CHLOE THOMPSON | DR BARRY DEVEREUX & DR JOANA CAVADAS

MACHINE LEARNING APPLIED IN THE CONTEXT OF QUESTION GENERATION AND QUESTION ANSWERING

INTRODUCTION

Project Aims

Aim is to leverage NLP techniques to develop a system that can consume text and return a set of comprehensive questions and answers

Question Generation (QG)

The process of generating questions based on a given context paragraph.

Question Answering (QA)

The process of generating an answer to a provided context paragraph and question.

SQuAD 2.0

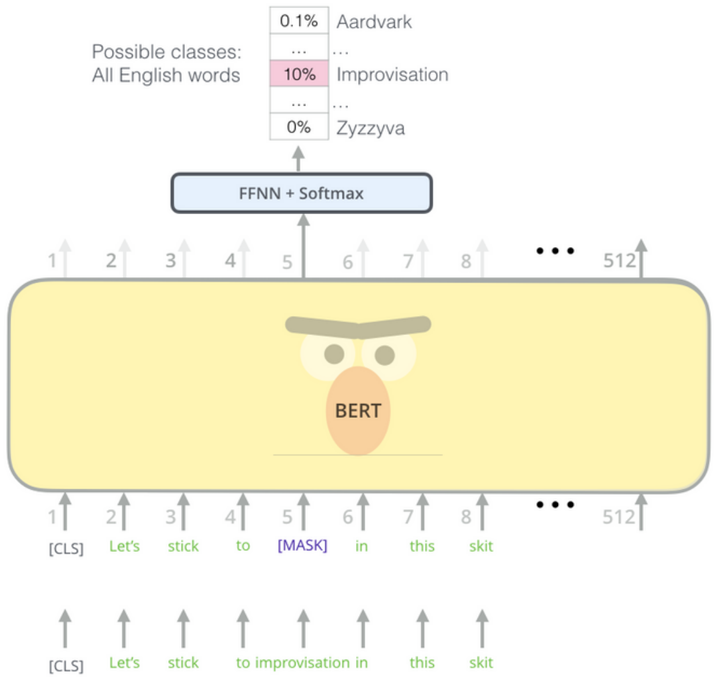
- Stanford Question and Answering Dataset.
- Created by crowdsourcing answers to over 100,000 questions on Wikipedia articles.
- Consists of context paragraphs, answer spans, answer text, questions and other data

QuAC

- Question and Answering in Context Dataset
- Built on the similar principle to that of SQuAD
- Contains extra style of questions such a multi-turn to have conversational question-answer dialogue
- Consists of context paragraphs, answer spans, answer text, questions and other data

