Machine Reading Comprehension using SQUAD

Sam Clastine Jesumuthu
220038747
MSc Data Science
sam.jesumuthu@city.ac.uk

1 Problem statement and motivation

Reading comprehension is a challenging task since it is important to read and understand natural language texts and responding to queries about it. The lack of precise and effective reading comprehension models made difficult to natural language processing field advancement. The goal of this study is to verify the predicted answer and say how correct is the answer. In this study we had focused on the the passage and the question is answerable or not. And find how Logistics regression and the state of the art models perform on the robust squad v2 dataset.For logistic regression, to investigate the negation feature and POS tag features varies the model performance in a positive or a negative way.We expect to gain insight into the strengths and shortcomings of these models, as well as the necessity of data prepossessing technique in constructing a accurate and effective reading comprehension classification models, by performing this study.

2 Research hypothesis

Questions for the given passage in SQUAD dataset always cannot be accurate there will be some wrong prediction, due to High complexity in human natural language. So How to identify the predicted Answer is wrong or does the question is Answerable or Not.
The SQUAD dataset is one of the most extensive and widely used for machine reading comprehension tasks. Despite efforts has been taken to increase accuracy, however errors and incorrect predictions may occur due to the high complexity of human natural language. As a result, it is necessary to identify and examine the predicted answers, as well as to establish if a question is answerable or not. This is especially crucial for applications like information retrieval and question-answering systems that rely largely on machine comprehension.

To reduce the error in machine reading comprehension, we need to create a model that predicts or verifies whether or not the question for the given paragraph is answerable. As there is lot of complication is human language, it is important to check that the given reasoning is relevant to the passage or not. The solution for this use machine learning algorithm and natural language understanding to classify the unanswerable questions based on the contextual aspects of passage and question. This approach has capability to reduce error that is mispredicting wrong information.

3 Related work and background

Many QA approaches and models had been made but most of them, is widely focused on correct answer is guaranteed. Nevertheless, they are unable to recognise unanswerable queries and instead return an unreliable text. Clark and Gardner (2018) illustrated two methods one is confidence method and the others No-Answer option, the algorithm used was Bidirectional GRU followed by linear model to compute scores. In confidence method, the algorithm then applied to the multi-paragraph context by measuring the model's confidence using the un-normalized and un-exponentiated (i.e., before the softmax operator is applied) score assigned to each span. And in No-Answer option the reader generates a no-answer probability after extracting a potential response (NA Prob). The answer verifier then determines if the extracted response is valid or not. Ultimately, the algorithm combines past data to get the final forecast. A function that combines a "no-answer" score and answer span scores to provide a probability that a question is unanswerable as well as output a proposed answer. Kundu and Ng (2018) did study on a nil-aware answer span extraction that may return Nil or a text span from the connected passage as a response in a single step.They had aggregated the orthogonally decomposed evidence vectors, it combines both the supportive and

unsupportive pieces of evidence for a particular passage word. To obtain the most impactful portions, they had performed a max-pooling operation over all the aggregated vectors. The resulting vector is denoted as the Nil vector.Both the papers have not, however, investigated further confirming the question's answerability by checking the legitimacy of the predicted answer. In this context, answerability refers to whether the question has an answer, and legitimacy refers to whether the extracted text is supported by the passage and the question.Hu et al. (2019b) proposed a system that solves the previous No-Answer method or NIL-Aware by read-then-verify as human does. The system consist 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.

Hu et al. (2019a) introduces a new neural network model for reading comprehension that can handle numerous answer spans and discrete reasoning issues.Many existing reading comprehension models, according to Hu et al. (2019a), are limited in their ability to handle questions that require discrete reasoning, such as counting, comparison, or logical operations. To solve this restriction, the authors offer a new model that identifies and extracts several answer spans from the input text using a multi-type multi-span network. The network is made up of two major components: The token-level answer span predictor predicts the beginning and ending positions of each response span, while the sentence-level answer span selector selects the most appropriate sentence(s) containing the answer spans.

Wang and Jiang (2016) demonstrates a neural network model for machine understanding that seeks to answer questions based on given passages. A Match-LSTM (Long Short-Term Memory) network and an answer pointer mechanism are the two key components of the suggested concept. The Match-LSTM network captures the semantic relationship between the question and the passage, and the response pointer mechanism locates the answer within the passage. The Stanford Question Answering Dataset (SQuAD), the CNN/Daily Mail dataset, and the Children's Book Test are among the benchmark datasets used by the authors to assess their approach. On these datasets, they demonstrate that their model outperforms several baseline methods, achieving state-of-the-art performance on the SQuAD dataset.

