A Comparative Approach to Training a Question Answering Model Using Transformers

MSIN0221 Natural Language Processing Group Assignment — Group 5 — 3,694 words

Kayi Yeung
University College London
ka.yeung.21@ucl.ac.uk

Julie Woo
University College London
lisa-julie.woo.21@ucl.ac.uk

Alysa Li
University College London
wai.li.21@ucl.ac.uk

Esther Tong
University College London
ching.tong.21@ucl.ac.uk

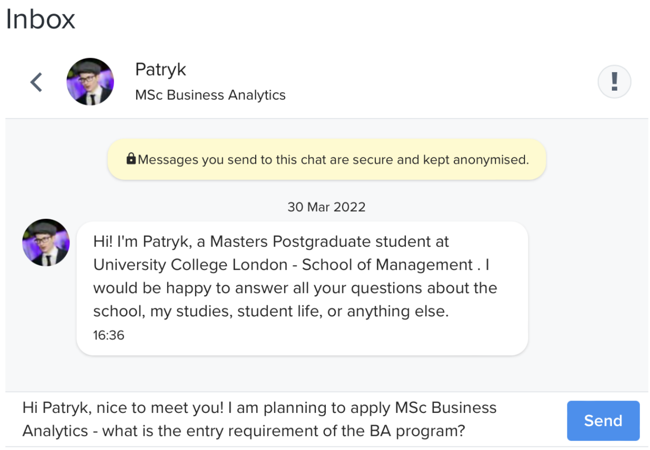
Russell Chou
University College London
russell.chou.17@ucl.ac.uk

Patryk Sobczak
University College London
patryk.sobczak.17@ucl.ac.uk

Akshat Biyani
University College London
akshat.biyani.21@ucl.ac.uk

1 Introduction

Chatbots are widely used across many industries and websites today, including on the UCL School of Management (SoM) website. The SoM currently appoints 12 student ambassadors from undergraduate and postgraduate studies every academic year to answer questions from prospective applicants via a chatbot.



🔒Messages you send to this chat are secure and kept anonymised.

30 Mar 2022

Hi! I'm Patryk, a Masters Postgraduate student at University College London - School of Management . I would be happy to answer all your questions about the school, my studies, student life, or anything else.
16:36

Hi Patryk, nice to meet you! I am planning to apply MSc Business Analytics - what is the entry requirement of the BA program?

Fig. 1. A snapshot of the SoM chatbot. A default greeting from the student ambassador is sent to prospective students who can then ask questions about the program, student life, or department.

Ambassadors are encouraged to reply within 3 working days depending on their availability and UCL pays them an hourly compensation of £15.66 in return. Over 30,000 enquiries have been received since March 2018, i.e. each ambassador receives 2 questions on average per day (UCL School of Management, 2022). After consulting one of the ambassadors, it is understood that around 70% of the questions were regarding generic admission requirements and programme information.

Assuming each question takes one minute to answer on average, a total of 87.5 hours per year (30,000 * 70% / 60 minutes / 4 years) could be saved by automating responses for these generic questions while student ambassadors can focus their efforts on more complicated questions. This can also lead to savings on operating cost by £1,370.25 per year (87.5 hours * £15.66 /hour).

In theory, the generic questions could be answered by a pre-trained language model that encodes the context, question and answer's start and end indices in the context, and then decode the answer. In practice, we test this hypothesis using three transformers - BERT-base, BERT-large and Electra-base to explore the capability of the model to answer questions accurately based on the programme information on the UCL SoM website. We find that the models need shorter answers in the training set due to the way they were trained. We then shorten the actual answers and retrain for several rounds.

This paper makes the following contributions:

1) We introduce the approach of automated responses to SoM's chatbot.
2) We train and compare several baselines models for the task and provide evaluation metrics.
3) We provide recommendations on the application and detail the challenges.

2 Methodology

2.1 Data Collection

Two main datasets are collected for this project - programme information and question answer sets. Given computing resource limitations and efficiency considerations, the focus was put on two postgraduate programmes in the School of Management - MSc Finance and MSc Business Analytics.

