Generating Questions with Custom Data

Emre Çoban1 Alihan Sağöz1

Abstract

Question generation given a context passage is one of the fundamental profits of artificial intelligence in education. In this paper, we try to introduce a sequence-to-sequence deep learning model using LSTM. On the other hand, as you can see from our title, we also will add some data by hand to extend our data from SQuAD (Stanford Question Answering Dataset) [[1](#page3)]. We think this project could help teachers or anyone who wants questions from the given context. A variety of structures can be used to achieve this project. We ran into a variety of choices. Of course, LSTM and GRU were the first ones we ran into, and we decided to use LSTM. But since it is a progression report we think we might change some things with our model. However, we believe that no matter what we achieve with our project still It will be a great experience for us.

1. Introduction

Questions are important for understanding everything around us. We can use them to understand and know more about the environment. That is the main reason why we chose this project as our topic. We wanted to discover the opportunities that artificial intelligence can provide in education. We are mainly interested in understanding Natural Language Processing and understanding more about processing sequence data. Of course, our project is considered a sequence-to-sequence model. This means we will input a sequence that we consider as our context passage and answer and we will get an output of sequence which will be our generated question. To achieve this, we at the first had a look at encoder-decoder models. Figure 1, for example, shows an example encoder-decoder structure.

We also encountered transformers architecture as an alternative to LSTM [2] (Long Short-Term Memory) which seems more efficient since it uses self-attention and positional embeddings.

Diagram

Description automatically generated

Figure 1: Sample encoder-decoder structure.

The idea is to use fixed or learned weights that encode information related to a specific position of a token in a sentence. The first point is the main reason why transformers do not suffer from long dependency issues. The original transformers do not rely on past hidden states to capture dependencies with previous words. They instead process a sentence as a whole. That is why there is no risk to lose (or "forget") past information. Moreover, multi-head attention and positional embeddings both provide information about the relationship between different words.

But when we kept working on it, for now, we decided to move on with LSTM (figure 2) encoder-decoder model. Because we understood and studied it a little bit more and we think it is a good model to start with this project.

Diagram, schematic

Description automatically generated

Figure 2: LSTM cell structure.

LSTM is an artificial neural network used in the fields of artificial intelligence and deep learning. Unlike feedforward neural networks, LSTM has feedback connections. LSTM is suitable for tasks like machine translation which is a sequence-to-sequence task just like our project.

For those who don’t know why is RNN (Recurrent Neural Network) [5] not enough for this kind of topic? To answer this question to start with, RNN also has feedback connections, and it was suitable for tasks that are suitable for LSTM in the past. The main problem of RNN is that when the input sequence gets longer and longer it has this vanishing gradient problem, figure 3 shows the visualization of this vanishing gradient problem.

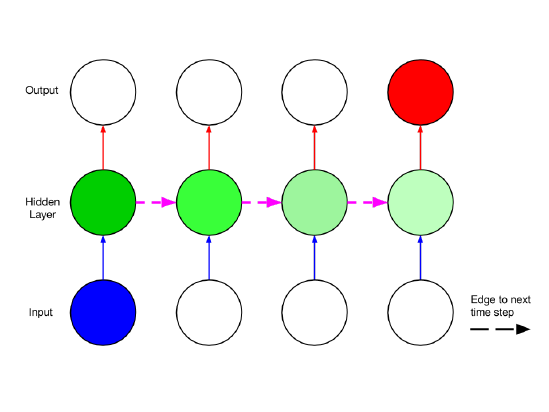


Figure 3: Visualization of Vanishing Gradient.

this problem causes the model to not update weights and biases accordingly for the task. So, we go with LSTM because it deals with this problem using more gates.

LSTM has 3 important gates an input gate, an output age and a forget gate. The input gate as it is obvious from its name it takes input from a cell before or from the input. The forget gate controls whether to forget or keep the information in the cell

To summarize what we mentioned in this section, we will move on with the encoder-decoder model using LSTM because we understood and studied it more right now. But maybe we can change our idea and use Transformers or fine-tune a pre-trained model such as BERT (Bidirectional Encoder Representations from Transformers) [8] if we can get permission from our associate professor.

2. Related Work

We found a lot of work related to our project. There are many studies for example on them is the LSTM network for question answering [8] which is very similar research to our project we tried to understand how to use LSTM for this kind of sequence-to-sequence project and there is also one study about question generation using transformers [9] which is way better performing than LSTM but just like we mentioned above right now we are still at the process of trying different models and trying to find best-suited one. The other related works we found with our project

### TRANSFORMERS, Bİ-LSTM QUESTİON GENERATİON, QUESTİON ANSWERİNG 4-5 PAPER NELER FARKLI, HANGİ PAPERLARDAN NELER ALDIK VE KISA SÜREDE ANLADIĞIMIZ KISMI ALDIK VS .VS GİBİ ###

3. Dataset

We will use is Stanford Question Answering Dataset (SQuAD) as our main dataset, but we are going to extend this data by adding our custom data generated by our hand.