The research done by (Gong et al., 2020) presents a novel recurrent chunking approach for long-text machine reading comprehension (MRC) challenges.Because of their limited memory capacity and lack of appropriate chunking methods, existing MRC models struggle to handle long-text sections. To solve this restriction, the authors suggest a novel model that divides the long-text passage into smaller, more understandable chunks using recurrent chunking processes. A chunk proposal module and a chunk reading module comprise the recurrent chunking system. The chunk proposal module generates a set of candidate chunks based on the input passage using a recurrent neural network (RNN), whereas the chunk reading module reads each candidate chunk and generates an answer using a separate RNN.

According to the (Zhuang et al., 2021), BERT model has a significant shortcomings, including a lack of diversity in training data and the use of a predetermined training schedule. To address these restrictions, the authors designed a new pretraining strategy that involves optimising numerous hyperparameters such as batch size, learning rate, and training step count. Furthermore, the authors uses more training data and eliminate certain training objectives that they discovered to be ineffective. During pretraining, they additionally employ a dynamic masking strategy that masks out distinct combinations of tokens randomly during each training cycle. RoBERTa was evaluated using a variety of benchmark datasets, including GLUE, SQuAD, and RACE. And it outperformed multiple state-of-the-art models on all benchmark datasets, achieving new state-of-the-art results.

Existing pretraining approaches for language models, such as ELMo and GPT, the (Devlin et al., 2019) believe, have drawbacks due to their unidirectional or shallow architecture. To overcome these constraints, the authors offer a new pretraining technique based on a deep bidirectional transformer architecture. The BERT model is trained on vast volumes of unlabeled text data by performing two tasks: masked language modelling and next sentence prediction. The masked language modelling job entails masking out some words in the input sentence at random and training the model to predict the masked words based on context. The following sentence prediction problem entails training the model to predict whether or not two in-

put sentences are sequential.The authors evaluates the BERT model's performance on multiple benchmark datasets, including GLUE and SQuAD. On these datasets, they show that the BERT model outperforms several state-of-the-art models, achieving new state-of-the-art results on several tasks.

The (Chen et al., 2016) look at how different neural network models perform on a reading comprehension assignment based on CNN and Daily Mail news items. They initially present a dataset of news stories linked with questions and responses, with the answer being a piece of text from the article. Then they compare the performance of numerous models, including a logistic regression model, a feedforward neural network model, and multiple types of a recurrent neural network model, including a Long Short-Term Memory (LSTM) model and a Gated Recurrent Unit (GRU) model, the LSTM model outperforms the other models. They also undertake multiple experiments to acquire insights into how the LSTM model works, such as analysing the model's attention weights and conducting ablation research to understand the impact of various model components. And the (Chen et al., 2016) also study the impact of different preprocessing strategies, such as stemming and stop-word removal, on the performance of the models in addition to analysing alternative neural network models. They demonstrate that these strategies have a minor influence on performance, implying that neural network models can learn to accommodate text variances.

This work done by Dibia (2020) proposes NeuralQA, a new open-source library for developing question-answering systems that can handle big datasets. Due to computational and memory constraints, existing question answering systems are limited in their ability to handle big datasets. To solve this constraint, the authors suggest a new library that improves performance on huge datasets by utilising contextual query expansion and BERT (Bidirectional Encoder Representations from Transformers). A data pre-processing module, a contextual query expansion module, and a BERT-based answer selection module comprise the NeuralQA library. The data pre-processing module cleans and prepares the input data, while the contextual query expansion module adds context to the input question.

The Simple Transformers library by Rajapakse is an open-source Python library that provides an easy-to-use interface for training and using cutting-edge transformer models for a variety of natural language processing (NLP) tasks such as text classification, question answering, named entity recognition, and more. The library is based on the Hugging Face Transformers library, which contains pre-trained transformer models as well as an interface for fine-tuning them for specific needs. Simple Transformers makes fine-tuning easier by offering a high-level API for data preparation, model training, assessment, and prediction, as well as a set of preset hyperparameters that are suitable for most jobs.

In terms of usability, both libraries strive to make training and deploying NLP models easier. Simple Transformers, on the other hand, is intended to be more user-friendly, with a high-level API that abstracts away many of the intricacies of model training and evaluation. The NeuralQA library is mainly focused on offering a collection of tools and models that are optimised for question answering, but using it efficiently may necessitate more specialised knowledge and skill.