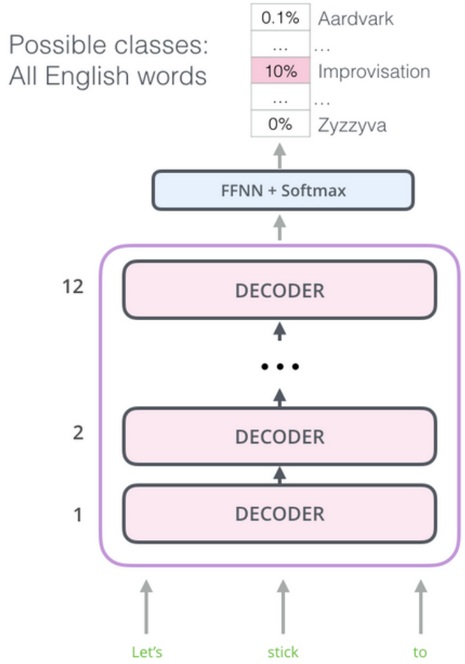
BERT

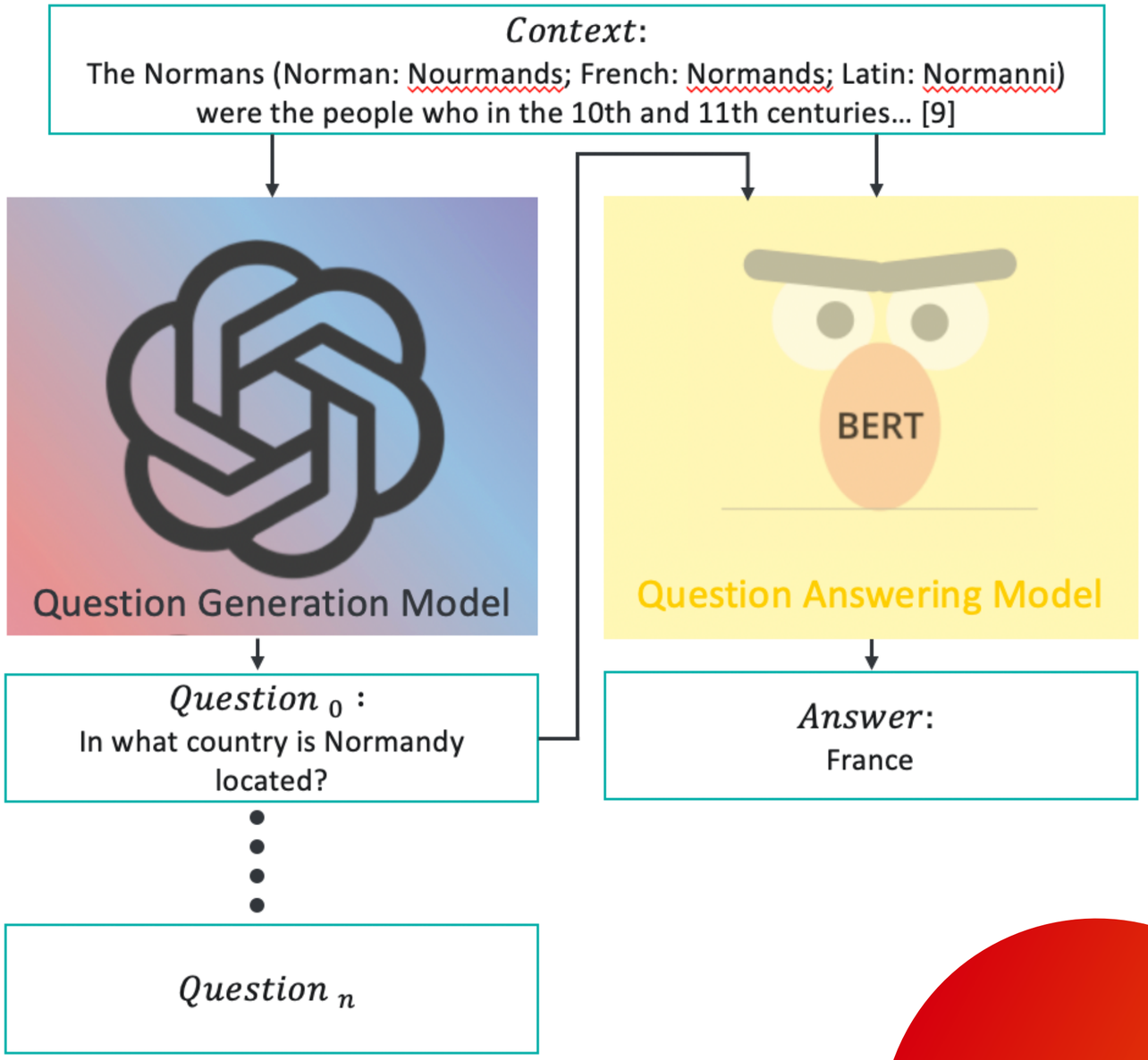
- A transformer model built on a number of encoder blocks.
- Two model sizes - Base and Large
- Trained to as a Masked language Model and for performing Next Sentence Predictions
- Specific encoder tokens



GPT-2

- A transformer model built on a number of decoder blocks.
- Four model sizes - 117 to 1542 Million Parameters
- Uses Masked Self Attention to mask tokens
- Produces one output token at a time as a generated word, that is then used for the next token.





METHODOLOGY

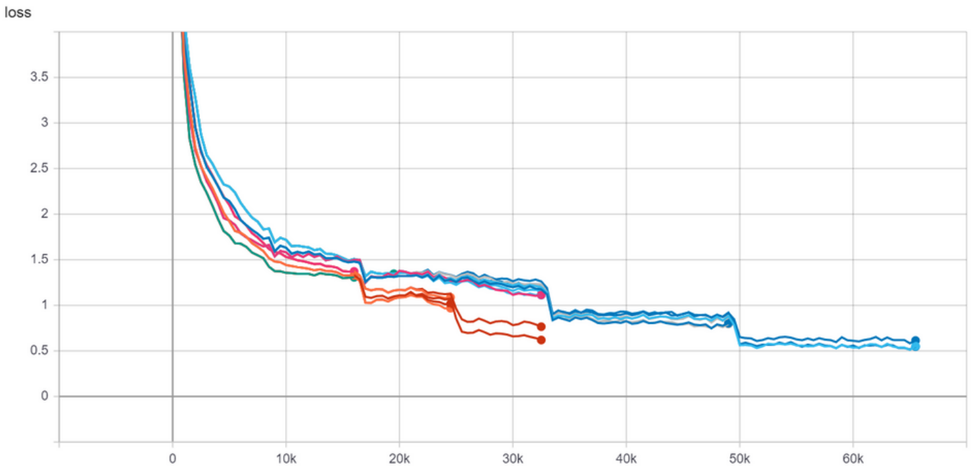
This project is based on a connected QG to QA pipeline.

The QA model can be used to both answer and evaluate the quality of generated questions.

