A Comparison of Approaches to Question and Answering Using a Novel Biology GCSE Data Set

Dhen Emmanuel Padilla
Thomas Collyer
Edward Crookenden
Arthur Turner

Abstract

A comparison of the different approaches for selecting the most appropriate contextual text to input into a BERT based question and answering model to achieve high performance on GCSE biology standard questions. In doing so, a novel GCSE biology dataset was produced from a textbook, split into sub-sections and accompanied with questions. The use of GCSE standard information and questions allows a direct comparison to human intelligence for short answer questions. The task requires a model which can not only be accurate but also be of reasonable computation. A two-layer model was used, the first being a form of passage retrieval or selection to feed into the second, a question and answering model using ALBERT. A baseline was achieved by inputting an entire textbook into the ALBERT. A BM25 algorithm in layer 1 to return 5 contextual texts was the best performing model and achieved an accuracy of 83% improving on the baseline by 4% as well as decreasing the computation time by a factor of 15. The high accuracy shows that although NLP is far from reasoning about science it can achieve strong performance compared to humans for short answer GCSE questions.

1 Introduction

Question Answering (QA) systems have become an increasingly prevalent and powerful platform over the past decade, allowing users to ask questions and receive answers automatically in natural language. As a result of more academic research, knowledge and techniques used in the field is growing rapidly. Large data sets are also helping develop the field, such as the Stanford Question and Answering Dataset (SQuAD) (Rajpurkar et al., 2016) and the Wiki Table Questions dataset (Pasupat and Liang, 2015), encouraging machine learning and Natural Language Processing methods to be utilised. One application of QA systems is the auto-completion of exam papers, to pass a paper students need to recall a varied knowledge contained within a textbook. This immediately presents a more complex problem as when the context passage length increases, the computation time of a state of the art QA models rapidly increases. On top of this, in some cases, the increased length can also cause a decrease in performance. For the auto-completion of an exam neither of these traits are characteristic, and therefore need a context passage with the relevant information associated with the question to answer a single question. Traditionally passage retrieval has been carried out using weighting factors such as term frequency-inverse document frequency (tf-idf). This study aims to compare a tf-idf variant, BM25, versus a raw NLP implementation for passage retrieval for auto exam completion. Methods will be compared through using our novel GCSE Biology Textbook QA dataset created specifically to add complexity to the passage retrieval due to high homogeneity and overlapping information contained within the document. The main outcomes of our work will include: a comparison of relevant passage retrieval accuracy and final performance of the novel models on a generated bespoke biology test.

2 Related Work

The question and answering problem has been a major challenge in Artificial Intelligence and Computer science for many years. This is due to the complex combination of information retrieval, natural language processing, machine learning and computer-human interfaces. After initial limited success, substantial progress was found through a major breakthrough by IBM's Watson system (Ferrucci et al., 2010) and the design of it's

DeepQA architecture. Watson's task was to "compete at the human champion level in real-time on the American TV quiz show, Jeopardy." This required the system to provide exact answers to complex natural language questions with high precision and speed while achieving an accuracy of at least 70%. The DeepQA approach was to use a parallel probabilistic evidence-based architecture composed of more than 100 different techniques to analyse natural language, identify sources, find and score evidence, and merge and rank hypotheses. The overarching principles in the DeepQA architecture are massive parallelism, many experts, pervasive confidence estimation and shallow and deep knowledge. In 2011 Watson competed on 'Jeopardy' against two previous champions and won'. This success provided a benchmark in the field for future projects to aim for and surpass.

After the success of Watson, the next substantial development came from the integration of deep neural architectures. End-to-end neural models require vast amounts of data to be trained effectively and it was not until the creation of datasets to facilitate this such as CNN/DailyMail (Hermann et al., 2015) and SQuAD (Rajpurkar et al., 2016) that deep architectures could be deployed.