Programme Information It is the programme information available on the UCL School of Management website (https://www.mgmt.ucl.ac.uk/study). We obtained permission from the school before performing the web scraping, and the

Beautiful Soup library was used to extract the data based on programme names. Afterwards, a list of desired HTML tags was targeted within the content section to get a full corpus for each degree. Each corpus had a set of cleaning steps applied to it to remove the undesired formatting, which are detailed in 2.2 Data Preparation.

Question & Answer Dataset The data was collected in two ways:

1) Questions received via UCL SoM chatbot, including past conversations between student ambassadors and prospective students.
2) Questions were generated by our team members, which each comes up with questions that answers are present in the programme information pages.

Around 50 unique Question and Answer sets for each postgraduate course were generated, the questions were then scaled up by slightly tweaking the wording and sentence structure. By doing so we are providing a larger Q&A database, to learn the questions and answers' patterns and keywords, in hopes of a better performing model. In total, 1,500+ and 500+ question-answer pairs were generated for MSc Finance and MSc Business Analytics respectively.

2.2 Data Preparation

After extracting and generating the required data, we conducted checks on its quality and identified the following data cleaning steps.

Programme Information Two steps are performed:

1) *Format Cleaning:* the text corpus scrapped is in its raw HTML format containing various unwanted formatting, e.g. additional spaces, non-breaking space characters, use of special characters etc. To ensure the corpus can be tokenized and interpreted by the model accurately, we have created a script to clean the corpus for the mentioned issues.
2) *Corpus Splitting:* we found that the size of each corpus is too long with over 2,000 words, which could cause issues during training due to the maximum context token length limitation of transformers. Therefore, each corpus was further split into a smaller context which was achieved by saving each HTML paragraph tag into a separate context file per degree. This resulted in 106 contexts for MSc Business Analytics and 95 contexts for MSc Finance.

Question & Answer Dataset Two steps are performed:

1) *Answer Matching:* the answers prepared for each question must appear exactly the same in the corpus, including spacing, capitalization and punctuation; therefore cross-checking was essential to ensure consistency.
2) *Answer Indexing:* the answer is fed into the model using its corresponding start and end indices in the corpus. Therefore, a script was developed to find the exact match of the answer and the indices within the given degree corpus, to add columns of the start and end index to store the two indices and to save all the data into a new file, which was formatted and prepared for training and testing of models.

2.3 Model Selection

Three transformers have been selected for the task - BERT-base, BERT-large and Electra-base for training our QA datasets. BERT is chosen because of its great capability on question and answering and the pre-training on the SQuAD dataset, while ELECTRA works well on a smaller scale of computational power, which can be trained on a single GPU core system to produce a better accuracy than the GPT. All of them are open-source and publicly available in the *Hugging Face* community.

TABLE 1
TRANSFORMER SPECIFICATIONS

	Transformer		
	BERT-base (baseline)	*BERT-large*	*ELECTRA-base*
Model Name	deepset/bert-base-cased-squad2	bert-large-uncased-whole-word-masking-finetuned-squad	deepset/electra-base-squad2
# Parameters	110 million	336 million	110 million
Dataset	SQuAD 2.0	SQuAD 2.0	SQuAD 2.0
Cased/Uncased	Cased	Uncased	-
Tokenization	Character-level	Word-level	Character-level

We have previously selected *BERT-tiny-5-finetuned-squadv2* as our model in the proposal, because it was trained on the SQuAD dataset and is designed to use fewer computing resources. However, it was discovered at the beginning of our model training process that it has limited capability in comprehending corpus with long sequence length. Provided that our text corpus consists of over 2,000 words (30,000 characters), it was decided to replace it with BERT large, and have BERT base as the baseline mode, and Electra base as a third model for comparison as it provides faster training speed and a higher accuracy score in theoretical understanding.

2.4 Model Training

We kick start by training the three transformers using most of the default parameters except batch size and number of epochs because of our relatively small training dataset with 1,633 questions, and evaluate their performance based on three metrics which are detailed in 3.1 Evaluation Metrics.

Below are the details of batch size and epochs:

Following this, we plot the training loss and validation loss over time to learn how well the model is fitting the training data and how well it fits new data respectively. The goal is to find out the optimal number of epoch as we do not set an Early Stopping function, so as to mitigate overfitting and increase generalisation capability on new data.