QA TRAINING

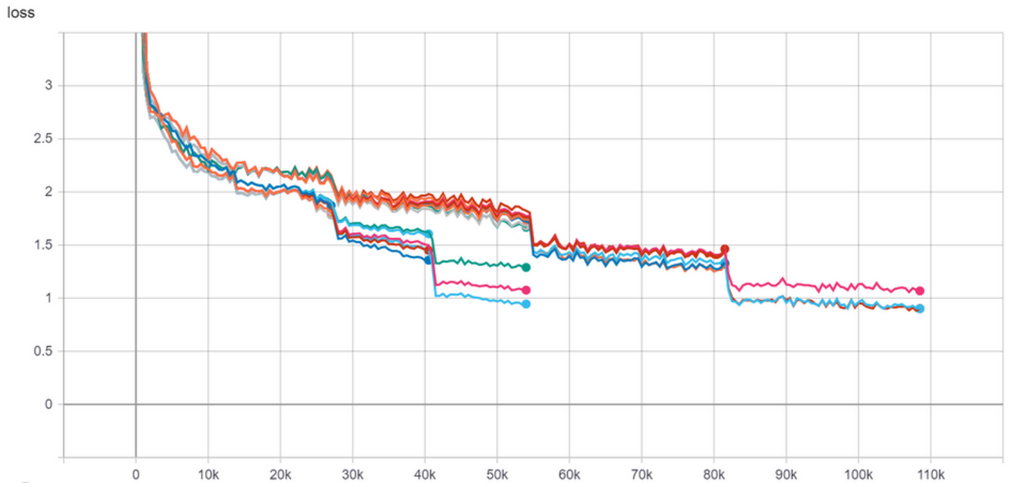
SQuAD

- Parameters were hypertuned using Grid Search - learning rate, epochs, and batch size
- Best results - F1: 71.84, EM: 68.17
- Achieved just below that of a leading paper in the space - F1: 76.7, EM: 73.85 - for a standard BERT implementation.
- Parameters {3e-05, 3, 2} created the best results.



QuAC

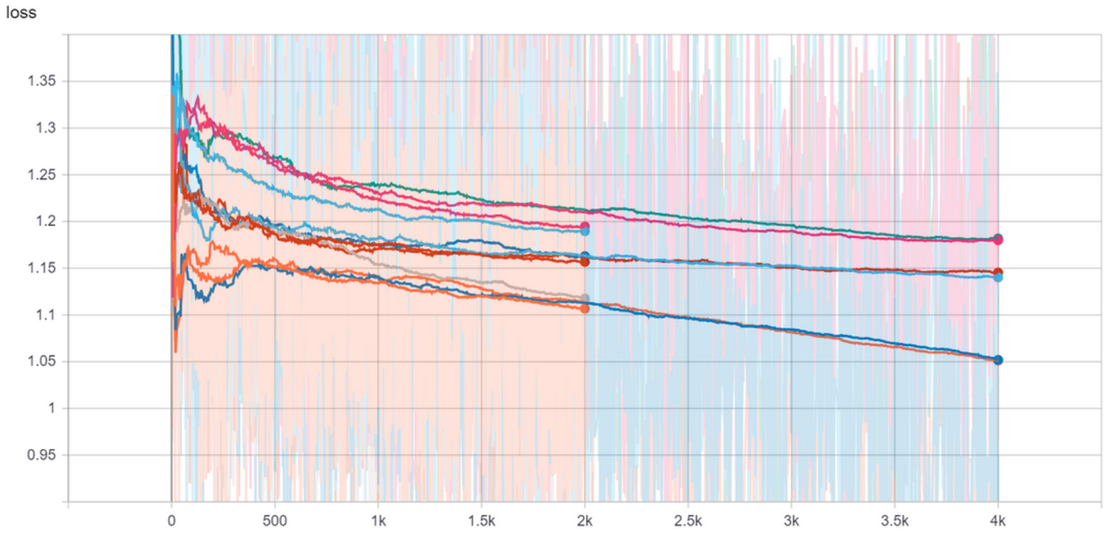
- Parameters were hypertuned using Grid Search - learning rate, epochs, and batch size
- Results were poor - F1: 39.74, EM: 25.78
- These results were higher than that of a similar paper that achieved F1: 33.3
- Further testing was done on hyperparameters, but no improvements could be made with current model