The Q&A deep learning realm has been predominantly occupied by attention-based deep neural networks. These models can be fine-tuned in various ways to achieve superior performance, this also includes the way attention weights are calculated and handled. 'Teaching Machines to Read and Comprehend' (Hermann et al., 2015) incorporates attention mechanisms into recurrent neural network architectures to achieve state-of-the-art performance. Attention weights are updated dynamically given the query, context and previous attention. Inspired by this, 'A thorough examination of the cnn/daily mail reading comprehension task' (Chen et al., 2016) uses the attentive reader model with the difference of computing attention weights using a bi-linear term instead of the simple dot-product to achieve significantly better performance. 'Text Understanding with the Attention Sum Reader Network' (Kadlec et al., 2016) uses recurrent networks to read the document and query, and uses attention mechanisms to directly select the answer in the context. A simplified model results as all transformations past the attention step are removed. The multi-hop approach was introduced (Hill et al., 2015). A variant of the memory network is applied to repeatedly compute attention vectors between the query and context through multiple layers. The Bi-Directional Attention Flow (Seo et al., 2016) uses a hierarchical multi-stage architecture including CNNs, RNNs, and word embedding models for modelling representations of the context paragraph. It uses a memory-less attention mechanism and lets the attention vectors flow into their RNN layer. The model achieved state-of-the-art results on SQuAD and on the CNN/Daily Mail cloze test.

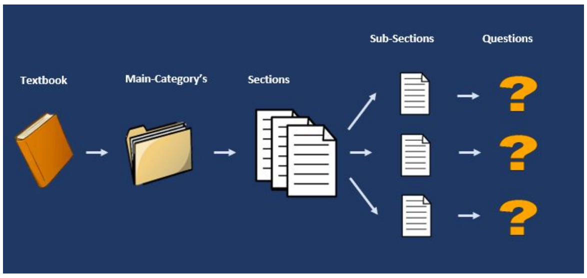
These approaches although not providing a full solution, achieved good accuracy in answering short questions provided the answer is given in the text. Benchmarks have been broken at a remarkable rate due to the fast development of deep learning models and their application to the natural language setting. The most recent development comes from the advent of large-scale language models such as ELMo (Peters et al., 2018), BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2019) etc. These are pretrained language representation models which can be used for a variety of natural language applications including the question and answering task. A notable success is 'The Aristo Project' (Clark et al., 2019) which uses BERT for answering non-diagrammatic multiple-choice-question based examinations. The progress represents a "significant milestone for the field" after achieving a mark greater than 90 % for the first time. The AristoBERT solver first uses an IR solver to retrieve up to 10 of the top sentences, truncated to fit into BERT. BERT is fine-tuned using a collection of science multiple-choice question sets. Lastly, the above is repeated with variations of BERT such as the original BERT-large-cased and BERT-large-uncased, as well as the later released BERT-large-cased whole-word-masking. These models are then ensembled together. The project uses an 8th Grade Science test dataset obtaining an independent benchmark which gives a direct comparison between computer performance and human performance.

The Q&A problem is part of the much wider objective of attaining human-like artificial intelligence. Although impressive accuracy has been reached the models cannot reason about science, only have a good enough probability distribution to select the correct answer. This shows the progress still to be made in Artificial Intelligence and full natural language understanding systems.

3 Methods

As part of this project, we produced a novel data set composed of texts from a Biology GCSE textbook. It was decided that the GCSE standard of information and question was a good benchmark to aim for. Any lower, we believed, would be too simple and would not fully utilise the contextual advantages of the BERT model. We also concluded that the subject of biology would be most fitting to analyse passage retrieval methods. With homogeneous texts and biology topics which reference and build upon other sections all adding to the overall complexity of passage retrieval. Although it may be possible to achieve success at a harder level such as A level, for this purpose we believed it was not suitable. Primarily because of the required background knowledge needed to answer questions sufficiently is not present in A level textbooks and would therefore not provide the necessary information to the model. The textbook was split up into different texts characterised by the first section and then in turn sub-section. Questions were composed for each of the sub-sections in the style of GCSE questions. It was not possible to collect official exam board GCSE questions due to both time constraints and a lack of availability given the amount we need. In total there are 4 main category's, 12 sections, 75 sub-sections which contain 300-400 tokens each and 1236 questions. A visualisation can be seen in figure 3. This was split into a train and test set with a ratio 80:20. A final "exam" was sampled from the test set to obtain an accuracy and observe a "realistic performance on an exam paper" for each of the models.