We observe that the following for the three models:

TABLE 2
INITIAL HYPERPARAMETERS

	Transformer		
	BERT-base (baseline)	*BERT-large*	*ELECTRA-base*

1st Training	batch_size = 8
	epoch = 8
		epoch = 8	batch_size = 8
			epoch = 4

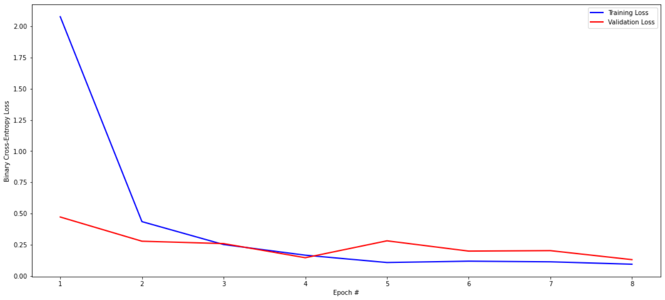


Fig. 2. BERT-base model training and validation loss graph

- *BERT-base*: With the number of 8 epochs, the optimal number is 4 as the model starts overfitting from the 5th epoch. Therefore, we will stick to it in the fine-tuning rounds.
- *BERT-large*: The validation loss is not higher than the training loss until the 7th epoch, but it is hard to tell if it is already the optimal number without running more epochs. Therefore, we have an exploration step to run more epochs or leverage open source pre-trained trainers.
- *ELECTRA-base*: It is not within expectation that the validation loss is smaller than the training loss. The fact that validation loss is lower than training loss could indicate the model has difficulty to learn the patterns in the training dataset for prediction. The plan is to increase the number of epochs in the optimization rounds to get more insights.

2.5 Model Optimization

We decide to adopt different approaches to optimise the models based on the analysis above.

1) **Batch Size and Epoch Combinations:** The BERT-base and ELECTRA-base models use this tactic based on the findings above.
2) **Pre-trained Open Source Trainer:** The BERT-large model utilises the BERT Miniatures, which is within the set of 24 BERT models referenced in Well-Read Students Learn Better: On the Importance of Pre-training Compact Models (Turc, Chang, Lee and Toutanova, 2022). It is intended for limited computational resources and is designed for fine-tuning in the same manner as the original BERT models.

In total, there are 3 rounds of fine-tuning and below are the hyperparameter details.

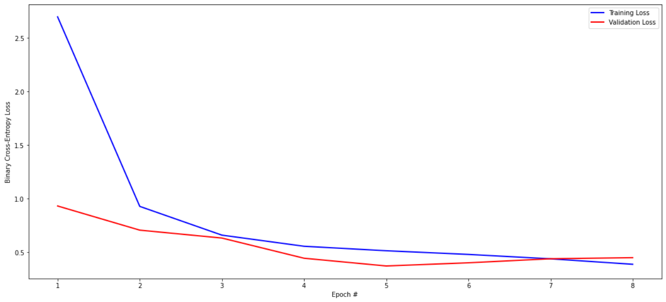


Fig. 3. BERT-large model training and validation loss graph

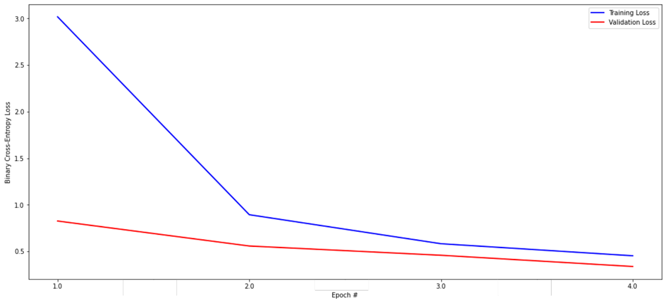


Fig. 4. Electra-base model training and validation loss graph

3 EVALUATION

3.1 Evaluation Metrics

When testing the models on the validation dataset, another script is written to store the answers generated by the model and the confidence score in a table, alongside the questions and the actual answers. In order to gain a better understanding of the three models' performance, strengths and weaknesses, three metrics have been defined for evaluations - F1 score for Word-overlap, Accuracy Score for exact correct answers, and Accuracy score by manual evaluation, i.e. a human compares the two answers and determine if the model provides a TRUE/FALSE answer.