Right now, we have around 150 context passages, questions, and answers. Unlike SQuAD right now we don’t have the answer start indexes as our feature but if we need to use it when creating our model, we believe that we can extract that feature easily. Figure 4 shows an example data point that we generated by hand. Just like we said there are only 2 differences first is that for some questions SQuAD has more than one answer such as alternative answers for “1915” is “in 1915” and “March 1915” etc. for some questions, also the new version of SQuAD 2.0 has a new feature which specifies whether the given question is answerable or not.

And the other difference is the answer start index which is mentioned above.

How human beings converse with each other is by using a common language. It is A body of words for their use common to people of the same community or nation, the same geographical area, or the same cultural tradition. The methodological study of language is called linguistics. The language distinguishes a country or in a region or a community.

Human language has the properties of capacity and movement and relies entirely on social convocation and learning. The Language was originated when preceding hominids gradually started changing their hominoid communication systems, acquiring the ability to approach other minds and lived experiences.

According to the philosophical angle, the definition of language and meaning, when used as a general concept, “language” may refer to the cognitive ability to learn and use systems of composite communication, or to describe the set of rules that constructs these systems, or the set of utterances that can be produced from those rules.,

- How do human beings converse with each other?

by using a common language

- Which is a common language?

A body of words

- What is the methodological study of language called?

Linguistics

-What did hominids gradually change?

hominoid communication systems

Figure 4: Question-answer pairs for a sample passage in the dataset we generated. Each of the answers is a segment of text from the passage.

Right now, we have around 150-200 question-answer pairs in our dataset if we have enough time, we plan to extract in total more than 500 question-answer pairs.

4. Methodology and Experiments

As the first step of our project, we tried to understand and investigate the Stanford question-answering dataset we wanted to create a baseline model which works with this data and later we will add our data to it.

We preprocessed data by removing unnecessary characters and punctuation. We tokenized contexts and questions and turned them into sequences of numbers. We did not use pre-trained embedding for this experiment and for the first experiment, we didn’t delete stop words either from context or questions and we get bad outputs. For example, figure 5 shows an example output.

Context:

-On July 1, 2014, the University of Notre Dame and Under Armour reached an agreement in which Under Armour will provide uniforms, apparel, equipment, and monetary compensation to Notre Dame for 10 years. This contract, worth almost $100 million, is the most lucrative in the history of the NCAA. The university marching band plays at home games for most of the sports. The band, which began in 1846 and has a claim as the oldest university band in continuous existence in the United States, was honored by the National Music Council as a "Landmark of American Music" during the United States Bicentennial. The band regularly plays the school\'s fight song the Notre Dame Victory March, which was named as the most played and most famous fight song by Northern Illinois Professor William Studwell. According to College Fight Songs: An Annotated Anthology published in 1998, the "Notre Dame Victory March" ranks as the greatest fight song of all time.

Original Question:

-What is the value of the contract between Under Armour and Notre Dame?

Generated Question:

-What is the name of

Figure 5: Example output without removing stop words

As we can see from figure 5 generated questions are not meaningful. Questions do not extract important words for example Notre Dame or Armour etc.

We think the main reason for these bad outputs is we didn’t preprocess data very well. However, we did another experiment too where we removed stop words from both context passages and questions. But the results were still bad. Later, after researching more and more we decided that we should remove stop words from context but not from questions because we think that stop words in questions are important for our model to understand. Figure 6 shows an example output with stop words removed from contexts but not removed from questions.

Context:

types energy varying mix potential kinetic energy example mechanical energy sum usually kinetic potential energy system elastic energy materials also dependent upon electrical potential energy among atoms molecules chemical energy stored released electrical potential energy electrons molecules atomic attract need the list also necessarily complete whenever physical scientists discover certain phenomenon appears law energy conservation new forms typically added account

Original Question:

what is dependent upon electrical potential energy?

Generated Question:

what is the name of the of the united league

Figure 5: Example output of model when removing stop words from context

As we can see from the figure 6

Going on we researched papers related to transformers and pretrained transformers such as BERT to see if we can have a better model structure instead of an LSTM-based encoder-decoder. Since our contexts are too large maybe LSTM is not enough for projects like this, we need way more improved architectures to get better results.

References

[1] Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). *SQuAD: 100,000+ Questions for Machine Comprehension of Text*. doi:10.48550/ARXIV.1606.05250

[2] Hochreiter, S., & Schmidhuber, J"urgen. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

[3]

[4]

[5]

[6]

[7]

[8]

References

[[4](#page3)]. https://www.researchgate.net/figure/Structure-of-the-LSTM-cell-and-equations-that-describe-the-gates-of-an-LSTM-cell\_fig5\_329362532

[[5](#page3)]. <https://en.wikipedia.org/wiki/Long_short-term_memory>

[6]

[7] <https://arxiv.org/abs/1810.04805>

[8] https://arxiv.org/abs/1707.01961

[9] https://arxiv.org/pdf/1909.05017.pdf