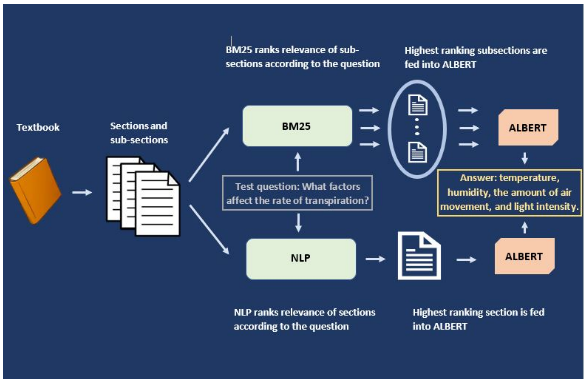
Figure 1: Visualisation of Dataset



The models constructed for the question and answering task centre around the use of BERT (Devlin et al., 2018), by which we fine-tune using text from the textbook and then feed a question for it to answer. We hypothesise that it is ineffective both in accuracy and computation to input the whole textbook as context to questions. A discussion of this can be found in the experiments section. In order to verify and test our hypothesis, a method is required to choose suitable text as input.

We therefore need a preliminary method of finding and choosing a suitable text. Thus, introducing a two-layer method. A first layer is created to take a question as input, and produce a relevant passage of text, based on the question. Layer 1 allows an entire textbook to be partitioned into smaller passages predicted to be relevant to a question. These texts are then fed into layer 2 which is an independent Q&A application of ALBERT(Lan et al., 2019), pre-trained using the Stanford Question and Answer Datasets (SQuAD) (Rajpurkar et al., 2016). For layer 2, the ALBERT model takes contextual data alongside the question to then answer the question. We adopted this synthesis in response to the limitations in accuracy and computation-time for very large text (40,832 tokens). Aiming to find a suitable trade-off between processing time and the final test score. A graphical representation of this can be seen in figure 2.

Figure 2: Visualisation of 2 Layer Method


BM25 ranks relevance of sub-sections according to the question
Highest ranking subsections are fed into ALBERT
Textbook
Sections and sub-sections
BM25
ALBERT
Test question: What factors affect the rate of transpiration?
Answer: temperature, humidity, the amount of air movement, and light intensity.
NLP
ALBERT
NLP ranks relevance of sections according to the question
Highest ranking section is fed into ALBERT

Before implementing more complex models, a baseline model and performance was required. We opted to use a simple 1 layer approach, where the ALBERT model, documented above, was used to answer exam questions. Contrary to the other models proposed, this model would be fed the entire textbook corpus as context. Experiments and results are produced later in the paper.

The reduced dependence of the two-pronged approach to the task allows a focused comparison of effectiveness between various approaches within the first layer mechanism. Although the correctness of answers provided by layer 2 is dependent on the passages retrieved in layer 1, numerous methods of passage retrieval may be explored without necessitating changes in layer 2. This investigation proposes a comparison between two separate passage retrieval techniques.

One approach to the first layer was to use the BM25 algorithm to rank the different texts depending on the given query. The BM25 algorithm was chosen due to it's performance for ranking shorter texts like the ones given for this task. The BM25 weighting scheme is a method of constructing a probabilistic model in order to perform passage search and ranking. The BM25 scheme dictates a formula for calculating the weight for a given token in a given query. For each token in a passage, the corresponding BM25 weights are summed to give an overall score which is then used for ranking in descending order. There are various versions and forms of the weight formula. In this case where no relevancy information is given, the following was used:

$$\log \frac{(N-n+0.5)}{(n+0.5)} \cdot \frac{(k_1+1)f_i}{K+f_i} \cdot \frac{(k_2+1)qf_i}{k_2+qf_i}$$

where $K = k_1((1-b) + b \cdot \frac{dl}{avdl})$, N is the total number of paragraphs, n is the number of paragraphs containing the token, f_i is the number of times the token appears in the given paragraph, dl is document length, $avdl$ is average document length and b, k_1 and k_2 are parameters set as seen below:

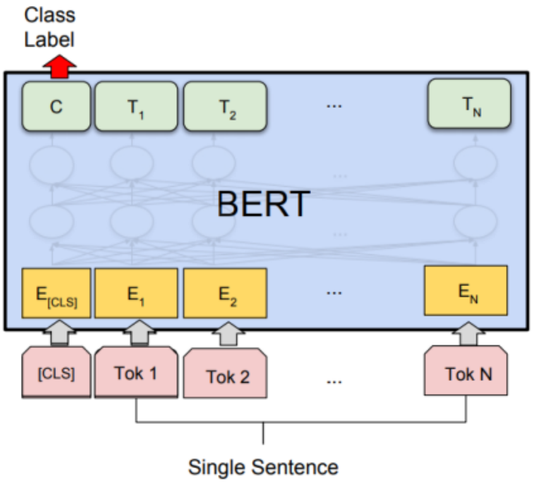
$$b = 0.75\,, \quad k_1 = 1.2, \quad k_2 = 100$$

The value of each weight is given a floor of 0. Without this if the word appears in over half the passages then it would be given a negative weight, which is undesirable.

From the final BM25 ranking, the text to be given to layer 2 can be extracted. The number of texts to feed to layer 2 was examined and a discussion can be found in the ”Experiments” section.

Another model was constructed with an entirely NLP-based approach to layer 1. As the objective of layer 1 is to retain a passage for a respective question, we employ a sequence classifier as our NLP layer 1 model (Figure 4), aiming to classify a subsection label given a question. This subsection label is used as an index to a passage of text to be passed to layer 2.

Figure 3: Sequence Classification



However, the data set we utilise poses a challenge for an NLP-based 1st layer, as the granularity of subsections increases the number of classes in which a 1st layer must classify. This can often lead a decrease in overall performance especially if data points become too sparse. In response, we exploit the hierarchical structure of the data (Main-Category, Section, Subsections) by trialling a chain of sequence classifiers reflecting the data's hierarchical nature. We employ separate premade RoBERTa models for classifiers at each layer of the hierarchy. RoBERTa is an extension of the BERT Bidirectional Transformer, which allows the unambiguous input of one or pair of sentences in one token sequence (Devlin et al., 2018). This sequence is inputted BERT's model architecture which involves multiple Transformer encoders. We omit an exhaustive background description of transformer architecture and in a similar fashion to the BERT publication, we refer to (Vaswani et al., 2017). BERT employs Transformer encoders to allow Masked Language Models (MLM) which randomly masks a small percentage of input tokens. We employ this tokeniser mechanism within BERT and add a classification layer which uses the MLM output (Figure 4). The classification layer is the only parameter added to the model having dimensions $K * H$, where K is the number of classified sections and H is the number of hidden layers. We fine-tune identical **RoBERTa**$_{BASE}$ models (Layers = 12, Hidden Dimension = 768, Attention heads = 12, 110M params) (Liu et al., 2019) for each classification task: Main-Category, Section, Sub-section, utilising the HuggingFace(Wolf et al., 2019) library of pre-trained NLP models to focus our study on fine-tuning . In total, 15 models were produced. A dictionary was created to hold models for each classification task. Each model requires an independent set of data, consisting of different questions and label-classes. Therefore, having various models allows for different NLP approaches to layer 1. We trial the following:

- Main-Category classification. This model will find a Category based on the question, and return a large text corpus consisting which is a concatenation of all passages within the classified category
- Section classification. RoBERTa is similarly fine-tuned on (Question, Section) pairs and returns the concatenated passages within the classified section
- Sub-Section classification. We fine-tune a RoBERTa on (Question, Sub-section) pairs and return the passage of the classified sub-section
- Hierarchical classification. This model is the most complex as it consists of a multi-model structure. RoBERTa models are fine-tuned for each parent in the hierarchy. A Main Category classifier to classify Sections. At the Section layer, each respective Section has an individually trained classifier aimed to classify a subsection within that section. Synthesising a model-chain, where an explicit model is applied conditioned on the output of a previous model.