Matrix 1: Accuracy Score for Exact Correct Answers The answers generated by the model are compared to the actual answers, to determine whether it is an exact match. A new *comparison* column is created in the table to store the result. It is calculated by the number of exact answers by the total number of questions.

Matrix 2: Accuracy Score by Manual Evaluation It is noted that some model generated answers were indeed contextually correct, however they were not exactly the same, word by word when comparing to actual answers. This has an immense effect on model performance when evaluation is made with Matrix 1, hence results were manually assessed by the team members with cross-validation rounds to produce a second accuracy score. It is calculated by the number of contextually correct answers by the total number of questions.

TABLE 3
HYPERPARAMETER TUNING

	Transformer		
	BERT-base (baseline)	*BERT-large*	*ELECTRA-base*

1st Optimization	batch_size = 16
	epoch = 4
		google/bert_uncased_L-2_H-128_A-2	batch_size = 16
			epoch = 4

2nd Optimization	batch_size = 32
	epoch = 4
		google/bert_uncased_L-4_H-512_A-8	batch_size = 16
			epoch = 8

3rd Optimization	batch_size = 16
	epoch = 8
		google/bert_uncased_L-12_H-768_A-12	batch_size = 32
			epoch = 13

Matrix 3: F1 Score for Word-Overlapping In an effort to statistically measure the model performance, the F1 score for word overlapping has been calculated. The more words that are overlapped between model generated and actual answers, the higher the F1 score would be. It is calculated by taking the mean of the F1 scores of all questions.

3.1.1 *Evaluation Metric Selection*

To determine which matrix is the best for performance evaluation, we calculate each of them for the 1st training round as below.

TABLE 4
MODEL EVALUATION METRICS

	Transformer		
	BERT-base (baseline)	*BERT-large*	*ELECTRA-base*
Matrix 1 - Accuracy Score for Exact Correct Answers	0.0098	0.41	0.0049
Matrix 2 - Accuracy Score by Manual Evaluation	0.26	0.40	0.28
Matrix 3 - F1 Score for Word-Overlapping	0.26	0.44	0.22

By looking at matrices 1 and 2, we observe that the Accuracy Score for Exact Answers does not seem to be a good measure because the BERT-base and ELECTRA-base models get very low scores compared to BERT-large, whereas the Accuracy Score by Manual Evaluation shows otherwise. For matrix 3, it is fairly similar to Matrix 2 and is the highest for two of the models.

After careful consideration, we decided to use **Matrix 3: F1 Score** for Word-Overlapping as the primary KPI because it is able to capture the model performance in a statistical and automatic way, no matter whether the answers were exactly correct or not, and it is more efficient than manual evaluation.

3.2 *Performance Comparison*

Below summarizes the **F1 Score for Word-Overlapping** on the validation dataset for each model in the 4 rounds of training and fine-tuning.

TABLE 5
F1 SCORE FOR WORD-OVERLAPPING (VALIDATION DATASET)

	Transformer		
	BERT-base (baseline)	*BERT-large*	*ELECTRA-base*
1st Training	deepset/bert-base-cased-squad2	bert-large-uncased-whole-word-masking-finetuned-squad	deepset/electra-base-squad2
1st Fine-tuning	0.26	**0.44**	0.22
2nd Fine-tuning	0.34	**0.47**	0.23
3rd Fine-tuning	0.23	**0.46**	0.23
Tokenization	0.27	**0.44**	0.23

Based on the above, we find that BERT-large outperforms the other two models in all rounds with the best one on the 1st round of fine-tuning using the pre-trained model *google/bert_uncased_L-2_H-128_A-2*, achieving an F1 word-overlap score of 0.47. The BERT-base model also has similar findings as BERT-large, i.e. the 1st fine-tuning round is the most outstanding. ELECTRA-base, however, does not have obvious performance fluctuation in all rounds.

Given the leading performance of BERT-large, it is decided to run the test dataset on two of the best-performing fine-tuned models, i.e. the 1st and the 2nd round.

TABLE 6
F1 SCORE FOR WORD-OVERLAPPING USING BERT-LARGE

		F1 Score for Word-Overlapping	
Transformer	**Hyperparameters**	**Validation**	**Test**
BERT-large	google/bert_uncased_L-2_H-128_A-2	0.47	0.39
BERT-large	google/bert_uncased_L-4_H-512_A-8	0.46	0.39

The two models obtain the same F1 Score for Word-Overlapping on the test set, which is 0.39. It is slightly lower than that of the validation set, which is within expectation when a model explores unseen data.