QGEN TRAINING

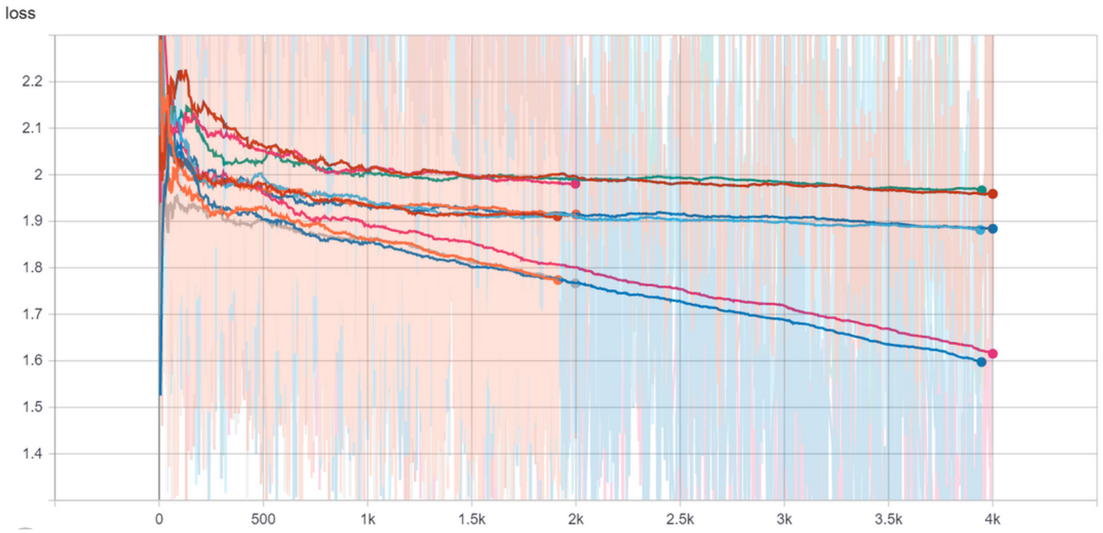
SQuAD

- Hypertuned Batch Size, Learning Rate and Steps with the same Grid Search approach
- Trained on p3.16xlarge instance in Sagemaker - 8 NVIDIA GPU 's taking 20-50 minutes per training
- Best results - loss of 1.17
- Parameters {6, 1e-4, 4000} created the best results in terms of loss

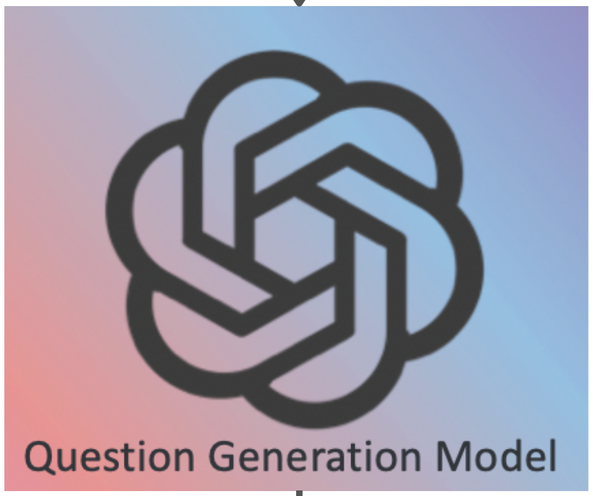


QuAC

- Hypertuned Batch Size, Learning Rate and Steps with the same Grid Search approach
- Trained on p3.16xlarge instance in Sagemaker - 8 NVIDIA GPU 's taking 20-50 minutes per training
- Best results - loss of 0.949
- Parameters {6, 1e-4, 4000} created the best results in terms of loss



<|startoftext|> [CONTEXT]:
Context Paragraph
[QUESTION]:



<|startoftext|> [CONTEXT]:
Context Paragraph
[QUESTION]: Ground Truth
Question <|endoftext|>

QGEN GENERATION

- Input a tokenised context that is to have a question generated.
- The QGen model generates a question up to a certain length between the [QUESTION]: and <|endoftext|> tokens.
- The creativity of the question generated depends on the temperature of the model

BLEU

- Bi-Lingual Evaluation Understudy
- Analyses the quality of machine generated text by comparing against human judgement.
- Achieved a correlation of 0.817.
- It is a form of modified precision based on the n-grams in a sentence.

Meteor

- Metric for Evaluation of Translation with Explicit ORdering
- An improvement on the BLEU metric to adapt to its shortfallings
- Achieves a correlation between sentences that is closer to human judgement, 0.964.

