SQuAD: 100,000+ Questions for Machine Comprehension of Text

Alireza Moradi

The article is introducing a QA dataset for the Reading Comprehension task which is pretty challenging for machines. The dataset contains more than 100,000 question-answer pairs on 536 articles with 20% baseline F1 score. Human performance in this dataset is 86.8%. One of the challenges of this dataset is that the model should understand a passage and provide an answer based on that (*span*?). Then some existing datasets in this field are named which are 3 types:

- **Reading Comprehension**
- **Open-Domain Question Answering**: which the goal is to answer a question from a large collection of documents.
- **Cloze Datasets**: which the goal is to predict the missing word (or entity) in a passage (unlike SQuAD in which the answers are often much longer and include non-entities)

Then it gets to collection of dataset which is mainly 3 stages:

- **Passage Curation**: retrieving *wikipedia*'s top 10000 English articles based on page ranks. 536 articles were sampled resulting in 23,215 paragraphs.
- **QA Collection**: where crowdworkers (filtered through a certain criteria) were employed to ask and answer up to 5 questions from each paragraph.
- **Additional Answers Collection**: To get better evaluations and get an overall of human performance on the dataset.

In the 4th part three subjects were analysed in the dataset:

- **Diversity in Answers**: the answers in the dataset were automatically categorized as below:

    - Numbers
    - Proper Noun Phrases
    - Common Noun Phrases
    - Adjective Phrases, Verb Phrases, Clauses etc.
- **Reasoning Required to Answer Questions**: which shows that all examples have lexical/syntactic divergence between the question and the answer in the passage.
- **Stratification by Syntactic Divergence**: the syntactic divergence between a question and the sentence containing the answer was quantized, which provides a way to measure the difficulty of the question. (I didn't exactly understand how they measure this.)

In the 5th part the accuracy of the developed model (logistic regression) was compared to the 3 baselines.

Finally the model was evaluated using 2 metrics: accuracy and F1 score. Human performance was also evaluated on the dataset and the results were 77% accuracy and 86.8% F1 were the mismatches were not important because they were in non-essential parts of the phrase.

While evaluating the model performance it was seen that the sentence with the correct answer was chosen with almost 80% accuracy, So the main difficulty for the model was finding the right answer in those sentences. Also in order to find out the performance of the model on different features one group of them were removed from the model at a time. The results were also reported based on the type of the answer. At last, the results were stratified by syntactic divergence which showed that more divergence equals lower performance. (unlike human performance which suffered the least amount of performance loss in this case)