Question generation with Text-to-Text Transfer Transformer

KWATE DASSI Loïc
BS in Computer Science at National Advanced School of Engineering Yaounde, Cameroon,
16p043@polytechnique.cm

1 Problematic

In our educational system, assess the students through a question form is commonly used by the teachers. Due to the difficulty of question generation task, we note that the evaluations cover the same question over time. Throughout this paper we are going to show how to leverage question generation task with Natural Language Processing technique. We will present in this study, Our approach to solve the problem of question generation, how can we generate targeted question according to a given context. Our system tries to generate the questions whose answers fit into a given context. This kind of questions are useful to train the remembering cognitive skills of students after reading a given text, thus it can be helpful to improve our educational system by decrease the pain of building the test for knowledge assessment.

2 Motivation

The health of some countries almost relies entirely on educational system because it's an essential element in the country's development. Having educated citizens in a country can help it to grow up very quickly. To attend to development of some countries, specially in educational field, we try to design one system that generates question according a given context, and this will be help teachers (to design examination forms) and it will encourage taught self students in the sense that they wouldn't wait for examination to evaluate themselves. The question forms is the common way of examination in our educational system, then serve the relevant questions to the students in a specific domain is useful to evaluate their cognitive skills and reach the learning purpose. Generally in our educational system it's usually accepted that having a great performance at examination reflects a *good learning*. Regarding the Bloom's Revised Taxonomy, the questions are categorized according to the level of thought required to answer them[1]. There are six different levels of cognitive skills:

Remembering :

The questions are designed to evaluate knowledge from long-term memory. The students have to exhibit memory of previously learned material by recalling facts.

understanding :

The Questions are designed to assess the understanding level of student. The students have to demonstrate their understanding of facts and ideas by organizing, comparing, translating, interpreting, giving description.

Applying :

The questions here asses the ability of students to use of build an elaborated procedure to solve problem.

Analyzing :

The Questions evaluate the ability of the students to dive into concepts, break it down and establish the link between them in order to solve a given problem.

Evaluating :

Theses questions evaluate the ability of the students to make judgments based on criteria, standards and objectives.

Creating :

Theses questions evaluate the ability of student to re-organize, merge something to form useful compound for the situation faced.
We firstly focus on the remembering skills, and we designed one system that tries to generate questions to tackle this learning skill

3 Related work

Recent word was developed to solve the problem of question generation based on text corpus and answer features question generation by RNN[2], question generation by transformers[3], question generation by BERT[4], took in this order we highlight the growth of performance measure by BLEU score.

4 Approach

Over the past few years, Natural Language Processing gain significantly improvements in transfer learning. There are a lot of tasks that can in Natural Language Processing, perform these tasks independently will be expensive. These tasks often share certain sub-tasks i.e contextual representation and so on. To face this situation we generally use Transfer learning by using pre-trained model, these models are generally train on unlabeled data with supervised learning task such that Mask Language Modeling, Replace Token Detection, Next Sentence Prediction and so on. The recent success Transfer learning models were BERT, GPT, ELECTRA, RoBerta, XLNet, Reformer, LongReformer, T5 (Text-to-Test Transfer Transformer).
We observe question generation task as two sub-tasks : Natural Language Understanding and Contextual Natural Language Generation. To perform these sub-tasks we use a model inherit to encoder-decoder architecture, encoder is for NLU and decoder is for NLG. Question generation can also be view as Text-to-Text task in the sense that the input and output of the model if a text. Perform Natural Language Understanding requires a high level of comprehension of text, thanks to the self-attention [5] of Transformer, it helps us to encode the meaningful features (i.e linguistic feature, dependence parsing graph) necessary to understand the input text. Encoder-Decoder attention presents in Transformer Architecture is also useful to generate conditional text. Based on all the previous observations we **Google T5 (Text-to-Text Transfer Transformer)** [6]. with T5, all the NLP Tasks are reformulated into a unified text-to-text-format where the input and outputs are text strings. T5 afford the use of same model, loss function and hyperparameters on any NLP task. The figure below presents different tasks that are unified in the same model.

