Assignment 3: Question Answering via Reading Comprehension

CS525: Natural Language Processing
Thabsheer Jafer Machingal

Dataset

Stanford Question Answering Dataset (SQuAD)[1] is a reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or span, from the corresponding reading passage, or the question might be answerable.

SQuAD1.1

The dataset has two versions. The original, first one is SQuAD v1.1, which contains a training and a validation set. Each question in the dataset has an id with references to it. The training set has 87,599 questions in total, similarly, there are 10,570 questions in the validation set.

All the questions in this dataset, unlike SQuAD2.0, are answerable and have been provided with an answer to it. On an average, each context has 4.63 questions in the training and 5.11 in the validation dataset.

SQuAD2.0 (SQuADRUn)

SQUADRUN, a new dataset that combines the existing Stanford Question Answering Dataset (SQuAD) with over 50,000 unanswerable questions written adversarially by crowdworkers to look similar to answerable ones.

Task 1 - Implement a question-answering model

Model

DistilBertForQuestionAnswering with fine-tuning

Result

Accuracy metric: The starting and end position of the ground truth and predicted answer is compared and a binary label is given to the prediction(either True or False).

SQuAD1.1

Using the SQuAD1.1 dataset and the same network as below, I got an F1 score of 0.64. The F1 score of version 1,1 is higher and is expected as there is no unanswerable question in the first version.

SQuAD 2.0

Using this metric and the model I got a **F1 score of 0.54**

	precision	recall	f1-score	support
2	0.81	0.81	0.81	95
3	0.70	0.69	0.70	91
4	0.57	0.57	0.57	138
5	0.69	0.77	0.73	131
6	0.62	0.55	0.58	140
7	0.53	0.56	0.55	96
8	0.60	0.57	0.58	93
9	0.62	0.53	0.57	100
10	0.62	0.63	0.62	95
11	0.56	0.72	0.63	78
12	0.68	0.60	0.64	92
13	0.69	0.54	0.60	69
14	0.65	0.57	0.61	100
15	0.63	0.63	0.63	71
16	0.68	0.63	0.65	133
17	0.70	0.62	0.65	104
18	0.51	0.56	0.53	82
19	0.53	0.54	0.53	65
20	0.70	0.60	0.64	112
21	0.63	0.52	0.57	73
22	0.61	0.68	0.64	117
23	0.57	0.68	0.62	78
24	0.52	0.53	0.53	85
...				
accuracy			0.54	20302
macro avg	0.44	0.45	0.43	20302
weighted avg	0.55	0.54	0.54	20302

Task 2 - Unanswerable question detection

Identifying unanswerable questions in a Reading comprehension problem is surely a challenging problem. As mentioned in this paper[2], creating rule based questions for the dataset would have been an easy problem for a deep learning model to solve. However, SQuADRUn has questions that does not follow a specific structure (simulating natural language).
There have been many attempts in the recent past year to solve the problem, most of them did not achieve accuracy comparable to human level performance.

Hypothesis

This solution could be language specific, the idea is to understand the relevance of words in a question with respect to the passage given (IDF) and identify a key text with respect to the passage in the question that determines whether the question is answerable.

Motivation

The following data is from [3] and is used to explain the motivation behind the proposed model.

Article: Endangered Species Act

Paragraph: " …Other legislation followed, including the Migratory Bird Conservation Act of 1929, a 1937 treaty prohibiting the hunting of right and gray whales, and the Bald Eagle Protection Act of 1940. These later laws had a low cost to society—the species were relatively rare—and little opposition was raised."

Question 1: "Which laws faced significant opposition?"

Plausible Answer: later laws

Question 2: "What was the name of the 1937 treaty?"

Plausible Answer: Bald Eagle Protection Act

Reading comprehension is fairly an easy task for a human reader, we can answer the above question or we can tell if the questions are unanswerable. Both question1 and question2 above are unanswerable and given a wrong answer, our task is to develop a model that can classify answerable questions from unanswerable questions. A question can be unanswerable due to various reasons, we cannot simply look for matching words (just TFIDF). Although that might help in some cases. An example is given below.

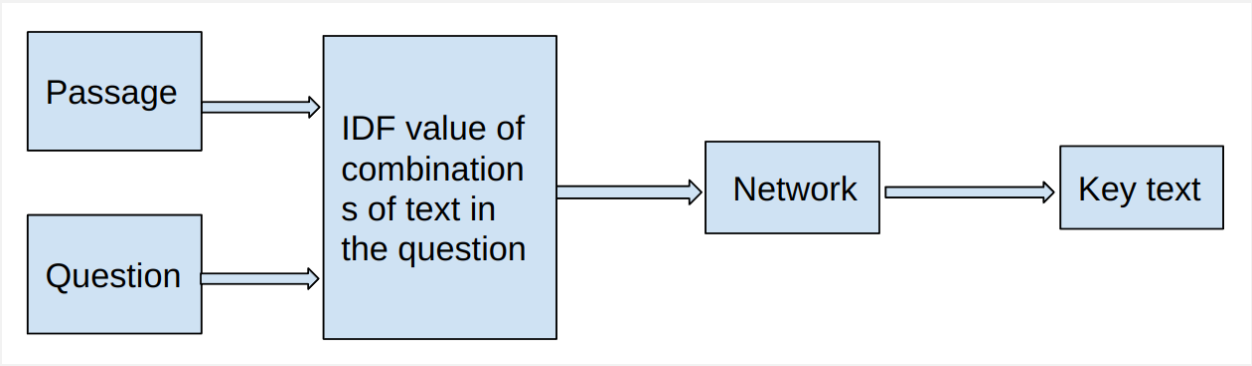
Question 3 : What was the name of the 2020 treaty?

