Question Answering on SQuAD dataset

SQuAD: The Stanford Question Answering Dataset

100,000+ question-answer pairs on 500+ articles.

Objective: to create an NLP system that, given a paragraph and a question regarding it, provides a single answer, which is obtained selecting a span of text from the paragraph.

SQuAD data format:

List of articles from Wikipedia. Each article is divided in paragraphs. For each paragraph, there is a set of questions (“qas”).

Each question has an id, the text of the question (“question”), and a list of possible answers. Some question have only one answer, other may have multiple correct answers. Each answer is characterized by the text of the answer itself and by the offset where the span starts.

Example of data:

{"data": [{"title": "University_of_Notre_Dame", "paragraphs": [{"context": "Architecturally, the school has a Catholic character. Atop the Main Building's gold dome is a golden statue of the Virgin Mary. Immediately in front of the Main Building and facing it, is a copper statue of Christ with arms upraised with the legend \"Venite Ad Me Omnes\". Next to the Main Building is the Basilica of the Sacred Heart. Immediately behind the basilica is the Grotto, a Marian place of prayer and reflection. It is a replica of the grotto at Lourdes, France where the Virgin Mary reputedly appeared to Saint Bernadette Soubirous in 1858. At the end of the main drive (and in a direct line that connects through 3 statues and the Gold Dome), is a simple, modern stone statue of Mary.", "qas": [{"answers": [{"answer_start": 515, "text": "Saint Bernadette Soubirous"}], "question": "To whom did the Virgin Mary allegedly appear in 1858 in Lourdes France?", "id": "5733be284776f41900661182"}, {"answers": [{"answer_start": 188, "text": "a copper statue of Christ"}], "question": "What is in front of the Notre Dame Main Building?", "id": "5733be284776f4190066117f"}, {"answers": [{"answer_start": 279, "text": "the Main Building"}], "question": "The Basilica of the Sacred heart at Notre Dame is beside to which structure?", "id": "5733be284776f41900661180"}, {"answers": [{"answer_start": 381, "text": "a Marian place of prayer and reflection"}], "question": "What is the Grotto at Notre Dame?", "id": "5733be284776f41900661181"}, {"answers": [{"answer_start": 92, "text": "a golden statue of the Virgin Mary"}], "question": "What sits on top of the Main Building at Notre Dame?", "id": "5733be284776f4190066117e"}]} …

Submission format (code): a python script called “compute_answers” that given a json file formatted as the training set, creates a prediction file in the desired format. Obviously the given json file will not contain the answers, only the contexts and the questions.

```
python3 compute_answers.py *path_to_json_file*
```

Provided files:

Training set (29 MB), Evaluation script, Example of prediction file

Performance evaluation method: the system will be evaluated using the attached script, that you can use to test your own method, on a test set that is not available to you.

```
python3 evaluate.py *path_to_ground_truth* *path_to_predictions_file*
```

Evaluation criteria are published in virtuale.

For training and validation purpose, you are not allowed to use other data other than the training set. This means also that you can't use any other dataset to train the models.

If you split the dataset in training and validation, we suggest you to do the splitting based on the title: all the questions/paragraphs regarding the same title should be in the same split.

You can approach the problem with any NLP technique and rely on any form of external background knowledge.

As a reference, be aware that the human performance on this task has a F1 around 0.90