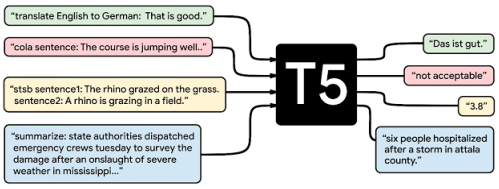

Figure 1: Architecture

4.1 Architecture

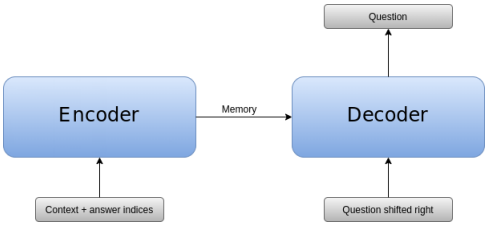

Figure 2: Architecture

4.2 Description of architecture

input :

The first form the input in text string, this text string is tokenized by T5-tokenizer that uses Byte Pair encoding algorithm, it breaks down input text into tokens that lie in word level as well as sub-word level. Theses tokens are then fed to encoder. Before pass the input text through tokenizer we add some meaningful text that tell the model which task it is going to perform. The input text has the form as follows : **"task_to_perform : input_text eos_token"**. the label *"task_to_perform"* here is *"ask_question"* and the label *"eos_token"* is **"</s>"**. We also add in *"input_text"* some special characters to show the model the answer of the question that it will generate. This special character is $<extra\_id\_1>$. The final form of the input is as follows : *"ask_question : context_1* $<extra\_id\_1>$ *answer* $<extra\_id\_1>$ *context_2*

output :

The outputs are only fed at the training step. The form of output is described as follow: *"output_text </s>"*

Encoder:

As we note earlier, Encoder is for Natural Language Understanding, then it consists of Transformer-Encoder linked in BERT fashion. the architecture of Encoder is close to BERT architecture.

Decoder:

The decoder also consists of Transformer-decoder linked in BERT fashion. It is used for language modeling of the question generation task.

forward through network:

let's consider :
$\mathcal{X}$ as encoded inputs of context and answer
Enc as Encoder layer
Dec as Decoder layer
$Softmax$ as Softmax layer
$\mathcal{Q}_{SR} = q_1...q_t$ as output of question shifted right

The equations of our model are :

$\mathcal{X}_1 = Enc(\mathcal{X})$
$\mathcal{Q}_i = Dec(\mathcal{Q}_{SR})$
$q_n = Softmax(Dec(\mathcal{X}_1, q_{t<n}))$ until $q_n =< /s >$

4.3 Data

we trained our model on recently updated question answering dataset SQUAD2.0 consists of question designed by crowdworkers on a set of Wikipedia articles. it consists of 100K+ questions about more than 500+ article in Wikipedia. We use SQUAD train (73K) for training set and SQUAD dev (11K) for test set. This dataset is only consists of question designed to evaluate remember skills, it's the reason why we use it to train our system to generate question to reach this learning purpose.

5 Implementation

5.1 Pre-processing data

Based on SQUAD dataset we constituted the triplets of (context, answer, question) that are fed to model for training Once we constituted these triplet on data, we used WordPiece tokenizer provided by T5-Tokenizer model to tokenize sentence into words and sub-words of vocabulary. The size of vocabulary is around $\sim$ 32000 words, that is very small that the vocabulary of previous word embedding methods such as Glove, Word2Vec, Fasttext.

5.2 Coding step

we used Pytorch Framework to implement our model. We used pre-trained model T5-base that we fine tune according to the purpose of generate contextual question.