3.3 *Error Analysis*

To further understand the possible reasons the model generates correct and wrong answers, we perform error analysis on the validation set for each model after the 1st training.

BERT-base model The model only gets very few exact correct answers and the generated ones can be ambiguous and only partially correct after validating by team members, showing a very different pattern with Bert Large. The model has a clear struggle on a continuous sentence divided by a full stop or comma, which can be reflected by Q1 in Table 7 as it only returns one word from the actual answer. Also, it is not capable of generating correct answers when it is too long and is separated in more than 1 sentence - Q2 is an example. Lastly, when the keyword on the question and answer is not self-descriptive enough, it has difficulty to give a meaningful answer. In Q3, the model struggled to understand what ”quantitative” is referring to, and fail to provide correct answers; unlike other questions in the validation set, with easy-to-understand keywords like ”start date”, ”deadline” and ”campus” etc.

BERT-large model It performs exceptionally well for questions related to tuition fees and dates. It is noted that, if the question and answer contain the exact keyword, the model tends to get the answers correctly. There is also this pattern that the answers generated from the model are either exactly the same or completely wrong, meaning there is not even a word overlapping on the actual and generated answers.

The model struggles when there is no common keyword between questions and answers which is indicated in Q1 in Table 8. Another pattern is when the answer is a continuous sentence divided by a full stop or comma, the model fails to find the correct answer or give the complete answer - Q2 is an example of the former and 15% of the validation set suffers this type of failure; Q3 is an example of the latter which only returns the first chunk as an answer.

Electra-base model Within Table 9, Question 1, the model returns an incorrect answer. It can be assumed that the model thinks the word “student” is related to the entry requirement and returns the answer for the course requirement as the intake. In the second case, the model fails to answer the question correctly. A possible reason might be the word ‘minimum requirement’, which can apply to other requirements, not only the English language requirement. In Q3, the model seems to be confused with rhetorical questions since it can be tricky to process two sentences from the question. Therefore, it only returns the non-related answers on the model answers.

4 Limitations

There are multiple limitations that need to be taken into account for this project

Limited Dataset Size The Questions and Answer dataset used in this project contains around 2,100 rows of data, however a more extensive dataset of at least 3000 rows per programme would have been favourable, to provide more training data for the models, in hopes to increase the accuracy and F1 scores.

TABLE 7
BERT-base detailed error analysis

	Question	Correct Answer (from context)	Model Answer	Possible error justification
1	What academic background are you looking for, for MSc Finance?	economics, finance, mathematics, mathematical economics, econometrics, economic theory, statistics, engineering, or any combination of thereof	mathematics	The model does not understand the association of words that are separated by commas.
2	Do you have a list of documents that are mandatory for the application of MSc BA Programme?	You will need to include a degree transcript in one of the suggested disciplines, two references, and a personal statement. Overseas applicants should also supply evidence of their English language ability usually via an IELTS test.	Overseas applicants should also supply evidence of their English language	The answer is too long and separated in two sentences, which confuses the model of the actual answers.
3	Do I have to be very quantitative for MSc Finance?	we stress to our applicants that this programme is highly mathematical and quantitative, so you should be prepared to be pushed in these areas.	prepare	The model struggles to understand the question and the answer's context. It produces a one-word answer that is close to the keyword but is not meaningful on its own.

Model Limitation The original project idea is that - once the model is trained with a smaller corpus, it can take general program questions about any programme in UCL SoM; and by using ELMo word embeddings and text similarity, the question would be matched to its respective full corpus as context, and in turns return a correct answer. However, during the first round of model testing, it becomes apparent the maximum context token length limitation not

TABLE 8
BERT-LARGE DETAILED ERROR ANALYSIS

	Question	Correct Answer (from context)	Model Answer	Possible error justification
1	What is the desired academic background for MSc Finance?	economics, finance, mathematics, mathematical economics, econometrics, economic theory, statistics, engineering, or any combination of thereof	thereof	The keyword in the question is 'Background', but this word is not present in the answer.
2	Is it required to have some sort of work experience or education background in the finance and finance-related sector for the Finance master?	Demonstration of an appropriate academic background is essential. Experience in the finance and finance-related sector is beneficial but is not a prerequisite.	. (i.e. a full stop)	The question is too long and contains too many keywords that confuse the model.
3	What academic background are you looking for, for MSc Finance?	economics, finance, mathematics, mathematical economics, econometrics, economic theory, statistics, engineering, or any combination of thereof	economics	The model does not understand the association of words that are separated by commas.