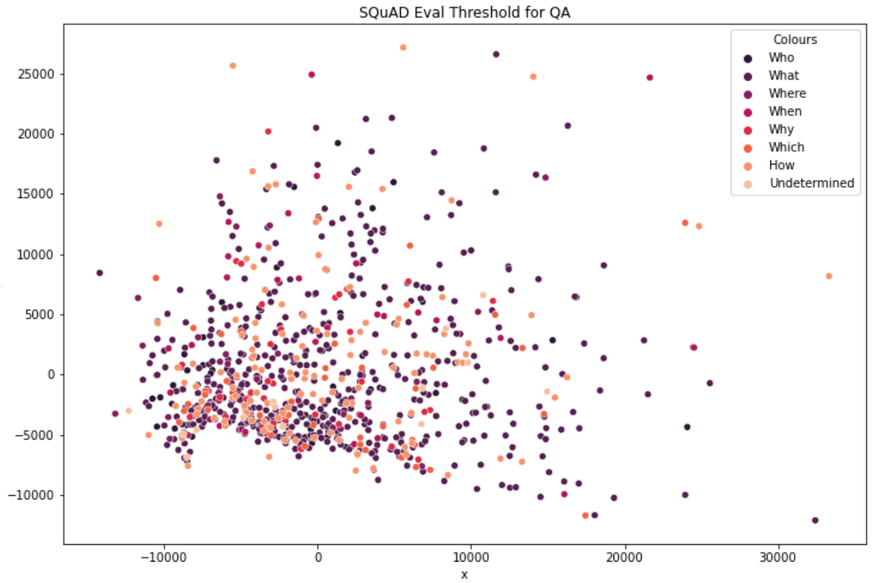
EVALUATION OVERVIEW

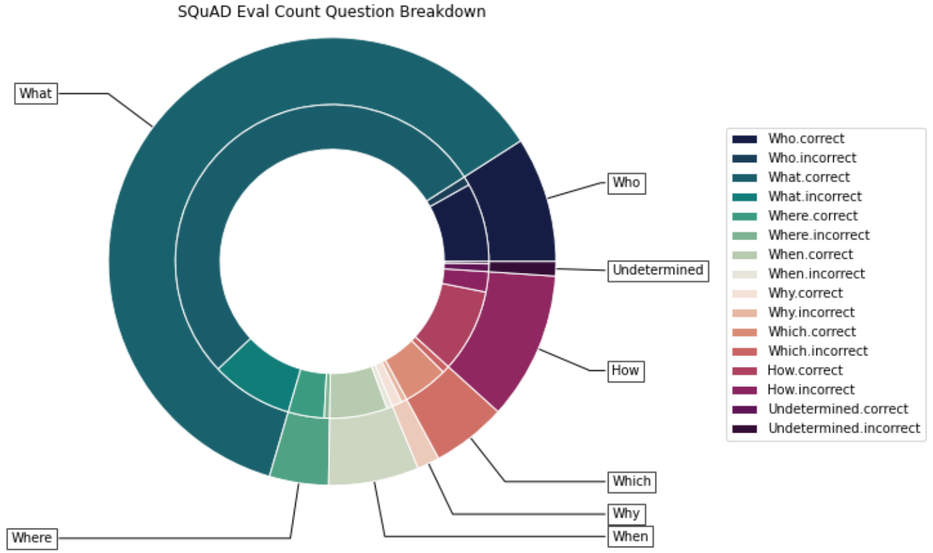
- QA on SQuAD and QuAC evaluation checks the metrics between the predicted answers and ground truth answers
- QG are evaluated with the metrics using the generated questions and all ground truth questions for the context paragraph
- Data with missing context or missing questions are excluded
- Evaluate the number of blank predicted answers and those that are the [CLS] token as these are not good answers

Type	Average BLEU	Average Meteor	Num below Threshold	% of Total	Num of [CLS]	Num of blanks
SQuAD	0.035	0.024	839	0.142	317	114
QuAC	0.038	0.021	4042	0.55	1641	525
SQuAD Gen	0.152	0.098	1020	0.181	395	51
QuAC Gen	0.176	0.088	394	0.429	121	0
QuAC Gen SQuAD Pred	0.176	0.088	394	0.429	112	0