3.1 Accomplishments

1. Task 1: Literature Review and Problem Definition – Completed
2. Task 2: Text cleaning like Removing special characters and lowercase all the text - Completed
3. Task 3: Tokenizing the passage and the question for logistic regression and pretrained models- Completed
4. Task 4: Training the baseline model for all 4 models - Completed
5. Task 5: Evaluating all 4 models - Completed
6. Task 6: Performed in-depth error analysis - Completed.

4 Approach and Methodology

For the hypothesis, we will using Machine learning Algorithms and Natural language processing to interpret analyze and extract important feature form the SQUAD dataset. Specifically for the Baseline we will be using Logistics regression and Pretrained Models like BERT-Base, ROBERTA-Base and ROBERTA-XLM-Base to extract features and train on the dataset.

4.1 Methodology

1. Data Collection: In this study, we would be using SQUAD 2.0 dataset, which consist of Context, Question and Answers along with the span. The dataset is download and load using Hugging face library 'datasets'.

2. Data Prepossessing: After data Collection, natural language text has to cleaned up to reduce noise. Some of the fundamental preprocessing technique we would be using are removing punctuations, converting all text to lower case and tokenizing the text and it changes based models we are training on some maybe added and some will be reduce or not needed. The library we will be using for this will be NLTK REGEX and Spacy.

3. Feature Extraction: For feature extraction, it depend on the model like for logistic regression we would be extracting negation feature and POS tag feature and vectorize using TFIDF. For pretrained model we would using Autotokenizer to tokenize, padding, and for extracting inputid and Attention mask for each passage.

4. Training and Evaluation: For training the logistics regression we would be using sklearn and for pretrained model we will be using Transformer library. In evaluation Part, will assessing the model based on recall, F1 Score and ROC curve.

5. Limitations in Baseline and Drawback in pretrained Models: The implication of baseline (logistics regression) fails when i give the whole dataset, so i just trained the model using 5000 sample. While the pretrained mode approximately took more that One hour for a epoch, so it pushed me to reduce the data size same as the baseline to 5000.

6. working implementation: As the model accuracy can be improved and optimized but computional cost issue and time constraints the model is not ready for full end to end implementation.

7. Code File : Training.ipynb consist of all 4 models preproccessed, trained and evaluated. Testing.ipynb consist of all 4 models loaded from the saved weight, evaluated on classification metrics like f1 score, recall and ROC curve.

8. The baseline saved model did not functioned in testing.ipynb, it was given the shape issue the error is shown in the testing.ipynb file.

5 Dataset

SQUAD 2.0 is reading comprehension dataset which consist of 130319 rows in train set and 11873 in the dev set. The figure 1 show s the overall distribution of the class ie (Unaswerable - 1 , answerable -0). Potential avenues for the datasets are: https://huggingface.co/datasets/squad_v2, https://rajpurkar.github.io/SQuAD-explorer/.

5.1 Dataset preprocessing

Data Preprocessing Technique applied for the Baseline and Pretrained Model are Listed Below:

Logistics Regression: For Logistic Regression,the initial steps are mentioned, after that we had to extract the negation feature and POS tag feature and concatenate to the text which is Passage and Question. Then the dataset is vectorized using TFIDF vectorizer.

Pretrained Models: We had used Transformer library to perform tokenization and vectorization for each pretrained model.And it also padding, adding tokens and other basic preprocessing tectniques. And the output will be inputid and attention mask which is need to train BERT based model. After that it is convert to torch tensor and then load to dataloader of batch size 16.

6 Baselines

The Baseline which is considered for the study is logistics regression, the major advantage of this baseline is time efficiency, for train a pretrained model with same 5000 examples took more than 2 minutes for a single epoch with high loss while it trained in a minute is validation accuracy is 66 percent.

	Anserable	Not Answerable	Total
Train	86821	43498	130319
Dev	5945	5982	11873

Figure 1

7 Results, error analysis

In baseline, We got validation accuracy 0.66 which is quiet less than the other pretrained model, the highest validation accuracy among all is ROBERTA model is 0.795 and its f1 score for both the classes is fair good but not great than the BERT-Base it has predicted the highest Unanswerable class in the test set with a recall of 35 percent true positive rate and the ROBERTA comes the second with 22percent. While the baseline recall rate for unanswerable class is 0.08 which is the least performed model and the XLM ROBERTA model fails to predict the unanswerable class its recall value is 0. Based on the result and findings the Report the performance of your model and compare it to the baselines.