TABLE 9
ELECTRA-BASE DETAILED ERROR ANALYSIS

	Question	Correct Answer (from context)	Model Answer	Possible error justification
1	What are the main programming language that BA course used?	The core elements of the programme are delivered in Python. In other modules we also use; R, Stata and JavaScript but Python is the main language used.	Python.	The model only returns part of the answer because it is closer to the keyword, i.e. "main programming language".
2	What information do I need to provide for my MSc ba' application?	You will need to include a degree transcript in one of the suggested disciplines, two references, and a personal statement. Overseas applicants should also supply evidence of their English language ability usually via an IELTS test.	a personal statement. Overseas applicants	The answer is too long and separated in two sentences, which confuses the model of the actual answers.
3	For the master programme in Finance, is working experience essential?	Experience in the finance and finance-related sector is beneficial but is not a prerequisite.	Demonstration of an appropriate academic background is essential	The model struggles to understand the question's context and confuses the actual answer.

only affects model training - the model is simply incapable of reading a lengthy context, no matter how well-trained it is. After going through multiple documents, we have found out that most NLP Transformer models are only capable of reading a maximum of 512 tokens (words) in one corpus (Briggs, 2021), yet on average, the UCL SoM Programme page contains around 4000-4500 tokens; hence the model struggles to answer questions where the answers are from the lower part of a context. For the purpose of this report, the models are tested on a smaller corpus, but this approach would have to be refined for real-life QA chatbot applications.

Additionally, the models that are chosen are trained for Extractive QA, meaning the answer has to be exactly the same, word by word (spacing, punctuation and capitalization included) from the context. For a QA chatbot, this might not be the best methodology, as most answers are from the middle of a sentence, it would not be a full, grammatically correct sentence on its own. To overcome such limitations, abstractive models such as BART can be experimented with in the future.

It is noted that all models, including the best performing Bert-large, were performing comparatively poorly. In general, the models could have achieved an F1 accuracy score over 0.9. This is caused by our very specific and application driven dataset, and neither Bert nor Electra models were pre-trained for such purposes; which explains the models' subpar performance.

Lastly, the performance of the ELECTRA model seems to create an unusual pattern between training loss and validation loss. In theory, the validation loss should be greater than the training loss because validation error is a method of estimating generalisations without training the test set in the Neural Network. However, our research shows the opposite with training loss higher than validation loss. The cause of this can indicate the model has difficulty learning the patterns in the training dataset for prediction. To improve the performance, we can reassess our data splitting process or increase the dataset size as mentioned in the first limitation.

Computational Limitation The resources of computational support are also limited in this project. For example, the best computational system we can use is a GPU with 4 cores and 61 GB memory on faculty.ai, which may not be the best performer in terms of running time. If a more advanced computational system is available, it will be possible to run a larger dataset with more complex models, so that this research can be conducted with better time efficiency and higher accuracy.

Early Stopping for Epoch Currently the models rely on manual fine-tuning to find the optimal hyperparameters. This could be further integrated and automated into the model training process by adapting early stopping in the epoch process. Such automation can help to select the optimal number of epochs without human intervention and streamline the whole training and fine-tuning process.

5 Conclusion and Recommendation

To summarise, a Question and Answering model has been developed for answering generic information automatically on SoM's website, for the two programs - MSc Business Analytics and MSc Finance programmes. Three different NLP models - Bert Base, Bert Large and Electra Base have been trained and evaluated by F1 word-overlapping score and two kinds of accuracy scores to find out which is the best for SoM's chatbot. Overall, the Bert Large model (bert-large-uncased-whole-word-masking-finetuned-squad) is deemed to be the best performing model out of the three, with F1 word-overlapping accuracy scores of 0.48 on validation set and 0.39 on test set respectively. The result is not at the same level or even close to the F1 accuracy score of Bert Large reaches, which is 0.93 (Devlin, et al., 2018). This can be explained by the limitations detailed in the previous section.