A-Lite-BERT (ALBERT) was used for layer 2 of the models due to it's state-of-the-art language representation (Soricut and Lan). The model is a derivative of BERT with superior performance, achieved by allocating it's capacity more efficiently and then scaling up. For the implementation, we again took advantage of the HuggingFace library (Wolf et al., 2019) of pre-trained models. The specific model can be found with the following path: ”ktrapeznikov/albert-xlarge-v2-squad-v2” and the following configuration: n_best_size = 1, max_answer_length = 30, do_lower_case = True and null_score_diff_threshold = 0.0. It was then fed a question along with the corresponding contextual text from layer 1 of the model to use to fine-tune in order to produce an answer.

4 Experiments

In order to measure our final results for the models, we sample our test set to construct a 50 question ”exam paper”. The exam paper was then inputted to each of the models which all returned 50 answers. The time taken to do this was recorded to give a measure of computational expense for each model. We created a mark scheme for the ”exam paper”, which was then used to manually calculate the raw mark for each model. This raw mark was then used as an estimate for the accuracy of the model for answering GCSE biology standard questions.

Our baseline model for the comparison was feeding the entirety of the textbook into Layer 2. This is a good control as no passage selection or retrieval has taken place, giving an accuracy and computation expense to aim to surpass.

An assessment of the computation time for the number of sub-sections given to ALBERT in layer 2 was conducted. The computation time is important given the task is to answer Biology GCSE standard questions, of which the exams are generally under timed conditions. The number of sub-section passages inputted was incrementally increased and the computation time for ALBERT to answer a single randomly selected question was measured.This experiment was run from inputting 1 sub-section to inputting all 75. The computation time was calculated 4 times for each input (each with a random question) and then averaged.

The effect of the length of the contextual text given to ALBERT in layer 2 on the accuracy of question answering was also explored. Due to the amount of ’marking’ needed to calculate the accuracy for the different input lengths an automatic marking system was created. This system is not as accurate as marking it manually but gives a close estimate of the real accuracy. It works by searching for keywords in an answer. For example, the auto-marker searches for the keywords ’protein synthesis’ in the output to the question ’What is the purpose of a chromosome?. If the output has the keywords in it is marked as correct otherwise it is marked as incorrect. We selected 42 questions relating to one sub-section, then the text for this sub-section along with the questions was fed into layer 2, the outputs were then marked giving an accuracy. This was then repeated with the same questions but the text padded by an increasing amount of text from other sub-sections. The accuracy was recorded with padding of size 0, 5, 10,..., 50. The padding was stopped at 50 due to the increased computation time of ALBERT with the input text.

To calculate the accuracy of the BM25 layer 1, the full test set was used. Whether the algorithm had correctly identified the correct sub-section a question belonged to is a binary statistic, and was summed for all the questions. The percentage correct could then be calculated as an estimate for the accuracy in sub-section classification.

RoBERTa accuracy of the layer 1 main-category, section and subsection classification will be provided through creating a train validation split. Therefore The final performance of the model's passage retrieval will be the validation accuracy. The overall accuracy of the hierarchical model correctly classifying a subsection, S may be formalised as: $A(Main) \times A(Section|Main) \times A(Sub - Section = S|Section)$

Where $A(Model)$ represents the accuracy for a single model

5 Results and Discussion

As our investigation consists of multiple analyses, we compile a review of performance for each model and aim to examine the reason behind our findings. This section discusses 3 groups of results. First, we will look at section classification results of the RoBERTa and BM25, critiquing the mechanisms respectively applied. Secondly we will analyse the limitations of the baseline ALBERT model justifying the requirement of the 2 Layer system. Finally, we will review the overall accuracy and completion time of the biology papers sat by the different section classification models using ALBERT for the 2nd layer, followed by a discussion of the overall results.