Considering figure 3 which shows the ROC Curve, it show that how all models have high false positive rate only for pretrained models. Test data error in baseline could able to plot due to shape issues. But according to the table it recall score is 0.08 we can assume that it would also ploted under the diagonal, Based on all model ROBERTA performed well with low false positive rate when compare to the other model, the second model performed good is BERT with less false positive rate than Baseline and XLM ROBERTA.

7.1 Sample 1

Predicted : Answerable **Truth** : Unanswerable
In figure 4, the Question is "What increased about Harvardś student body in 2005?" According to the Rajpurkar et al. there are 20percent antonyms used questions in SQUAD for example "decline", "increased" or "decreased" this type of question can be analyzed qualitatively to answer it is difficult to model say this is answerable or not if such kind of Antonyms are used.

7.2 Sample 2

Predicted : Answerable **Truth** : Unanswerable
In figure 5, the Question is "What borders did the French complete during the Middle Ages?" This a neutral question where the passage doesn't imply any answer, this type of question have high vocabulary relation but doesn't have an answer to the question. Thus the based on context there are lot of similarities maybe this the reason it has predicted Answerable.

Models	Classes	Recall	F1 Score	Val Accuracy	Test Accuracy
Logistic Regression	Unanswerable	0.08	0.14	0.66	N/A
	Answerable	0.96	0.79		
BERT-Base	Unanswerable	0.35	0.45	0.7	0.53
	Answerable	0.75	0.59		
ROBERTA-Base	Unanswerable	0.22	0.35	0.795	0.552
	Answerable	0.97	0.66		
XLM-ROBERTA	Unanswerable	0	0	0.67	0.446
	Answerable	1	0.66		
Best Model		The two best model are BERT (0.53) and ROBERTA (0.552)			

Figure 2

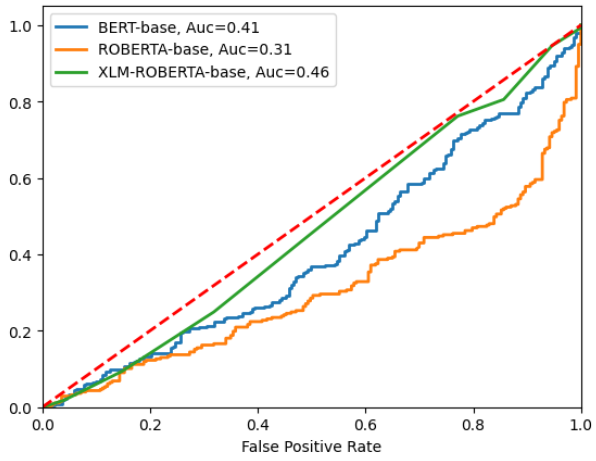

Figure 3

Harvard\'s academic programs operate on a semester calendar beginning in early September and ending in mid-May. Undergraduates typically take four half-courses per term and must maintain a four-course rate average to be considered full-time. In many concentrations, students can elect to pursue a basic program or an honors-eligible program requiring a senior thesis and/or advanced course work. Students graduating in the top 4–5% of the class are awarded degrees summa cum laude, students in the next 15% of the class are awarded magna cum laude, and the next 30% of the class are awarded cum laude. Harvard has chapters of academic honor societies such as Phi Beta Kappa and various committees and departments also award several hundred named prizes annually. Harvard, along with other universities, has been accused of grade inflation, although there is evidence that the quality of the student body and its motivation have also increased. Harvard College reduced the number of students who receive Latin honors from 90% in 2004 to 60% in 2005. Moreover, the honors of "John Harvard Scholar" and "Harvard College Scholar" will now be given only to the top 5 percent and the next 5 percent of each class. What increased about Harvard\'s student body in 2005?

Figure 4

7.3 Sample 3

Predicted : Answerable **Truth** : Unanswerable
In figure 6, the Question is "What commission did not resign even though faced corruption allegations?" The passage have the answer it is clearly mentioned " the **Santer Commission** was censured by Parliament in 1999, and it eventually resigned due to **corruption allegations**. " It is vague that the truth value is Unanswerable and the predicted is answerable. According to Rajpurkar et al. this type reasoning called mutual exclusion that is Word or phrase is mutually exclusive with something for which an answer is present.