In the SQuADRUn paper[2], this problem is referred to as the RULEBASED problem. Most real world reading comprehension tasks involve critical analysis or summarization of the passage, this would make the problem difficult to solve. Consider the '**Question 2'**, this is unanswerable, because the name of the treaty is not mentioned in the passage, making **"Name of the 1937 treaty"** non answerable text in the passage(let's call it "***key text***"). My idea is to search for the IDF value of the *key tex*t in the passage. Use a neural network to train on the SQuAD2.0 to learn this key text from a question, given the passage to a question.

Model Design

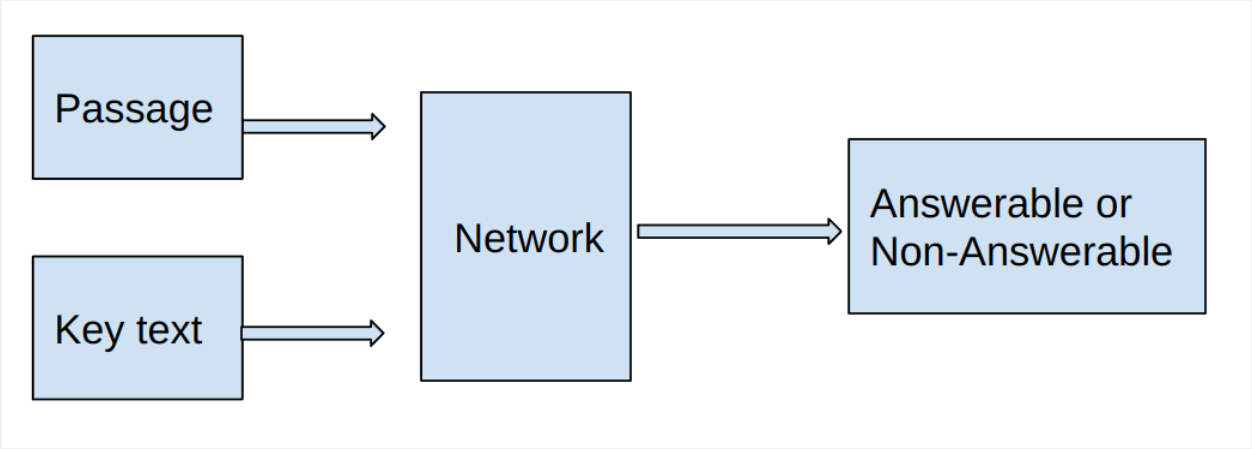
Network 1

Using the passage and question find the IDF score for a combination of texts in the question. Find the highly relevant text(group of words) in the question and use the network to learn it from the SQuAD2.0, although this might require modifying the dataset(for the ground truth).



Now that we have a network that can identify a *key tex*t in any given question we can use the text to search if the passage can find the answer to the *key text.*

Network 2



Using the key text we can find the embedding of the text and that of the passage and look if the text exists in the passage.

Discussion

This approach might not be scalable to problems involving summarization questions. Also the synonyms of words might trouble the model with False positive answers.

Learning of the two networks might be costly and also might need to modify the dataset(unless we can identify the key text from the question without the network 1). It would also be interesting to learn a different parameter instead of IDF.

It would also be interesting if we can generate key text for a question that has words that are not in the question but very relevant to the question [4],[5] (modifying inputs to the network 1). We could essentially modify the network and generate the key text that is relevant in a question. There are numerous methods, each of with its own challenges, for text generation including employing a reinforcement learner[5].

References

[1] “b’SQuAD Dataset’,” *DeepAI*. https://deepai.org/dataset/squad (accessed Nov. 29, 2022).

[2] P. Rajpurkar, R. Jia, and P. Liang, “Know What You Don’t Know: Unanswerable Questions for SQuAD,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Melbourne, Australia, Jul. 2018, pp. 784–789. doi: 10.18653/v1/P18-2124.

[3] P. Rajpurkar, R. Jia, and P. Liang, “Know What You Don’t Know: Unanswerable Questions for SQuAD.” arXiv, Jun. 11, 2018. doi: 10.48550/arXiv.1806.03822.

[4] X. Zhang, Y. Yang, S. Yuan, D. Shen, and L. Carin, “Syntax-Infused Variational Autoencoder for Text Generation,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, 2019, pp. 2069–2078. doi: 10.18653/v1/P19-1199.

[5] G. Yasui, Y. Tsuruoka, and M. Nagata, “Using Semantic Similarity as Reward for Reinforcement Learning in Sentence Generation,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, Florence, Italy, 2019, pp. 400–406. doi: 10.18653/v1/P19-2056.