5.1 RoBERTa Section Classification

As aforementioned, 4 different approaches were employed in our pure NLP Layer 1, all utilising the RoBERTa model. In the process of our analysis, we found an optimum learning rate of 1e-5 for each RoBERTa model to reach convergence. Each approach; *RoBERTa-Main*, *RoBERTa-Section*, *RoBERTa-Subsection* and *RoBERTa-Hierarchy* gradually decreased in training and validation accuracy. This was anticipated in our study, as each approach required different subsets of the main data set. From Table 1. we outline the RoBERTa-Main classifier to have the best performance, with a final validation accuracy of 73. With the highest frequency of data-per-label. We also see a slight increase in performance of the hierarchical model over the subsection model, from 2 to 16%.

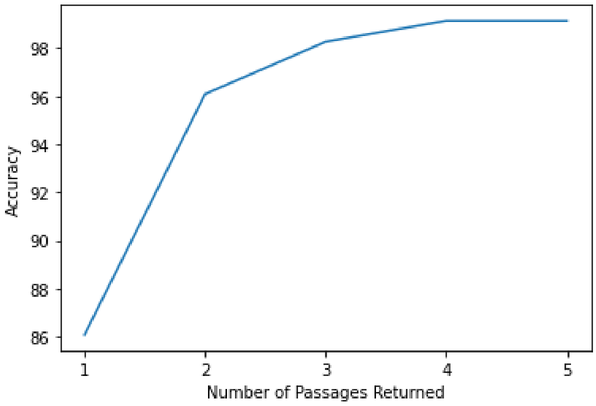
RoBERTa Model	Validation Accuracy (%)	Average No. Data Points per Class
Main	73	205.5
Section	42.6	68.5
Subsection	2	11.3
Hierarchy	Avg. 16	N.A.

Table 1: RoBERTA Layer 1 Section Classification Accuracy's and Dataset sizes (Hierarchy Average Taken from Combination of Classifier Accuracy's)

We see final validation accuracy's decrease for the subsequent branches in the hierarchy. This could be expected due to smaller training dataset sizes as you descend the hierarchy. Albeit, we emphasise the limitations in the small dataset size we draw on and insist the utilisation of a larger dataset will likely improve performance of each RoBERTa model (Roh et al., 2018). We also hypothesise that due to the questions being shorter than that of a typical comprehensive exam, accuracy suffers. We propose that questions containing more tokens would provide more context to boost sequence classification performance.

5.2 BM25 Section Classification

Figure 4: BM25 n Passage Subsection Classification Accuracy (1: 86.1%, 2: 96.1%, 3: 98.3%, 4: 99.1%, 5: 99.1%)



There are 6 models which use the BM25 algorithm for the first layer. Each differs by how many passages the BM25 algorithm returns, and therefore how many passages are fed into the 2nd layer. The highest classification accuracy achieved was 99.1% through the 4 and 5 passage model. Figure 4 shows that the more passages that BM25 is allowed to return, the better the sub-section classification accuracy is. This is only true up to returning 4 passages and more where increasing the number of passages has little to no effect on the classification accuracy. This is because the keywords in a question will often only appear in around 4 or 5 texts and therefore be easily narrowed down by the weighting scheme.

5.3 Limitations of ALBERT

To complete the test paper the QA system ALBERT is required, we now discuss the model's limitations. Figure 5 shows how the number of passages concatenated and given to the ALBERT layer 2 affects the computation time. As our task is to answer GCSE biology questions, it seems important that our overall model can answer questions quickly. The figure demonstrates the inefficiency of feeding the model the entire textbook and justifies the requirement of the two-layer system. From the graph, we can see that ALBERT took 154.81 seconds to answer 1 question with the whole textbook, scale this up to an exam of 100 questions or increase the number of text sources and the computation time will greatly increase.

