**Building dataset Based on SQUAD 2.0**

For NEF6001

**Name:** Valentine Somtochukwu Okeke

**Student ID:** 4589611

**Supervisor:** Dr Gongqi Lin

**Date:** 20/08/2019

**1. LITERATURE REVIEW**

* 1. **SQUAD 2.0 Datasets**

This research is to provide innovative solutions to build datasets based on SQUAD 2.0. A possible solution can be by applying general knowledge rule with deep learning. To do well on SQUAD 2.0, systems must not only answer questions when possible, but also determine when no answer is supported by the paragraph and abstain from answering.

However, searching IEEE explore research database with keywords “SQUAD2.0 datasets” returns no result. This illustrates that the research on SQUAD 2.0 is still very new. Researchers may have applied datasets building but did not base it on SQUAD 2.0. Given that IEEE Explore research database mainly focuses on building datasets, the articles it includes do not relate to SQUAD 2.0. Searching with keywords “dataset building” returns 7150 articles with 2,865 published in the last two years. This illustrates that the research on dataset building is very active particularly in the past two years. The following is a list of the articles:

1. Semantic Segmentation based building extraction method using multi-source GIS map datasets and satellite imagery.
2. Building a dataset for personalized learning recommendation system.
3. Yarmouk Arabic OCR dataset.
4. EMG dataset augmentation approaches for improving the multi-DOF wrist movement regression accuracy and robustness.
5. Building reference datasets to support social-bots detection.
6. Optimization tuning on the realistic cyber dataset CSE-CIC-IDS2018 using cloud computing.
7. Building a semi-supervised dataset to train journalistic relevance detection models.
8. A method to build multi-scene datasets for CNN for camera pose regression.
9. Building test speech dataset on Russian language for spoken document retrieval task
10. Building a labelled dataset for recognition of handball actions using mask RCNN and STIPS
11. Creation of datasets from open sources.
12. designing a novel dataset for non-intrusive load monitoring.
13. Building a benchmark dataset and classifiers for sentence-level findings in AP chest x-rays.

It shows that the current researches have a specific focus, which is building of dataset for a specific purpose. For instance, personalized learning activity system, non-intrusive load monitoring, for recognition of handball actions etc. None of the article had anything on SQUAD datasets.

Furthermore, a further search was carried out on Google Scholar, with keywords “SQUAD 2.0 dataset” which returns 1460 articles with 718 published in the last four years. This shows that the research on SQUAD 2.0 is very active particularly in the last four years. The following is a list of the articles:

1. Know what you don’t know: Unanswerable questions for SQuAD
2. Stochastic answer networks for SQuAD 2.0
3. Exploring Neural Net Augmentation to BERT for Question Answering on SQUAD 2.0
4. A qualitative comparison of COQA, SQUAD 2.0 and QUAC
5. Read + Verify: Machine reading comprehension with unanswerable questions.
6. A BERT baseline for the natural questions.
7. Learning to ask unanswerable questions for machine reading comprehension.
8. QA Diver: Interactive framework for diagnosing QA models
9. Unified language model pre-training for natural language understanding and generation
10. BERT: pre-training of deep bidirectional transformers for language understanding
11. MS MARCO: A human generated machine reading comprehension dataset.
12. U-Net: Machine reading comprehension with unanswerable questions.
13. Transfer learning for question answering on SQUAD
14. Sogou machine reading comprehension toolkit
15. Pair2vec: compositional word-pair embeddings for cross-sentence inference.
16. Modified VS3-NET for reading comprehension question answering with no answers
17. Contextual aware joint probability model towards question answering system
18. A simple but effective method to incorporate multi-turn context with BERT for conversational machine comprehension.
19. EQuANt (Enhanced Question Answer Network)
20. ReQA: An evaluation for end to end answer retrieval models
21. Quac: Question answering in context
22. Controlling risk of web question answering
23. SpanBERT: improving pre-training by representing and predicting spans
24. BoolQ: Exploring the surprising difficulty of natural yes / no questions.
25. RoBERTa: A Robustly Optiomized BERT Pretraining Approach.
26. Dialog state tracking: A neural reading comprehension approach.

This shows that the current research articles on Google Scholar have a similar focus, which is unanswerable questions and questions with answers datasets.

