Katharina Alefs, Finn Freiheit, Raphael Reimann

Exercise 2 - Assignment 5

Question 2
Perform an explanatory data analysis (EDA) on the train split of the dataset. Train Split:
Distribution of Number of Questions per Article:
1 questions: 220 articles
2 questions: 174 articles
3 questions: 203 articles
4 questions: 160 articles
5 questions: 69 articles
6 questions: 32 articles
7 questions: 14 articles
8 questions: 9 articles
9 questions: 4 articles
10 questions: 1 articles
11 questions: 1 articles
12 questions: 1 articles

Average Number of Answers per Question: 1.032
Number of Free-form Answers: 622
Number of Extractive Answers: 2675
Number of Unanswerable Questions: 281

Distribution of Abstract Lengths:
Minimum length: 279 Maximum length: 2022 Mean length: 982.027
Distribution of Abstract Word Counts:
Minimum word counts: 45 Maximum word counts: 324 Mean word counts: 143.546
Distribution of Answer Word Counts:
Minimum word counts: 1 Maximum word counts: 275 Mean word counts: 10.147

Question 3
Perform an explanatory data analysis (EDA) on the test split of the dataset.
Distribution of Number of Questions per Article:
1 questions: 32 articles
2 questions: 80 articles
3 questions: 115 articles
4 questions: 100 articles
5 questions: 49 articles
6 questions: 23 articles
7 questions: 8 articles
8 questions: 7 articles
9 questions: 1 articles
10 questions: 1 articles
Average Number of Answers per Question: 2.44934527911785
Number of Free-form Answers: 878
Number of Extractive Answers: 3554

Number of Unanswerable Questions: 366

Distribution of Abstract Lengths:
Minimum length: 252 Maximum length: 1909 Mean length: 930.930

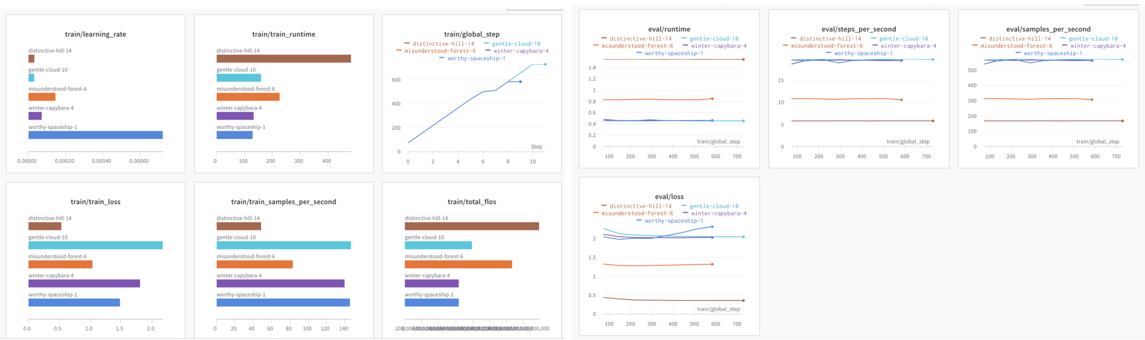
Distribution of Abstract Word Counts:
Minimum word counts: 36 Maximum word counts: 293 Mean word counts: 136.428

Distribution of Answer Word Counts:
Minimum word counts: 1 Maximum word counts: 340 Mean word counts: 8.899

Question 6

Do these maximum sequence values result in truncating a lot of your sequences?
The truncation of the input values, i.e. the "Question: {question} Context: {abstract}" the truncation is significant since we find in our EDA that the distribution of Abstract Word Counts has a mean word count of 143.5 with min 45 and max 324. Keeping in mind that the input value consists of the question, which in itself is mostly short, and the abstract for context we must conclude that for most input sequences the abstract is truncated (since mean word count of abstract = 143.5 > max_length = 128) and therefore not the whole context is provided for the model. The truncation for the answer is not as severe because we find in our EDA that the mean answer word count is 10.15 with a maximum of 275. Since our max_length for the answers is 32 we know that some answers will be truncated, but not the majority. We know that the tokenizer does not strictly produce one token per word but we make this approximation to roughly analyze the impact of the truncation on our sequences.

Question 10



Question 11

Explain your observations. How does your model perform when answering? Does the output make sense? What do you think should be done for improving the prediction?

The model does not perform well for question answering. We see for some questions from the test set that the model actually uses the context to generate an answer but the answer is not true for most cases.

question: Who made the stated claim (that "this is because character-level models learn morphology")?

context: When parsing morphologically-rich languages with neural models, it is beneficial to model input at the character level, and it has been claimed that this is because character-level models learn morphology. [...]

Predicted Answer: <pad> morphologically-rich</s>

We see that the model uses words from the context to generate an answer, however, the answer does not relate to the question in any logical way. For many questions the model actually does not generate a new answer. This could be due to the fact that the training data includes many questions that have no answer. To fix this one could think about cleaning the data so that fewer questions are included that have no answer, if our goal is to train the model to generate answers.

For some questions the model generates answers that are actually close to the ground-truth answers.
question: what combination of features helped improve the classification?

context: Corpora and web texts can become a rich language learning resource if we have a means of assessing whether they are linguistically [...]

Predicted Answer: <pad>a combination of features, compared to using lexical features alone, resulted

Answer: length-based, lexical, morphological, syntactic and semantic features