Since the Peace of Westphalia, the Upper Rhine formed a contentious border between France and Germany. Establishing "natural borders" on the Rhine was a long-term goal of French foreign policy, since the Middle Ages, though the language border was – and is – far more to the west. French leaders, such as Louis XIV and Napoleon Bonaparte, tried with varying degrees of success to annex lands west of the Rhine. The Confederation of the Rhine was established by Napoleon, as a French client state, in 1806 and lasted until 1814, during which time it served as a significant source of resources and military manpower for the First French Empire. In 1840, the Rhine crisis, prompted by French prime minister Adolphe Thiers's desire to reinstate the Rhine as a natural border, led to a diplomatic crisis and a wave of nationalism in Germany. What borders did the French complete during the Middle Ages?

Figure 5

Commissioners have various privileges, such as being exempt from member state taxes (but not EU taxes), and having immunity from prosecution for doing official acts. Commissioners have sometimes been found to have abused their offices, particularly since the Santer Commission was censured by Parliament in 1999, and it eventually resigned due to corruption allegations. This resulted in one main case, Commission v Edith Cresson where the European Court of Justice held that a Commissioner giving her dentist a job, for which he was clearly unqualified, did in fact not break any law. By contrast to the ECJ's relaxed approach, a Committee of Independent Experts found that a culture had developed where few Commissioners had 'even the slightest sense of responsibility'. This led to the creation of the European Anti-fraud Office. In 2012 it investigated the Maltese Commissioner for Health, John Dalli, who quickly resigned after allegations that he received a €60m bribe in connection with a Tobacco Products Directive. Beyond the Commission, the European Central Bank has relative executive autonomy in its conduct of monetary policy for the purpose of managing the euro. It has a six-person board appointed by the European Council, on the Council's recommendation. The President of the Council and a Commissioner can sit in on ECB meetings, but do not have voting rights. What commission did not resign even though faced corruption allegations?

Figure 6

7.4 Sample 4

The use of remote sensing for the conservation of the Amazon is also being used by the indigenous tribes of the basin to protect their tribal lands from commercial interests. Using handheld GPS devices and programs like Google Earth, members of the Trio Tribe, who live in the rainforests of southern Suriname, map out their ancestral lands to help strengthen their territorial claims. Currently, most tribes in the Amazon do not have clearly defined boundaries, making it easier for commercial ventures to target their territories. **On-site sensing is being used by indigenous tribes for what?**

Figure 7

Predicted : Answerable **Truth** : Unanswerable
In figure 7, the Question is "On-site sensing is being used by indigenous tribes for what?" When the user is not satisfied with the paragraph or sentence this question may implies and this is impossible to answer. The make machine to understand this type of context is difficult, as type of reasoning comes under impossible condition and getting ground truth is quiet difficult as it has high similarities.

7.5 Sample 5

Predicted : Answerable **Truth** : Unanswerable
In figure 8, the Question is "What commission

The University of Chicago (UChicago, Chicago, or U of C) is a private research university in Chicago. The university, established in 1890, consists of The College, various graduate programs, interdisciplinary committees organized into four academic research divisions and seven professional schools. Beyond the arts and sciences, Chicago is also well known for its professional schools, which include the Pritzker School of Medicine, the University of Chicago Booth School of Business, the Law School, the School of Social Service Administration, the Harris School of Public Policy Studies, the Graham School of Continuing Liberal and Professional Studies and the Divinity School. The university currently enrolls approximately 5,000 students in the College and around 15,000 students overall. What types of other schools is the city of Harris well known for?

Figure 8

did not resign even though faced corruption allegations?" The passage have the answer it is clearly mentioned " the **Santer Commission** was censured by Parliament in 1999, and it eventually resigned due to **corruption allegations**. " It is vague that the truth value is Unanswerable and the predicted is answerable. According to Rajpurkar et al. this type reasoning called mutual exclusion that is Word or phrase is mutually exclusive with something for which an answer is present.

7.6 Sample 6

The immune system is a system of many biological structures and processes within an organism that protects against disease. To function properly, an immune system must detect a wide variety of agents, known as pathogens, from viruses to parasitic worms, and distinguish them from the organism's own healthy tissue. In many species, the immune system can be classified into subsystems, such as the innate immune system versus the adaptive immune system, or humoral immunity versus cell-mediated immunity. In humans, the blood–brain barrier, blood–cerebrospinal fluid barrier, and similar fluid–brain barriers separate the peripheral immune system from the neuroimmune system which protects the brain. What separates the neuroimmune system and peripheral immune system in humans?