Another further search was carried out on Google Scholar, with keywords “Reading Comprehension System” which returns 16,800 articles with 1218 published in the last three years. This shows that the research on reading comprehension system is very active particularly in the last four years. The following is a list of the articles:

**1.** Adversarial examples for evaluating reading comprehension systems

**2.** Deep read: A reading comprehension system

**3.** Qanet: Combining local convolution with global self-attention for reading comprehension

**4.** Triviaga: A large scale distantly supervised challenge dataset for reading comprehension.

**5.** Reading Wikipedia to answer open-domain questions.

**6.** How reading comprehension is embodied and why that matters.

**7.** The narrative QA reading comprehension system.

**8.** Stochastic answer networks for machine reading comprehension.

**9.** Learning to ask: Neural question generation for reading comprehension.

**10.** Race: Large-scale reading comprehension dataset from examinations.

**11.** Constructive processes in prose comprehension and recall

**12.** Constructing datasets for multi-hop reading com prehension across documents

**13.** Gated self-matching networks for reading comprehension and question answering.

**14.** Dureader: a Chinese machine reading comprehension dataset from real-world applications

1**5.** Contextualized word representations for reading comprehension

**16.** Zero shot relation extraction via reading comprehension

**17.** Multi-granularity hierarchical attention fusion networks for reading comprehension and question answering

**18.** Looking beyond the surface: A challenge set for reading comprehension over multiple sentences

**19.** How much reading does reading comprehension require? A critical investigation of popular benchmarks.

**20.** Does writing system influence the associations between phonological awareness, morphological awareness and reading? A meta-analysis.

**21.** Towards human-level machine reading comprehension: Reasoning and inference with multiple strategies

**22.** Identifying where to focus on reading comprehension for neural question generation.

**23**. Web-based text structure strategy instruction improves seventh graders’ content area reading comprehension.

**24.** Subword-augmented embedding for cloze reading comprehension.

**25.** Difficulty controllable question generation for reading comprehension.

This shows that the current research articles on Google Scholar have a similar focus, which is reading comprehension systems with questions generation.

**1.2 Unanswerable Questions for SQUAD**

Stanford Question Answering Dataset (SquAD) 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 unanswerable.

Extractive reading comprehension systems can often locate the correct answer to a question in a context document, but they also tend to make unreliable guesses on questions for which the correct answer is not stated in the context. Existing datasets either focus exclusively on answerable questions or use automatically generated unanswerable questions that are easy to identify. The authors addressed these weaknesses by presenting SQUAD 2.0. The authors believed that for a system to do well on SQUAD 2.0, the system must not only answer questions when possible, but also determine when no answer is supported by the paragraph and abstain from answering.

This will be discussed further in focused reading.

* 1. **Machine Reading Comprehension with Unanswerable Questions**

Machine reading comprehension with unanswerable questions focuses on abstaining from answering when no answer can be inferred. In this article, Minghao Hu, Furu Wei, Yuxing Peng, Zhen Huan, Nan Yang, Dongsheng (2018) proposed a novel read-then-verify system that utilizes a neural reader to extract candidate answers and form no answer probabilities and also leverages an answer verifier to decide whether the predicted answer is entailed by the input snippets. The authors also introduced two auxiliary losses to help the reader better handle answer extraction as well as no answer detection.

This will be further discussed in focused reading.

* 1. **Adversarial Examples for Evaluating Reading Comprehension Systems**

Standard accuracy metrics indicate that reading comprehension systems are making rapid progress, but the extent to which these systems truly understand language remains unclear. They proposed an adversarial evaluation scheme for the Stanford Question Answering Dataset (SQuAD) in which systems are instead evaluated on adversarial-chosen inputs. This in turn will reward systems with real language understanding abilities.

This will be further discussed in focused reading of the related works.

**1.4 Learning to Ask Unanswerable Questions for Machine Reading Comprehension.**

In this work, Haichao Zhu, Li Dong, Furu Wei, Wenhui Wang, Bing Qin, Ting Liu (2019) proposed a data augmentation technique by automatically generating relevant unanswerable questions according to an answerable question paired with its corresponding paragraph that contains the answer.

