Exam generator through question generation
Machine Learning for Natural Language Processing 2020

Mohamed Farhat
ENSAE Paris
mohamed.farhat@ensae.fr

Papa Mademba Gaye
ENSAE Paris
papamademba.gaye@ensae.fr

Abstract

We explore the feasibility of generating different types of reading comprehension questions (normal, MCQ and fill-in) by means of a neural network architecture. We build a model derived from (Du et al., 2017) that learns from the context and the question. The model was capable of generating questions to a certain degree but can't be used to generate MCQ or fill-in question. We conclude that the model needs further work and modification to achieve satisfactory results.

1 Problem Framing

In this project[1][2], we aimed to use NLP techniques, more precisely question generation as a basis for an exam generator in the sense that we want to use a model that can generate different question styles from a context (a paragraph or a sentence). We are interested in 3 main forms of questions : direct, fill-in and MCQ.

The chosen model should be able to generate a reading comprehension question given a context and/or an answer. The question should be coherent and understandable without interpretation. The model will be evaluated qualitatively and quantitatively.

2 Experiments Protocol

Given that our goal is to generate reading comprehension questions, we retrieved and adapted what are arguably the two most used datasets for question answering from a reading comprehension point of view which is the inverse task of the model we build. The datasets we used are : The Stanford Question Answering Dataset (SQuAD) in its second iteration and the News Question Answering or NewsQA in its lone version. The SQuAD is a reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or span, from the corresponding reading passage, or the question might be unanswerable. The second dataset is also a crowd-sourced machine reading comprehension dataset of 120,000 question-answer pairs. Documents are CNN news articles and questions are written by human users in natural language. The answers may be multi-word passages of the source text and questions may be unanswerable.

Both datasets have a context that is coupled with a question answer pair. Questions can be valid or not and answers can be absent, multiple or in some cases have a consensus around a valid one.

We retrieve all components of each dataset : The context, the sentences, the questions and the answers. We then merge the same components of each dataset to build our final one. This dataset is then pre-processed (using Spacy tokeniser and GLoVe (Wikipedia 2014-Gigaword-5) embedding of dimension 300).

We then build a sequence to sequence model following (Du et al., 2017) that incorporates an embedding layer, a Bi-LSTM encoder-decoder, an attention layer and a generator for the final output. We train the model on the contexts, questions and the answers of the training dataset and evaluate on the validation part for 15 epochs. We use the negative likelihood loss with a learning rate decay and gradient clipping to avoid gradient problems. We also employ dropout regularisation with a probability of 0.5. Once training is done, we generate some questions on the validation dataset. We can generate MCQ questions by firstly generating a question then take the nearest neighbours to the

[1] https://colab.research.google.com/drive/1GJ6570__wBRYZgUAcTc0eiB5BZfTNKE3?usp=sharing

[2] https://github.com/mohamedfarhat-github/NLP-Project-exam-generator.

answer using a k-nearest neighbour classifier fitted on the embeddings.

We have also used Transfer learning in a pre-trained BERT model to answer questions in an attempt to generate better answer alternatives to those generated by the first method.

3 Results

Our model was able to learn how to generate a question to a certain extent although learning doesn't improve beyond the 15th epoch and the computed loss and accuracy doesn't decrease or increase respectively which prompts us to suspect that the model can't learn properly the structure of a question. The perplexity also doesn't improve in a significant way on both the training and the validation datasets.

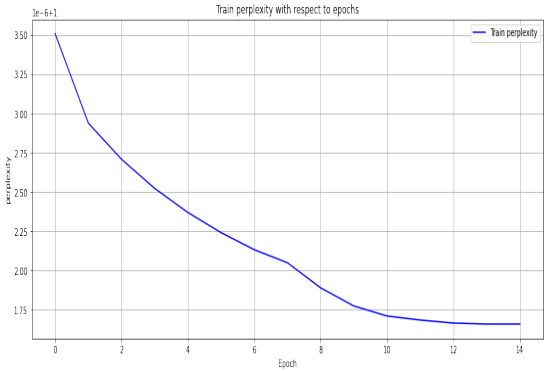

Figure 1: Perplexity with respect too epochs on training data

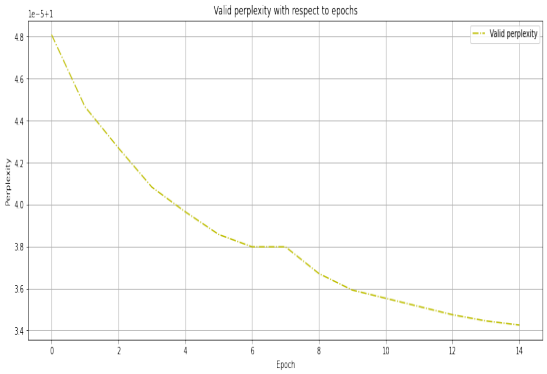

Figure 2: Perplexity with respect to epochs on validation data

Inspecting the generated questions, we can see that many questions lack the WH word that makes them coherent and instead take the inverted form which proves that the model can't differentiate between the two. It can't also make the distinction between the different uses of WH words, for example why and what. The questions can sometimes be meaningless and/or need interpretation to be understood. Arguably the most apparent mistake the model makes is not including the question mark at the end which is essential for a question. We have also noticed that the model in some instances keeps repeating words at the end of a question. For the MCQ questions with the KNN classifier, the method proved to be unstable as alternative answers were not suitable for the question when the question itself made sense. The pre-trained BERT model is a more robust alternative as it generates acceptable alternatives.

4 Discussion/Conclusion

Throughout this project, we build a model that learned, to a certain degree, to generate a reading comprehension question given a context. There is no doubt that the model needs further work to generate proper questions (understandable, coherent ones). Our model succeeded partially in the task of generating questions but it fails to be scaled to generate different kinds of questions. We can propose to inspect a different kind of approach that incorporates a question generator with an question answering component which might be a more robust approach. This method was explored in this repository[3] where they employed pre-trained transformers with the T5 model.

[3] https://github.com/amontgomerie/question_generator

.

References

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1342–1352, Vancouver, Canada. Association for Computational Linguistics.