Figure 5: Analysis of Layer 2 Computation Time

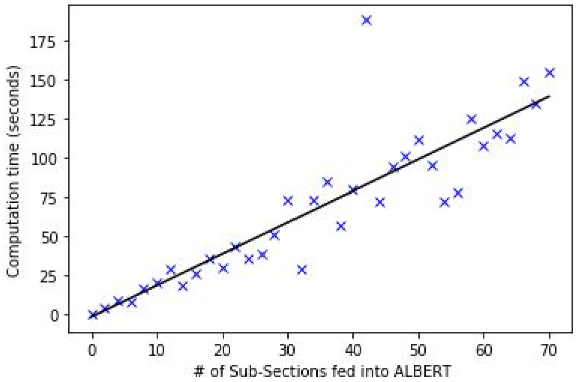
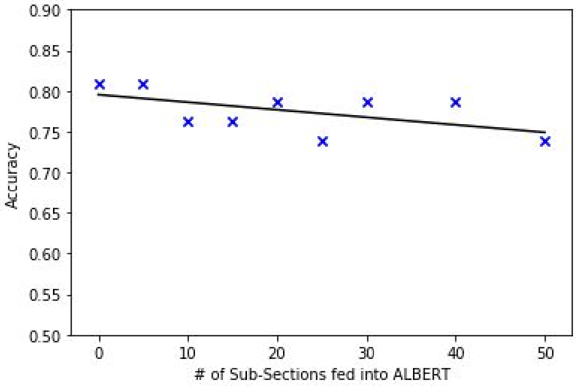


Figure 6 shows how increasing the number of input texts reduces the accuracy of the answer. The more text ALBERT is fed the more likely it chooses the wrong context to answer the question, hence making a mistake. This again shows why the whole textbook should not be inputted to ALBERT and layer 2 is necessary. However, the change in accuracy has a reasonably small gradient and therefore we observe a minimal difference in performance when models are given multiple of subsections. This idea is further discussed and quantified in section 5.3.

Figure 6: Analysis of Layer 2 Accuracy



5.4 Biology 50 Question Test Paper Results

The overall performance of the section classification models when combined with ALBERT to answer the 50 question exam paper is presented in Table 2. 'Test accuracy' refers to the number of questions answered correctly and 'computation time' refers to the time the model took to complete the test. The ALBERT baseline model scored 79% accuracy with a 10286s for completion time. RoBERTa's highest achieved test accuracy of 41% was that of RoBERTa-Main, and a final completion time of 853.12s, although accuracy performance was far worse compared to the baseline, completion speeds were far quicker. Table 2 shows that score accuracy of hierarchical models greatly decrease as you increase the granularity of your classifier, whereas the computation time considerably decreases. We claim from this result, stronger NLP passage-retrieval systems, trained on more sizeable training sets, may be realised to explore this accuracy-duration trade-off and can influence significant contributions to this area. Most BM25 passage models (2-5, 10) achieved far superior completion times along with equal or increased accuracy's when compared to the baseline. Similar to the results of Table 4, test accuracy is seen to increase with more passages supplied to ALBERT. However as evidenced in 5, the larger the contextual text inputted, the more expensive the computation. We, therefore, have a trade-off between the accuracy and computation time. As the classification accuracy plateaus around the return of 5 passages and the computation time is still reasonable, we justify that the best BM25 model is that with a layer 1 which returns 5 passages, giving a final accuracy of 83%. Any less passages will reduce the accuracy, more passages will increase the computational time as well as potentially decreasing the accuracy as seen in figure 5. This can be seen with the BM25 10 passage model being slightly less accurate with 82% than the 5 passage model. The final accuracy of 83% and completion time of 659s outperforms the baseline in both metrics.