Looking forward, three crucial areas have been identified for future improvement.

Firstly, the QA model needs to be trained on a much larger dataset. Our current best model - Bert Large, after being trained on 1,800 rows of Question and Answer sets, only achieved an F1 accuracy score of 0.48. It is suggested to aggressively increase the size of the Question and Answer dataset to at least 3,000 per programme, in order to improve model performance, to return smarter, more accurate responses; and achieve a higher f1 score.

Furthermore, restructuring program information on the SoM webpages is recommended as the information is not as well structured as one may expect. Details regarding one topic are scattered over multiple different paragraphs, which in turn confuses the model in respect of where to extract the correct answers or could be difficult for prospective students to gather all information. It is suggested that the website shall be reorganised in a more structured and coherent way to facilitate the chatbot. Alternatively, a separate script can be written based on the scraped website content and make adjustments accordingly.

Finally, it would be quintessential to expand the application of Chatbot to other UCL SoM programmes. For this project, the scope is limited to two master programmes - MSc Business Analytics and MSc Finance due to time and computing resource constraints. If the chatbot were to roll out onto the UCL SoM website, the model will need to be trained on QA database and corpus from all other programmes. The model also needs to be able to read any corpus, nevertheless the length; by automatically choosing the appropriate chunk of the corpus as context. Open Domain QA technique can be adapted - running cosine similarity between questions and each corpus chunk, and the model would only be deployed to the chunk with the highest similarity score.

REFERENCES

Akhtar, Z., 2021. BERT base vs BERT large. [Online] Available at: https://iq.opengenus.org/bert-base-vs-bert-large/ [Accessed 15 March 2022].

Bartolo, M., 2021. HuggingFace Electra-Large-synqa. [Online] Available at: https://huggingface.co/mbartolo/electra-large-synqa [Accessed 16 March 2022].

Bartolo, M. et al., 2021. Improving Question Answering Model Robustness with Synthetic Adversarial Data Generation. arXiv.

Bartolo, M., Tylinski, K. Moore, A., 2019. Pre-trained Contextual Embeddings for Litigation Code Classification. LegalAIIA Workshop.

Briggs, J., 2021. How to Apply Transformers to Any Length of Text. [Online] Available at: https://towardsdatascience.com/how-to-apply-transformers-to-any-length-of-text-a5601410af7f [Accessed 17 March 2022].

Briggs, J., 2021. How to Train Bert For QA in Any Language. [Online] Available at: https://towardsdatascience.com/how-to-train-bert-for-q-a-in-any-language-63b62c780014 [Accessed 1 March 2022].

Devlin, J., Chang, M.-W., Lee, K. Toutanova, K., 2018. Pre-training of Deep Bidirectional Transformers for Language. [Online] Available at: https://huggingface.co/bert-large-uncased-whole-word-masking-finetuned-squad [Accessed 10 03 2022].

Google Research, 2020. BERT-Tiny (5) fine-tuned on SQuAD v2, Hugging Face. [Online] Available at: https://huggingface.co/mrm8488/bert-tiny-5-finetuned-squadv2 [Accessed 2 Feburary 2022].

Goswami, D. S., 2020. Introduction to Early Stopping: an effective tool to regularize neural nets. [Online] Available at: https://towardsdatascience.com/early-stopping-a-cool-strategy-to-regularize-neural-networks-bfdeca6d722e [Accessed 10 March 2022].

HuggingFace, 2021. ELECTRA. [Online] Available at: https://huggingface.co/docs/transformers/main/model_doc/electra#transformers.ElectraForQuestionAnswering [Accessed 5 March 2022].

Khanna, C., 2021. Question Answering with a fine-tuned BERT. [Online] Available at: https://towardsdatascience.com/question-answering-with-a-fine-tuned-bert-bc4dafd45626 [Accessed 13 March 2022].

Kiela, D. et al., 2021. Dynabench: Rethinking Benchmarking in NLP. arXiv.

Mikic, F. A. et al., 2009. CHARLIE: An AIML-based chatterbot which works as an interface among INES and humans. In: 2009 EAEEIE Annual Conference. Valencia: IEEE, pp. 1-6.

