneau, A. (2019). Cross-lingual language model pretraining. arXiv preprint arXiv:1901.07291.
Lee, K., Yoon, K., Park, S., & Hwang, S. (2018). Semi-supervised Training Data Generation for Multilingual Question Answering. LREC.

Lewis, P., Oğuz, B., Rinott, R., Riedel, S., & Schwenk, H. (2019). Mlqa: Evaluating cross-lingual extractive question answering. arXiv preprint arXiv:1910.07475.

Lim, S., Kim, M., & Lee, J. Korquad1.0: Korean qa dataset for machine reading comprehension. arXiv preprint arXiv:1909.07005.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.

Mihaylov, T., Clark, P., Khot, T., & Sabharwal, A. (2018). Can a suit of armor conduct electricity? A new dataset for open book question answering. In EMNLP.

Miller, A. H., Fisch, A., Dodge, J., Karimi, A.-H., Bordes, A., & Weston, J. (2016). Key-value memory networks for directly reading documents (emnlp16).

Mozannar, H., Hajal, K. E., Maamary, E., & Hajj, H. (2019). Neural arabic question answering. arXiv preprint arXiv:1906.05394.

Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., & Deng, L. (2016). Ms marco: A human-generated machine reading comprehension dataset.

Nir, S. M. H., Nayeem, M. K., & Islam, M. S. (2017). Question classification using support vector machine with hybrid feature extraction method. 2017 20th International Conference of Computer and Information Technology (ICCIT), Dhaka (pp. 1–6).

Onishi, T., Wang, H., Bansal, M., Gimpel, K., & McAllester, D. (2016). Who did what: A large-scale person-centered cloze dataset. Proceedings of the 2016 Conference on Empirical methods in Natural language processing (pp. 2230–2235), Austin, Texas, November 2016. Association for Computational Linguistics. https://doi.org/10.18653/v1/D16-1241

Ostermann, S., Modi, A., Roth, M., Thater, S., & Pinkal, M. (2018). MCScript: A novel dataset for assessing machine comprehension using script knowledge. Proceedings of the Eleventh International Conference on language resources and evaluation (LREC 2018), Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

Park, D., Choi, Y., Kim, D., Yu, M., Kim, S., & Kang, J. (2019). Can machines learn to comprehend scientific literature? IEEE Access, 7, 16246–16256. https://doi.org/10.1109/ACCESS.2019.2891666

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. arXiv preprint arXiv:1802.05365.

Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is multilingual BERT. arXiv preprint arXiv:1906.01502.

Rajpurkar, P., Jia, R., & Liang, P. (2018). Know What You Don’t Know: Unanswerable Questions for SQuAD. In: CoRR abs/1806.03822 (2018). arXiv: 1806.03822. http://arxiv.org/abs/1806.03822

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250.

Reddy, S., Chen, D., & Manning, C. D. (2018). CoQA: A conversational question answering challenge. In: CoRR abs/1808.07042 (2018). arXiv: 1808.07042. http://arxiv.org/abs/1808.07042

Richardson, M., Burges, C. J., and Renshaw, E. 2013. MCTest: A challenge dataset for the open-domain machine comprehension of text. In Proceedings of the 2013 Conference on Empirical methods in Natural language processing (pp. 193–203). Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

Saeidi, M., Bartolo, M., Lewis, P., Singh, S., Rocktäschel, T., Sheldon, M., Bouchard, G., & Riedel, S. (2018). Interpretation of natural language rules in conversational machine reading. Proceedings of the 2018 Conference on Empirical methods in Natural language processing. pp. 2087–2097, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1233

Saha, A., Aralikatte, R., Khapra, M. M., & Sankaranarayan, K. (2018). DuoRC: Towards complex language understanding with paraphrased reading comprehension. Proceedings of the 56th Annual Meeting of the Association for computational Linguistics (Volume 1: Long papers), pp. 1683–1693, Melbourne, Australia, July 2018. Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1156
Sanh, V., Debut, L., Chaumond, J., Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In: arXiv preprint arXiv:1910.01108.

Seo, M., Kembhavi, A., Farhadi, A., & Hajishirzi, H. (2016). Bidirectional attention flow for machine comprehension. arXiv preprint arXiv:1611.01603.

Shao, C. C., Liu, T., Lai, Y., Tseng, Y., & Tsai, S. (2018). Drqa: a chinese machine reading comprehension dataset. arXiv preprint arXiv:1806.00920.