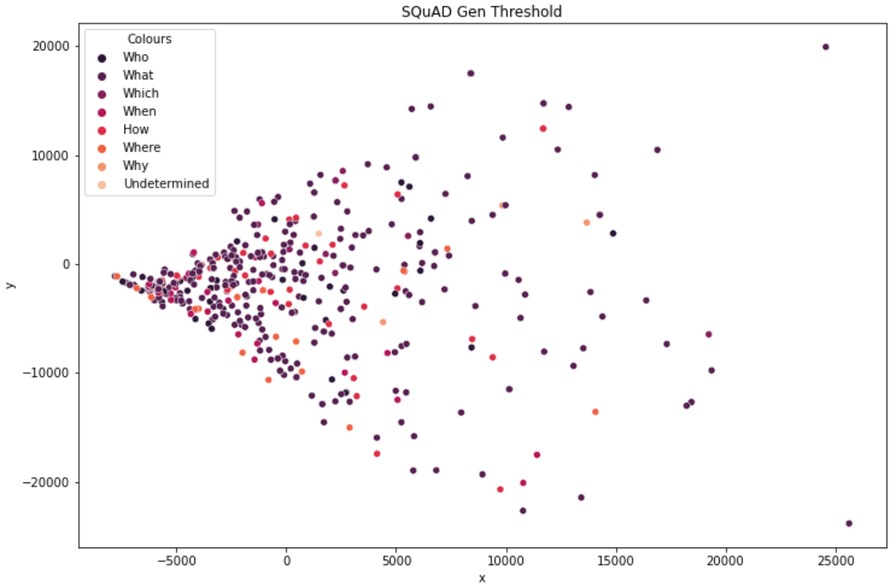
SQUAD EVALUATION

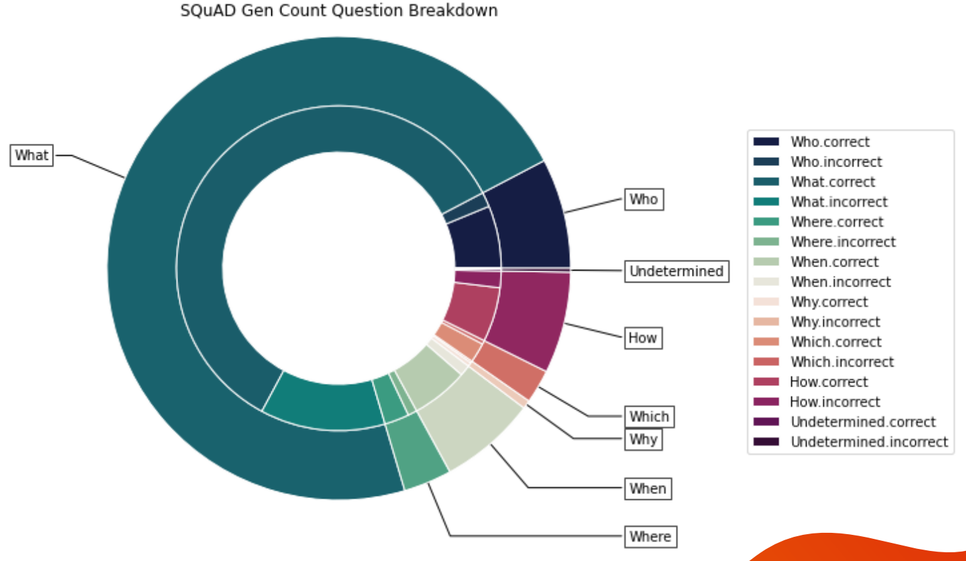
QA





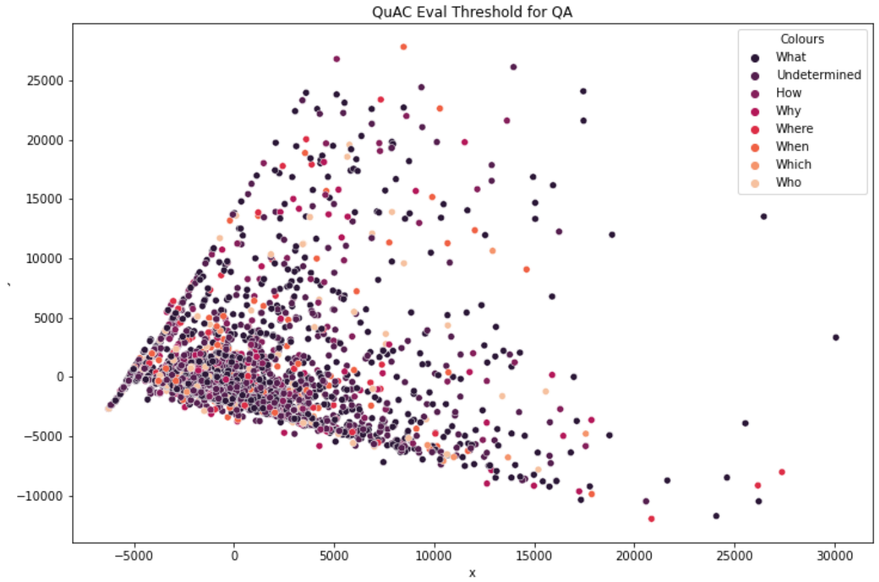
QGen

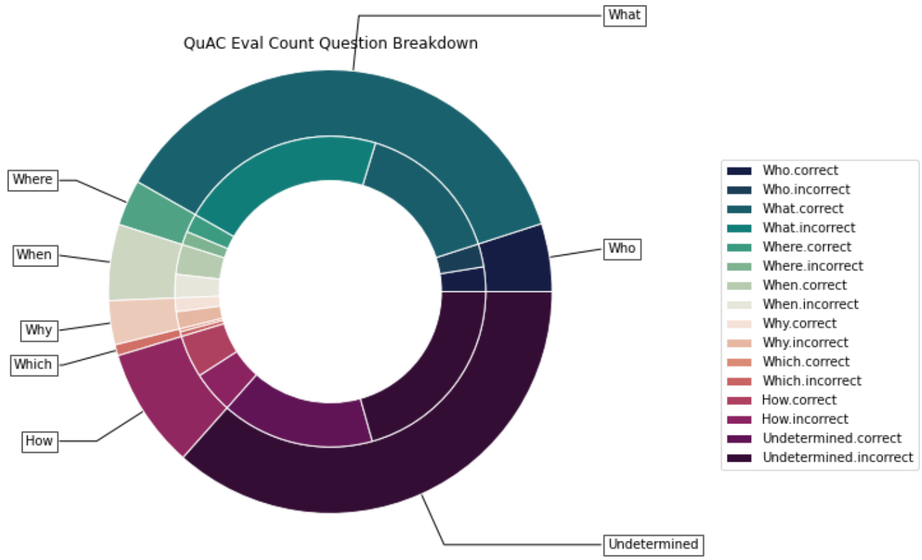




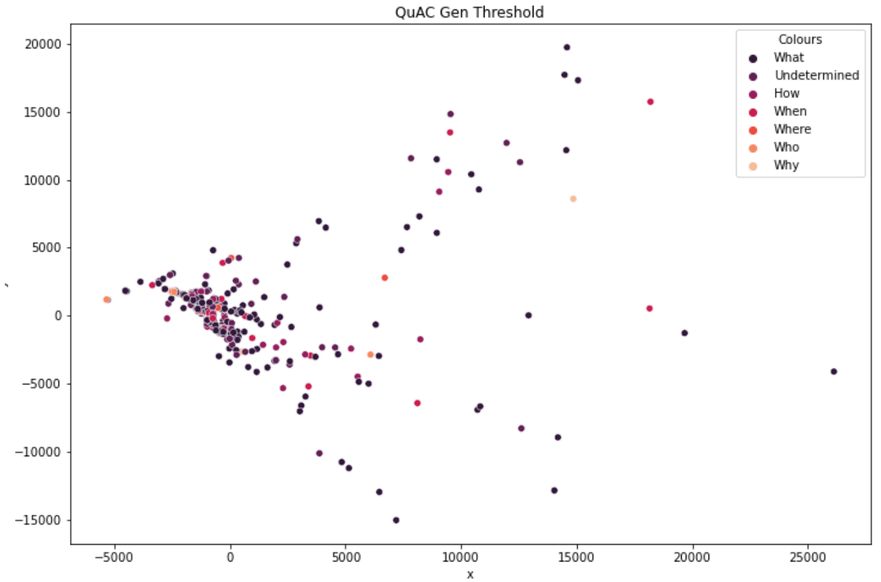
QUAC EVALUATION

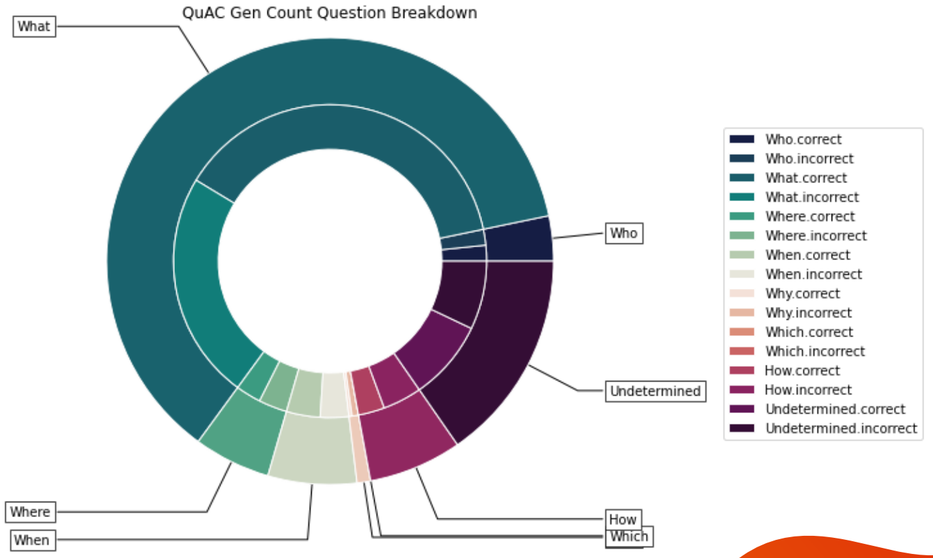
QA



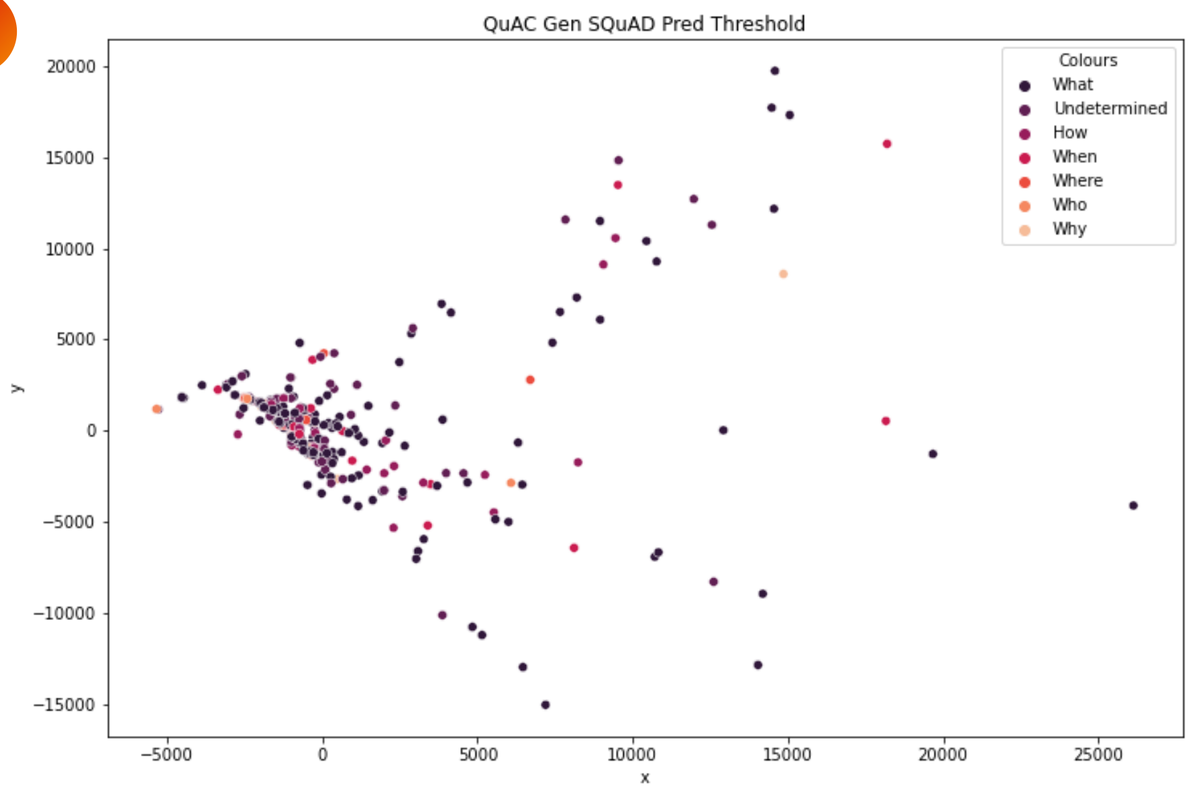


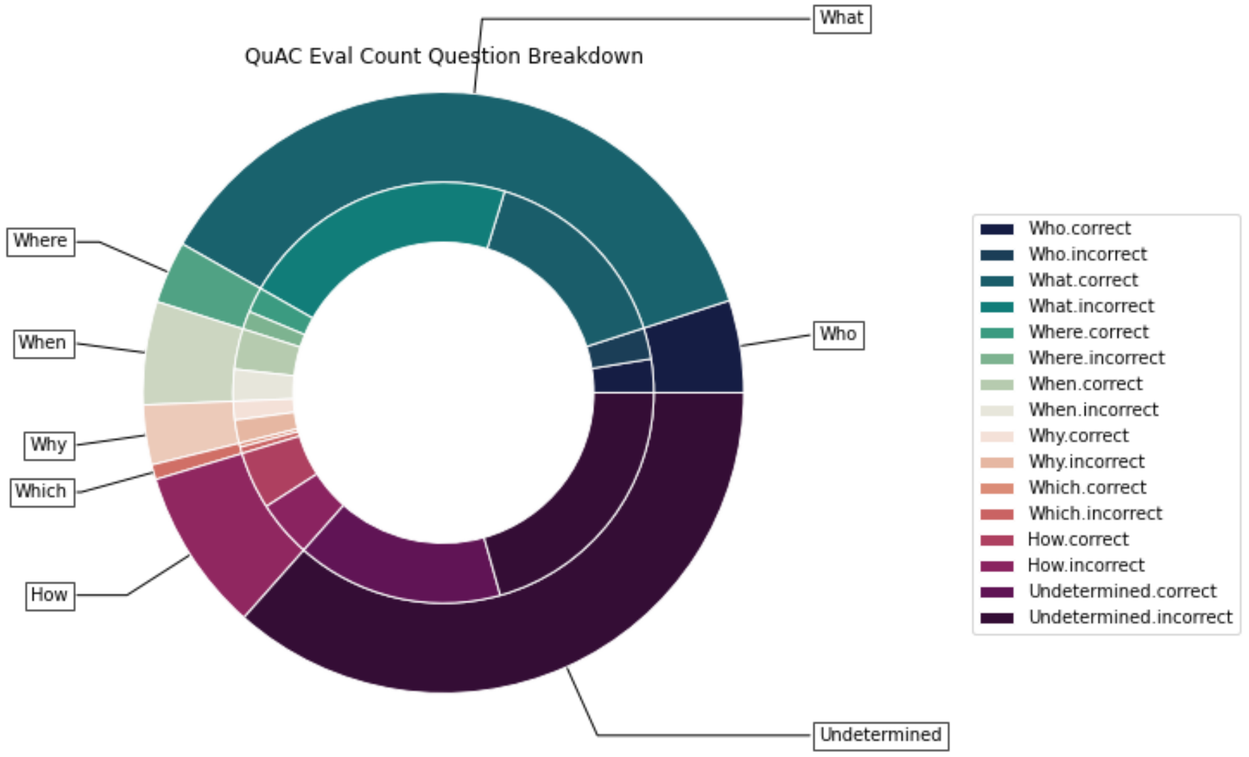
QGen





QUAC EVALUATION USING SQUAD





RESULTS AND CONCLUSIONS

Question Answering (QA)

Outstanding performance on questions that required factual answers and less context dependant questions.

Possible to achieve high results within a short amount of training time.

Question Generation (QG)

Found that more creative questions resulted in an undetermined category or no answer at all.

Similarly good results could be achieved in short time, also due to GPU training to accelerate

Conclusions

While the model process is not new, evaluating and understanding subsets is the novel approach.

Shows that QA and QG systems can be interlinked to produce robust conversational AI.

FUTURE WORK

1. Impact training with dataset combinations has for both QA and QG
2. Understanding affect of varying hypertuning techniques will have on model outcomes
3. Analysing subsets of poor performing questions and answers via clustering
4. Further model additions and layer variations for QuAC improvements
5. Investigation of GPT-2 temperature variation on generation of questions
6. Deployment of a QA and QG pipeline with a frontend interface

CHLOE THOMPSON | DR BARRY DEVEREUX & DR JOANA CAVADAS

MACHINE LEARNING APPLIED IN THE CONTEXT OF QUESTION GENERATION AND QUESTION ANSWERING