The authors introduced a pair-to-sequence model for unanswerable question generation, which captures effectively the interactions between the question and the paragraph. In addition, they present a way to construct training data for their question generation models by leveraging the existing reading comprehension dataset. They further used the automatically generated unanswerable questions as a means of data augmentation on the SQuAD 2.0 dataset, yielding 1.9 absolute F1 improvement with BERT-base model and 1.7 absolute F1 improvement with BERT-large model.

This will be further discussed in focused reading.

**2. Related work**

***2.1 Unanswerable Questions for SQUAD***

In this work, Pranav Rajpurkar, Robin Jia and Percy Liang (2016) constructed SQuAD 2.0, a new dataset that combines answerable questions from the previous version of SQuAD (SQuAD 1.1) with new, unanswerable questions about the same paragraphs.

Extractive reading comprehension systems can often locate the correct answer to a question in a context document, but they also tend to make unreliable guesses on questions for which the correct answer is not stated in the context. Existing datasets either focus exclusively on answerable questions or use automatically generated unanswerable questions that are easy to identify. The authors addressed these weaknesses by presenting SQUAD 2.0. They believed that for a system to do well on SQUAD 2.0, the system must not only answer questions when possible, but also determine when no answer is supported by the paragraph and abstain from answering. Many works have developed systems that surpass human-level exact match accuracy on the SQuAD (one of the most widely-used reading comprehension benchmarks).

However, the authors believe these developed systems are still far from true language understanding. Recent analysis shows that models can do well at SQuAD by learning context and type-matching heuristics and that success of the developed models on SQuAD does not ensure robustness to distracting sentences. The main cause of these problems is SQuAD’s focus on questions for which a correct answer is guaranteed to exist in the context document. Therefore, models only need to select the span that seems most related to the question, instead of checking that the answer is actually entailed by the text.

The authors developed SQuAD 2.0, a new dataset that combines answerable questions from the previous version of SQuAD (SQuAD 1.1) with 53,775 modern, unanswerable questions about the same paragraphs.

They achieved this by first, stating their objectives for SQuAD 2.0 by posting two desiderata specific to unanswerable questions. Then they surveyed existing reading comprehension datasets with the criteria of existence of plausible answers in mind. They referred the context passage paired with an unanswerable question as negative example. They created their new dataset to satisfy both the relevance and plausible answer desiderata by employing crowd workers on the Daemo crowdsourcing platform to write unanswerable questions. The workers were asked to form up to 5 questions that were impossible to answer for each paragraph of an article while ensuring that a plausible answer is present and referencing entities in the paragraph.

The authors filtered the questions by taking out questions from workers who had trouble understanding the task. They applied the filter on both existing unanswerable question and new data from SQuAD 1.1. They combined the existing data with their new data for each split and used the same partition of articles as SQuAD 1.1 to generate, train, develop and test splits. They removed articles for which they did not collect unanswerable questions. They hired additional crowd workers in order to confirm clean dataset. The workers were told that for each question, they highlight the answer in the paragraph or mark it as unanswerable and also to expect every paragraph to have some unanswerable questions. The authors also inspected 100 randomly chosen negative examples from their development set to understand the challenges these examples present.

The authors then evaluated three existing model architectures that learn to predict the probability of unanswerable question which are: the BiDAF-No-Answer (BNA), DocumentQA No-Answer (DocQA) with ELMo and DocQA without ELMo. They first trained and tested all three models on SQuAD 2.0 and reported average exact match and F1 scores. They reported that the best model is DocQA + ELMo which achieved only 66.3 F1 on the test set, 23.2 points lower than the human accuracy of 89.5 F1.

From their analysis, they concluded a state-of-the-art model achieves only 66.3% F1 score when trained and tested on SQuAD 2.0, whereas human accuracy is 89.5% F1, a full 23.2 points higher. The same model architecture trained on SQuAD 1.1 gets 85.8% F1, only 5.4 points worse than humans. The authors went further to illustrate that the unanswerable questions are more challenging than ones created automatically, either via distant supervision or a rule-based method. In conclusion, they believe that the new datasets will encourage the development of reading comprehension systems to know what they do not know.