Figure 9

Predicted : Unanswerable **Truth** : Answerable
In figure 9, the Question is "What separates the neuroimmune system and peripheral immune system in humans?" The passage has the answer which is pretty straight forward " fluid–brain barriers separate the peripheral immune system from the neuroimmune system".

7.7 Sample 7

Predicted : Answerable **Truth** : Unanswerable
In figure 10, the Question is "What commission did not resign even though faced corruption allegations?" The passage says about the second arm off 'Leukocytes (white blood cells) act like independent, single-celled organisms and are the second arm of the innate immune system.' Due the high

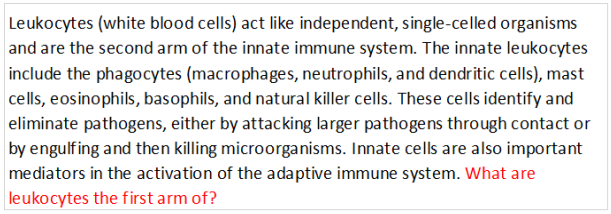

Figure 10

relevance in passage and question the high probable answer will be be Answerable. This type of reasoning comes under 'Other neutral' it says where the paragraph or a passage doesn't have any answer to the question.

7.8 Sample 8

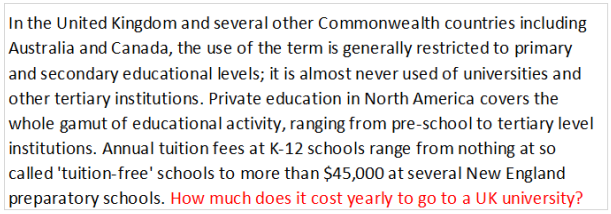

Figure 11

Predicted : Answerable **Truth** : Unanswerable In figure 11, the Question is "How much does it cost yearly to go to a UK university? " this is a quantitative reasoning this question is answerable to but the model predicted as answerable may be due to high relevance in the context.

8 Lessons learned and conclusions

SQUAD dataset was challenging such as the size and complextity in passages and question. I ahve learned how to handle large dataset how to process and use many linraries like transformers, simpletransformer and wandb for visualizing the loss in ever epoch. On of the challenge i faced while building logistics regression is extraction negation words, by using dependency parsing in spacy library, i got to know the tree of words and how it is correlated to each other. In pretrained model, i had used transformer to vectorizer based on the model which i had use in that i learned how add tokens and how to set maxLen and Ndimension in embedding layer and how padding is important is deep learning task. The inital goal of the project was create a model which predict the span and the answerable and answerable score due to time and machine constraints the task was simplified to build only the verification model. Overall the model training evaluation for the given hypothesis was successfully and the best model performed is ROBERTA which predicted the high unanswerable class. For the further research, we would be merging this model with the span prediction model for full end to end implementation.

References

Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016. A thorough examination of the cnn/daily mail reading comprehension task.

Christopher Clark and Matt Gardner. 2018. Simple and effective multi-paragraph reading comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 845–855, Melbourne, Australia. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Victor Dibia. 2020. NeuralQA: A usable library for question answering (contextual query expansion + BERT) on large datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 15–22, Online. Association for Computational Linguistics.

Hongyu Gong, Yelong Shen, Dian Yu, Jianshu Chen, and Dong Yu. 2020. Recurrent chunking mechanisms for long-text machine reading comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6751–6761, Online. Association for Computational Linguistics.

Minghao Hu, Yuxing Peng, Zhen Huang, and Dongsheng Li. 2019a. A multi-type multi-span network for reading comprehension that requires discrete reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1596–1606, Hong Kong, China. Association for Computational Linguistics.

Minghao Hu, Furu Wei, Yuxing Peng, Zhen Huang, Nan Yang, and Dongsheng li. 2019b. Read + verify: Machine reading comprehension with unanswerable questions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:6529–6537.

Souvik Kundu and Hwee Tou Ng. 2018. A nil-aware answer extraction framework for question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4243–4252, Brussels, Belgium. Association for Computational Linguistics.

T. C. Rajapakse. 2019. Simple transformers. https://github.com/ThilinaRajapakse/simpletransformers.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. 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)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Shuohang Wang and Jing Jiang. 2016. Machine comprehension using match-lstm and answer pointer.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.