We used Teacher forcing at time of text generation during training step.
We used Cross entropy loss (come from Information Theory, used to compute the dissimilarity between two distributions) to compute the loss function objective of our model.
We used Adam's algorithm to optimize the loss function. The hyper parameters are $\alpha = 5e^{-5}$, $\beta_1 = 0.99$, $\beta_2 = 0.999$, $eps = e^{-8}$
At the training step we used the batch size of 256, and 2 epochs
We use beam search algorithm at decoding step with beam size of 4.

6 Results and analysis

6.1 Score

We use the recent BLEURT [7] released by Google AI. BLEURT is a a metrics for Natural Language Generation based on Transfer Learning. BLEURT is an evaluation metric for Natural Language Generation. It takes a pair of sentences as input, a reference and a candidate, and it returns a score that indicates to what extent the candidate is grammatical and conveys the meaning of the reference. It is comparable to sentence-BLEU and BERTscore. the model of BLEURT is based on BERT [8].
The score we obtain with BLEURT metric is as follows.

BLEURT score : 0.133

Here is and example of the assessment of the model in text about the colonisation in Cameroon.

Context :

The Douala-Yaoundé railway line, begun under the German regime, had been completed. Thousands of workers were forcibly deported to this site to work fifty-four hours a week. Workers also suffered from lack of food and the massive presence of mosquitoes. In 1925, the mortality rate on the site was 61.7%. However, the other sites were not as deadly, although working conditions were generally very harsh

Question :

- *What was the mortality rate in 1925?*
- *How many workers were deported to this railroad site?*

- *Which railroad line in the German era was completed in 1925?*
- *Which regime began the Douala-Yaoundé railway?*

6.2 Analysis

The results we have shown earlier highlight, despite the fact that the model was trained on SQUAD, it performs well on the production of question to assess remembering skills after the reading of a given text.

7 Future work

As we mentioned earlier our system only focus to generate question to evaluate remembering skills of students. the future tasks we have to accomplish are as follows :

- design another system which will cover the others learning purposes such as **understanding, analyzing, applying, analyzing, evaluating and creating**. Thanks to **LearningQ**[1] **A Large-scale Dataset for Educational Question Generation** it contains around $\sim$ 230000 questions that fit learning purposes that we highlight here
- design another system for question answering problem, for this task will also use **LearningQ dataset**
- merge these two subsystems to form an autonomous evaluation system. Thus, instead of only generate question the system will also answer the self-generated questions. This global system will make learning easy by help students to constantly evaluate themselves. The added task to the two previous subsystems consists of finding a good way to encode the answer and compute the similarity between the reference and hypothesis answer regarding to the learning outcomes.

8 Conclusion

To conclude these writing but not the study of the subject, we have built a system able to generate questions according to a given context whose answers fit in the context, our system can be used by the teachers to generate the questions to evaluate remembering skills of the students after reading a given document. We will continue to push our skills in field of artificial intelligence and improve our model, to attend to the development of country where education is not widespread

References

[1] Guanliang Chen, Jie Yang, Claudia Hauff, and Geert-Jan Houben. *LearningQ: A Large-scale Dataset for Educational Question Generation*. Association for the Advancement of Artificial Intelligence, Fribourg, 2017.

[2] Qingyu Zhou, Nan Yang, Furu Wei, Chuanqi Tan, Hangbo Bao, and Ming Zhou. *Neural Question Generation from Text: A Preliminary Study*. arXiv, Harbin, Beijing, China, 2017.

[3] Kettip Kriangchaivech and Artit Wangperawong. *Question Generation by Transformers*. arXiv, New York, NY 10036, 2019.

[4] Ying-Hong Chan and Yao-Chung Fan. *BERT for Question Generation*. Association for Computational Linguistics, Taichung, Taiwan, 2019.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. *Attention Is All You Need*. Long Beach, CA, USA., 2017.

[6] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. arXiv, 2019.

[7] Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. *BLEURT : Learning Robust Metrics for Text Generation*. Association for Computational linguistics, New Yord, 2020.

[8] Jacob Devlin, Ming-Wei, Chang Kenton Lee, and Kristina Toutanova. *Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv, 2019.