The research can be further improved on few aspects:

* By applying further, the rule of general knowledge on some questions. It is possible to apply the rule of general knowledge to question and answer datasets.
  1. ***Adversarial Examples for Evaluating Reading Comprehension Systems***

In this article, Robin Jia and Percy Liang (2017) re-evaluated reading comprehension systems. Standard accuracy metrics indicate that reading comprehension systems are making rapid progress, but the extent to which these systems truly understand language remains unclear. They proposed an adversarial evaluation scheme for the Stanford Question Answering Dataset (SQuAD) in which systems are instead evaluated on adversarial-chosen inputs. This in turn will reward systems with real language understanding abilities.

They did not rely on semantics-preserving perturbations, rather they created adversarial examples by adding distracting sentences to the input paragraph. They automatically generate these sentences so that they confuse models, but do not contradict the correct answer or confuse humans. For their main results, they use a simple set of rules to generate a raw distractor sentence that does not answer the question but looks related, then fix grammatical errors via crowdsourcing. While adversarially perturbed images punish model oversensitivity to imperceptible noise, their adversarial examples target model overstability (the inability of a model to distinguish a sentence that actually answers the question from one that merely has words in common with it). They believe their experiments demonstrate that no published open-source model is robust to the addition of adversarial sentences. Across sixteen such models, adding grammatical adversarial sentences reduces F1 score from an average of 75% to 36%. On a smaller set of four models, they run additional experiments in which the adversary adds nongrammatical sequences of English words, causing average F1 score to drop further to 7%. To encourage the development of new models that understand language more precisely, they have released all of their codes and data publicly.

Their method tests whether systems can answer questions about paragraphs that contain adversarial inserted sentences, which are automatically generated to distract computer systems without changing the correct answer or misleading humans. In this adversarial setting, it showed that the accuracy of sixteen published models drops from an average of 75% F1 score to 36%; when the adversary is allowed to add ungrammatical sequences of words, average accuracy on four models decreases further to 7%. With the proposed SQUAD evaluation scheme, they hope their insights will motivate the development of new models that understand language more precisely. Below is an example of how they confused the system model (BiDAF Ensemble model) by the addition of an adversarial distracting sentence (in blue) using Squad dataset.

**Article:** Super Bowl 50.

**Paragraph:** “Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver’s Executive Vice President of Football Operations and General Manager. *Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV*.”

**Added adversarial sentence:** *Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*

**Question:** “What is the name of the quarterback who was 38 in Super Bowl XXXIII?”

**System first Prediction:** John Elway

**System Prediction under adversary:** Jeff Dean

From the result, we can see the system originally got the answer correctly but after the addition of an adversarial sentence, the system got the answer wrong.

* 1. ***Machine Reading Comprehension with Unanswerable Questions***

In this article, M+inghao Hu, Furu Wei, Yuxing Peng, Zhen Huan, Nan Yang, Dongsheng (2018) proposed a novel read-then-verify system that utilizes a neural reader to extract candidate answers and form no answer probabilities and also leverages an answer verifier to decide whether the predicted answer is entailed by the input snippets. The ability to comprehend test ansd answer is really vital for natural language processing.

The authors proposed a system that will address the issue of unanswerable questions. The system aims to be robust to unanswerable questions by validating the answerability of the question and by verifying the legitimacy of the predicted answer.

Their system not only utilizes a neural reader to extract candidate answers and produce no answer probabilities, but also leverages an answer verifier to decide whether the predicted answer is entailed by the input snippets. Their system consists of two components: (1) a no-answer reader for extracting candidate answers and detecting unanswerable questions, and (2) an answer verifier for deciding whether or not the extracted candidate is legitimate.

The key contributions of their work are in three folds. Firstly, they augmented existing readers with two auxiliary losses, to better manage answer extraction and no-answer detection respectively. Since the downstream verifying stage always requires a candidate answer, the reader must be able to extract plausible answers for all questions. However, they solved the problem of former approaches which is the inability to find potential candidates for unanswerable questions by introducing an independent span loss that aims to focus on the answer extraction task regardless of the answerability of the question. To avoid conflict with no answer detection, they leveraged a multi-head pointer network to generate two pairs of span scores, where one pair is normalized with the no-answer score and the other is used for our auxiliary loss. Moreover, they presented another independent no answer loss to further alleviate the confliction, by concentrating on the no-answer detection task without considering the shared normalization of answer extraction.