Siblini, W., Pasqual, C., Lavielle, A, & Cauchois, C. (2019). Multilingual question answering from formatted text applied to conversational agents. arXiv preprint arXiv:1612.04342.

Sun, K., Yu, D., Chen, J., Yu, D., Choi, Y., & Cardie, C. (2019). Dream: A challenge data set and models for dialogue-based reading comprehension. Transactions of the Association for Computational Linguistics, 7, 217–231. https://dataset.org/dream/

Šuster, S., & Daelemans, W. (2018). CliCR: A dataset of clinical case reports for machine reading comprehension. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human language Technologies, Volume 1 (long papers, pp. 1551–1563). New Orleans, Louisiana, June 2018. Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1140

Talmor, A., Herzig, J., Lourie, N., & Berant, J. (2019). CommonsenseQA: A question answering challenge targeting commonsense knowledge. Proceedings of the 2019 Conference of the North American Chapter of the Association for computational Linguistics: Human language Technologies, Volume 1 (long and short papers), pp. 4149–4158). Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1421

Trischler, A., Wang, T., Yuan, X., Harris, J., Sordoni, A., Bachman, P., & Suleman, K. (2017). NewsQA: A machine comprehension dataset. Proceedings of the 2nd Workshop on representation learning for NLP, pp. 191–200, Vancouver, Canada, August 2017. Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-2623

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 5998–6008.

Vig, J. A multiscale visualization of attention in the transformer Model. In: arXiv preprint arXiv:1906.05714. https://arxiv.org/abs/1906.05714

Wang, Alex, Pruksachatkun, Yada, Nangia, Nikita, Singh, Amanpreet, Michael, Julian, Hill, Felix, Levy, Omer, & Bowman, Samuel. (2019). SuperSquad: A stickier benchmark for general-purpose language understanding systems. Advances in Neural Information Processing Systems, 3266–3280

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). Glue: A multi-task benchmark and analysis platform for natural language understanding. arXiv preprint arXiv:1804.07461.

Welbl, J., Liu, N. F., & Gardner, M. (2017). Crowdsourcing multiple choice science questions. Proceedings of the 3rd Workshop on noisy User-generated Text, pp. 94–106, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-4413

Welbl, J., Stenetorp, P., & Riedel, S. (2018). Constructing datasets for multi-hop reading comprehension across documents. Transactions of the Association for Computational Linguistics, 6, 287–302.

Wu, S., & Dredze, M. (2019). Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. arXiv preprint arXiv:1904.09077.

Xie, Q., Lai, G., Dai, Z., & Hovy, E. (2018). Large-scale cloze test dataset created by teachers. Proceedings of the 2018 Conference on Empirical methods in Natural language processing (pp. 2344–2356). Brussels, Belgium, October–November 2018. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1257
Yagcioglu, S., Erdem, A., Erdem, E., & Ikizler-Cinbis, N. (2018). RecipeQA: A challenge dataset for multi-modal comprehension of cooking recipes. Proceedings of the 2018 Conference on Empirical methods in Natural language processing, pp. 1358–1368, Brussels, Belgium, October–November 2018. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1166

Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W. W., Salakhutdinov, R., & Manning, C. D. (2018). Hotpotqa: A dataset for diverse, explainable multi-hop question answering. arXiv preprint arXiv:1809.09600.

Yang, Y., Yih, W.-t., & Meek, C. (2015). Wikiqa: A challenge dataset for open-domain question answering. Proceedings of the 2015 Conference on Empirical methods in Natural language processing (pp. 2013–2018).

Yu, A. W., Dohan, D., Luong, M. T., Zhao, R., Chen, K., Norouzi, M., & Le, Q. V. (2018). Qanet: Combining local convolution with global self-attention for reading comprehension. arXiv preprint arXiv:1804.09541.

Zeng, C., Li, S., Li, Q., Hu, J., Hu, J. (2020). A survey on machine reading comprehension: Tasks, evaluation metrics, and benchmark datasets. arXiv preprint arXiv:2006.11880.

Zhang, S., Liu, X., Liu, J., Gao, J., Duh, K., & Van Durme, B. (2018). Record: Bridging the gap between human and machine commonsense reading comprehension. arXiv preprint arXiv:1810.12885.

Zhang, X., Yang, A., Li, S., & Wang, Y. (2019). Machine reading comprehension: a literature review. arXiv preprint arXiv:1907.01686.