Satu, M. S., Parvez, M. H. Shamim-Al-Mamun, 2015. Review of integrated applications with AIML based chatbot. In: 2015 International Conference on Computer and Information Engineering (ICCIE). Bangladesh: IEEE, pp. 87-90.

tagtog, 2022. The Text Annotation Tool to Train AI. [Online] Available at: https://www.tagtog.net/ [Accessed 21 January 2022].

The TensorFlow Authors, 2020. Model Maker Question Answer Tutorial. [Online] Available at: https://colab.research.google.com/github/tensorflow/tensorflow/blob/master/tensorflow/lite/g3doc/tutorials/model_maker_question_answer.ipynb#scrollTo=h2q27gKz1H20 [Accessed 1 March 2022].

The TensorFlow Authors, 2020. Train a QA model. [Online] Available at: https://colab.research.google.com/github/neuml/txtai/blob/master/examples/19_Train_a_QA_model.ipynb#scrollTo=4Pjmz-RORV8E [Accessed 1 March 2022].

Turc, I., Chang, M.-W., Lee, K. Toutanova, K., 2019. Well-Read Students Learn Better: On the Importance of Pre-training Compact Models. [Online] Available at: https://huggingface.co/google/bert_uncased_L-2_H-128_A-2 [Accessed 10 March 2022].

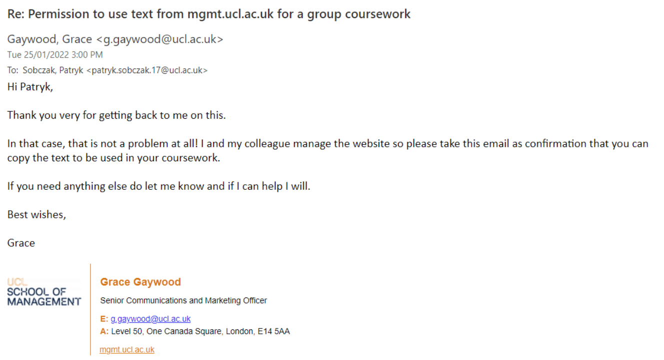
UCL Extend, 2021. UCLeXtend terms and conditions and website acceptable use policy. [Online] Available at: https://extend.ucl.ac.uk/admin/tool/policy/view.php?versionid=3&returnurl=https%3A%2F%2Fextend.ucl.ac.uk%2Fadmin%2Ftool%2Fpolicy%2Findex.php&numpolicy=1&totalpolicies=2 [Accessed 23 January 2022].

UCL School of Management, 2022. CHAT TO OUR STUDENTS. [Online] Available at: https://www.mgmt.ucl.ac.uk/chat-to-students [Accessed 23 January 2022].

UCL School of Management, 2022. STUDY. [Online] Available at: https://www.mgmt.ucl.ac.uk/study [Accessed 23 January 2022].

Appendix

Appendix A: *Email confirmation from School of Management for permission to scrape the website*



Re: Permission to use text from mgmt.ucl.ac.uk for a group coursework

Gaywood, Grace <g.gaywood@ucl.ac.uk>
Tue 25/01/2022 3:00 PM
To: Sobczak, Patryk <patryk.sobczak.17@ucl.ac.uk>

Hi Patryk,

Thank you very for getting back to me on this.

In that case, that is not a problem at all! I and my colleague manage the website so please take this email as confirmation that you can copy the text to be used in your coursework.

If you need anything else do let me know and if I can help I will.

Best wishes,

Grace

UCL SCHOOL OF MANAGEMENT

Grace Gaywood
Senior Communications and Marketing Officer
E: g.gaywood@ucl.ac.uk
A: Level 50, One Canada Square, London, E14 5AA
mgmt.ucl.ac.uk

Appendix B: *Links to the programme information pages used in the research.*

- MSc Business Analytics: http://www.mgmt.ucl.ac.uk/business-analytics
- MSc Finance: http://www.mgmt.ucl.ac.uk/finance

Appendix B: *Links to the models taken from Hugging Face*

1) BERT-base: deepset/bert-base-cased-squad2
2) BERT-large: bert-large-uncased-whole-word-masking-finetuned-squad
3) ELECTRA-base: deepset/electra-base-squad2
4) BERT Miniatures