Secondly, in addition to the standard reading phase, they introduced an extra answer verifying phase, which focuses on finding local entailment that supports the answer by comparing the answer sentence with the question. This is based on the observation that the core phenomenon of unanswerable questions usually takes place between a few passage words and question words. The authors went further to investigate three different architectures for the answer verifying task. The first one is a sequential model that takes two sentences as a long sequence, while the second one attempts to capture interactions between two sentences. The last one is a hybrid model that combines the above two models to test if the performance can be further improved. Lastly, they evaluated their system on the SQuAD 2.0 dataset. Their best reader achieves a F1 score of 73.7 and 69.1 on the development set, with or without ELMo embeddings. When combined with the answer verifier, the whole system improves to 74.8 F1 and 71.5 F1 respectively. Moreover, the best system obtains a score of 74.2 F1 on test set, achieving state-of-the-art results at the time of submission (Aug. 28th, 2018).

* 1. ***Learning to Ask Unanswerable Questions for Machine Reading Comprehension.***

In this work, Haichao Zhu, Li Dong, Furu Wei, Wenhui Wang, Bing Qin, Ting Liu (2019) proposed a data augmentation technique by automatically generating relevant unanswerable questions according to an answerable question paired with its corresponding paragraph that contains the answer.

The authors introduced a pair-to-sequence model for unanswerable question generation, which captures effectively the interactions between the question and the paragraph. In addition, they present a way to construct training data for their question generation models by leveraging the existing reading comprehension dataset. They further used the automatically generated unanswerable questions as a means of data augmentation on the SQuAD 2.0 dataset, yielding 1.9 absolute F1 improvement with BERT-base model and 1.7 absolute F1 improvement with BERT-large model.

The authors believed that for unanswerable questions, the systems are supposed to abstain from answering rather than making unreliable guesses, which is an embodiment of language understanding ability.

They tackled the issue by automatically generating unanswerable questions for data augmentation to improve question answering models. The generated unanswerable questions should not be too easy for the question answering model so that data augmentation can better help the model. In this work, they proposed to generate unanswerable questions by editing an answerable question and conditioning on the corresponding paragraph that contains the answer. So the generated unanswerable questions are more lexically similar and relevant to the context. Moreover, by using the answerable question as a prototype and its answer span as a plausible answer, the generated examples can provide more discriminative training signal to the question answering model. They used answer spans in paragraphs as pivots to align pairs of answerable questions and unanswerable questions, which in turn creates training data for unanswerable question generation.

With this, they obtained the data with which the models can learn to ask unanswerable questions by editing answerable ones with word exchanges, negations, etc. They introduced a pair-to-sequence model to better capture the interactions between questions and paragraphs. The proposed model first encodes input question and paragraph separately, and then conducts attention-based matching to make them aware of each other. The context-aware representations are then used to generate outputs. They also incorporated the copy mechanism (Gu et al., 2016; See et al., 2017) to facilitate the use of context words during the generation process. Their experimental results on the unanswerable question generation task shows that the pair-to-sequence model generates consistently better results over the sequence-to-sequence baseline and performs better with long paragraphs than with short answer sentences. Their further experimental results show that the generated unanswerable questions can improve multiple machine reading comprehension models. Even using BERT fine-tuning as a strong reading comprehension model, we can still obtain a 1.9% absolute improvement of F1 score with BERT-base model and 1.7% absolute F1 improvement with BERT-large model.

On their problem formulation, given an answerable question *q* and its corresponding paragraph *p* that contains the answer *a*, their objective is to generate unanswerable questions *q˜* that satisfies certain requirements. Firstly, it cannot be answered by paragraph *p*. Secondly, it must be relevant to both answerable question *q* and paragraph *p*, which refrains from producing irrelevant questions. Thirdly, it should ask for something of the same type as answer *a*. The authors investigated two simple neural models built upon encoder-decoder architecture (Cho et al., 2014; Bahdanau et al., 2015) to generate unanswerable questions. A sequence-to-sequence model takes the concatenated paragraph and question as input and encodes the input in a sequential manner. A pair-to-sequence model is further introduced to capture the interactions between inputs. The decoder of two models generates unanswerable questions sequentially. They factorized the probability of generating the unanswerable question.