Model (ALBERT Layer 2)	**Test Accuracy (%)**	**Computation Time (s)**
Baseline	79	10286.87
BM25		
1 Passage	77	536.65
2 Passage	79	548.24
3 Passage	79	586.90
4 Passage	81	614.92
5 Passage	83	659.61
10 Passage	82	944.34
RoBERTa		
Main	41	853.12
Section	8	309.92
Subsection	0	53.85
Hierarchy	0	83.85

Table 2: Model Accuracy and Computation Times

5.5 Results Discussion

The hierarchical NLP approach to section classification proved to be a difficult problem. The primary bottleneck was the lack of data points for the RoBERTa sequence classifier to train with. To allow the sequence classifier to train off of our biology dataset, around 15 questions were created and labelled for each subsection of text, this totalled to over 1000 questions for the entire corpus. After conducting the experiments it is clear far more labelled questions would be needed for the system to train and recognise the required passage. The hierarchical architecture did yield somewhat positive results, improving the accuracy from 2 to 16% (1). Better RoBERTa results were also shown by reducing the granularity of the dataset i.e. Main and Section models, however, this clearly showed a completion time trade off. Tf-idf techniques inherently specialise in passage retrieval from short texts and worked well in finding the needed passages in all experiments. It was also shown that concatenating passages of highly ranked text yields higher accuracy's at the cost of completion time. This was understood to be because the questions in the dataset only contained one or two keywords which could help identify the passage. We, therefore, propose the notion that a dataset comprising of longer and more complex questions would allow the RoBERTA model to potentially outperform tf-idf as context would be needed from the question to locate the needed section of text. Deep learning techniques used for section classification gives the primary advantage of low completion times. This is because the workload and performance overhead of training occurs prior to actual use. While accuracy levels are currently low, very high performance could be achieved with the changes mentioned made to the dataset.

6 Conclusion

Within our research, we touch upon the requirements necessary for accurate GCSE biology question-answering. Exploring this space foregrounded the influence of accurate passage retrieval in the trade-off between final test-score and computation-time. The model which scored the best was the BM25 5 passage model with a mark of 83% and a computation time of 659.61 seconds. The NLP model despite being the fasted model for classification lacked accuracy in comparison. The main reason for this was not only due to the small number of questions for each sub-section but also the length of each question.Additionally, this study provides a novel Biology GCSE data set used in the analysis, available for further development.A possible avenue for future work would be to build on the NLP layer 1 developed here by expanding the dataset in order to provide enough data to train the models effectively. This could be achieved by using multiple textbooks and spending time writing more questions. Another option for expanding the number of questions could be by creating a model to directly synthesise questions straight from textbooks. Another potential area for future work would be an exploration of a more general model which could perform over a wide variety of subjects possibly resulting in multiple classification layers.

References

Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016. A thorough examination of the CNN/daily mail reading comprehension task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2358–2367, Berlin, Germany. Association for Computational Linguistics.

Peter Clark, Oren Etzioni, Daniel Khashabi, Tushar Khot, Bhavana Dalvi Mishra, Kyle Richardson, Ashish Sabharwal, Carissa Schoenick, Oyvind Tafjord, Niket Tandon, Sumithra Bhakthavatsalam, Dirk Groeneveld, Michal Guerquin, and Michael Schmitz. 2019. From 'f' to 'a' on the n.y. regents science exams: An overview of the aristo project.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A. Kalyanpur, Adam Lally, J. William Murdock, Eric Nyberg, John Prager, Nico Schlaefer, and Chris Welty. 2010. Building watson: An overview of the deepqa project. *AI Magazine*, 31(3):59–79.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2015. The goldilocks principle: Reading children's books with explicit memory representations.

Rudolf Kadlec, Martin Schmid, Ondrej Bajgar, and Jan Kleindienst. 2016. Text understanding with the attention sum reader network.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Panupong Pasupat and Percy Liang. 2015. Compositional semantic parsing on semi-structured tables.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text.

Yuji Roh, Geon Heo, and Steven Euijong Whang. 2018. A survey on data collection for machine learning: a big data – ai integration perspective.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension.

Radu Soricut and Zhenzhong Lan. Google ai blog: Albert: A lite bert for self-supervised learning of language representations. `https://ai.googleblog.com/2019/12/albert-lite-bert-for-self-supervised.html`.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing.

A Appendix

Data set available at: `https://github.com/DhenPadilla/KSQuAD`