In the sequence-to-sequence model, paragraph and question pairs are cmobined into an ordered sequence x with a special separator in between. They introduced token type embeddings which can also be used to distinguish questions from paragraphs in sequence-to-sequence model which in turn indicates answers in paragraphs. For a given token, they formed the input representation ei by summing the corresponding word embeddings, character embeddings and token type embeddings. Here characters are embedded by an embedding matrix followed by a max pooling layer. They applied a single-layer bi-directional recurrent neural networks with long short-term memory units (LSTM; Hochreiter and Schmidhuber, 1997) to produce encoder hidden states. Moreover, they used an attention mechanism to summarize the encoder-side information into ct for current step.

In pair-to-sequence, the interactions make the paragraph and question aware of each other and help to predict the answer more precisely. Hence, they proposed a pair-to-sequence model, conducting attention-based interactions in encoder and subsequently decoding with two series of representations. In pair-to-sequence model, the paragraph and question are embedded as in sequenceto-sequence model, but encoded separately by weight-shared bi-directional LSTM networks, yielding h p i = fBiLSTM(h p i−1 , e p i−1 ) as paragraph encodings and h q i = fBiLSTM(h q i−1 , e q i−1 ) as question encodings. They used the same attention mechanism as in sequence-to-sequence model in the interaction layer to produce question-aware paragraph representations.

In generating unanswerable question and training data construction, they used plausible answer spans in paragraphs as pivots to align pairs of answerable questions and unanswerable questions. They sorted the pairs by Levenshtein distance (Levenshtein, 1966) and keep the pair with the minimum distance, and make sure that each question is only paired once as to the spans that correspond to multiple answerable and unanswerable questions. They got 20, 240 aligned pairs from the SQuAD 2.0 dataset in total. The Levenshtein distance between the answerable and unanswerable questions in pairs is 3.5 on average. They randomly sample 46 articles from the SQuAD 2.0 training set with 1, 805 (∼10%) pairs as holdout set and evaluate generation models with 2, 765 pairs extracted the SQuAD 2.0 development set since the SQuAD 2.0 test set is hidden.

Also, they conducted human evaluation on 100 samples in three criteria: (1) unanswerability, which indicates whether the question is unanswerable or not; (2) relatedness, which measures semantic relatedness between the generated question and input question answering pair; (3) readability, which indicates the grammaticality and fluency. They requested three raters to score the generated questions in terms of relatedness and readability on a 1-3 scale (3 for the best) and determine the answerability in binary (1 for unanswerable). The raters are not aware of the question generation methods in advance.

For the results of the automatic evaluation, they found that the proposed pair-to-sequence model that captures interactions between paragraph and question performs consistently better than sequence-to-sequence model.

On data augmentation setup, they first generated unanswerable questions using the trained generation model. They specifically used the answerable questions in the SQuAD 2.0 training set, besides ones aligned before, to generate unanswerable questions. Then they used the paragraph and answers of answerable questions along with the generated questions to create training examples. Lastly, they obtained an augmentation data containing 69, 090 unanswerable examples. Then they trained question answering models with augmentation data in two separate phases. In the first phase, they trained the models by adding the augmentation data and all 86, 821 SQuAD 2.0 answerable examples. They subsequently used the original SQuAD 2.0 training data alone to further fine-tune model parameters.

For the results, they used the exact Match (EM) and F1 as two metrics to evaluate model performance. EM measures the percentage of predictions that match ground truth answers exactly while F1 measures the word overlap between the prediction and ground truth answers. By default, they used pair-to-sequence model with answerable questions and paragraphs for data augmentation. They observed that the generated unanswerable questions can improve both specifically designed reading comprehension models and strong BERT fine-tuning models, yielding 1.9 absolute F1 improvement with BERT base model and 1.7 absolute F1 improvement with BERT-large model. Finally, their submitted model obtains an EM score of 80.75 and an F1 score of 83.85 on the hidden test set.

The research can be further improved on few aspects:

* By enhancing the ability to utilize antonyms for unanswerable question generation via leveraging external resources.

**3. Research Problem**

This research will focus on building Squad 2.0 datasets based on general knowledge rule.

Current NLP (Neuro-Linguistic Programming) model is given a paragraph, and a question about that paragraph, as input. The goal is to answer the question correctly. From a research perspective, this is an interesting task because it provides a measure for how well systems can ‘understand’ text. The goal of this project is to sort out the ground truth answer that has no answer and classify it whether can be solve by implicit general knowledge.