Ilkogretim Online - Elementary Education Online, 2021; Vol 20 (Issue 3): pp. 3251-3256
http://ilkogretim-online.org
doi: 10.17051/ilkonline.2021.03.338

A Mechanism Of Context-Aware Answer Extraction In Question Answering

Jaskaran Singh[1], **Dr Rakesh Patra**[2]

[1]Department of Computer Science and Engineering, Graphic Era Deemed to be University Dehradun, Uttarakhand, India 248002
[2]Assistant Professor, Department of Computer Science and Engineering, Graphic Era Hill University, Dehradun.

[1]Corresponding Author: jaskaran.jsk2001@gmail.co

Abstract—One of the most popular research topics in recent years has been the development of a Natural Language Processing system capable of extracting answers to natural language queries from a given context. Question Answering is a computer science discipline that focuses on building systems that automatically answer questions posed by humans in natural language. It is related to information retrieval and natural language processing. It aims to provide precise answers in natural language in response to the user's questions. We used various NLP datasets in this project to train an NLP model for developing a context aware question answering system. In order to help with the task, we trained multiple models using BERT and LSTM architecture and performed a variety of processing tasks. To fit our model, we converted data to textual data, token sized, stemmed, and embedded it. The final model was then deployed to a front end application, which takes context data and questions and, after processing and utilizing the deployed model, provides answers based on its learning.

Keywords—Machine learning, prediction, expert system, context-aware, natural language processing (NLP).

I. INTRODUCTION

Question Answering is a computer science discipline that focuses on building systems that automatically answer questions posed by humans in natural language. It is related to information retrieval and natural language processing. It aims to provide precise answers in natural language in response to the user's questions. We used various NLP datasets in this project to train an NLP model for developing a context aware question answering system. In order to help with the task, we trained multiple models using BERT and LSTM architecture and performed a variety of processing tasks. To fit our model, we converted data to textual data, tokesized, stemmed, and embedded it. The final model was then deployed to a front end application, which takes context data and questions and, after processing and utilising the deployed model, provides answers based on its learning.

A. Motivation

Fortune Business Insights reports “NLP market will witness an impressive 29.4 percent compound annual growth rate by 2028”. NLP makes it possible for machines to understand human language by developing intelligent systems capable of interpreting context, sentiment, and syntax. The primary motivation is NLP-based. QA systems are a developing field that will be the future of automation and will be critical for any type of organisation in any industry. These systems will be further restricted and could be better trained on their specific private data, similar to how Bio-BERT was trained for the medical and biotechnology domains. I was personally interested in this project because I am particularly interested in the domain of NLP, and because of its moderately researched work, it is wonderful to explore this domain and construct a question answering software.

II. Related work

Whereas Natural language processing through deciphering how a computer can read, understand and generate human understandable language has been researched for a long time [1], most of the work done in the domain of question answering is done after the advent of availability of pre trained bert models [2]. Paper [3] elaborates more on work done in Bert and its role in nlp pipeline with Qualitative analysis. Paper [4] uses target dependent variation of the basic BERT model to perform Sentiment Classification of text sequences. On a different model [5] utilizes a Transformer based Neural Network model for Answer Selection in a question answering system. Paper [6] performs NLP tasks with automatic feature engineering with question answering combining syntactic and semantic information. Paper [7] provides and evaluates a deep learning approach for Visual Question Answering. Paper [8] uses an n-gram match and an attention-based siamese bidirectional long-short term memory model for open-domain question answering. Paper [9] uses a Bidirectional Long Short-Term Memory (BiLSTM) model that uses co-attention mechanism and cosine similarity to derive the relation between questions and answers. Paper [10] also utilises the same co-attention mechanism for answer selection and performed experiments on WikiQA dataset.

III. Proposed Methodology

A. Data preparation

For our training and fine-tuning processes, we used two different datasets. Our BiLSTM model was trained on CORD 19 data to become a closed domain NLP model (one that deals with questions within a specific domain) that deals with health and hospitalisation data. In the SQuAD dataset, we fine-tuned our Bert Model. The Stanford Question Answering Dataset (SQuAD) is a reading comprehension dataset made up of crowdworker-posted questions on a set of Wikipedia articles. Every question has an answer that is a text segment, or span, from the corresponding reading passage. All of this data was in the form of JSON, which was converted into textual data using a custom Python function before being sent to the next stage of data pre-processing.

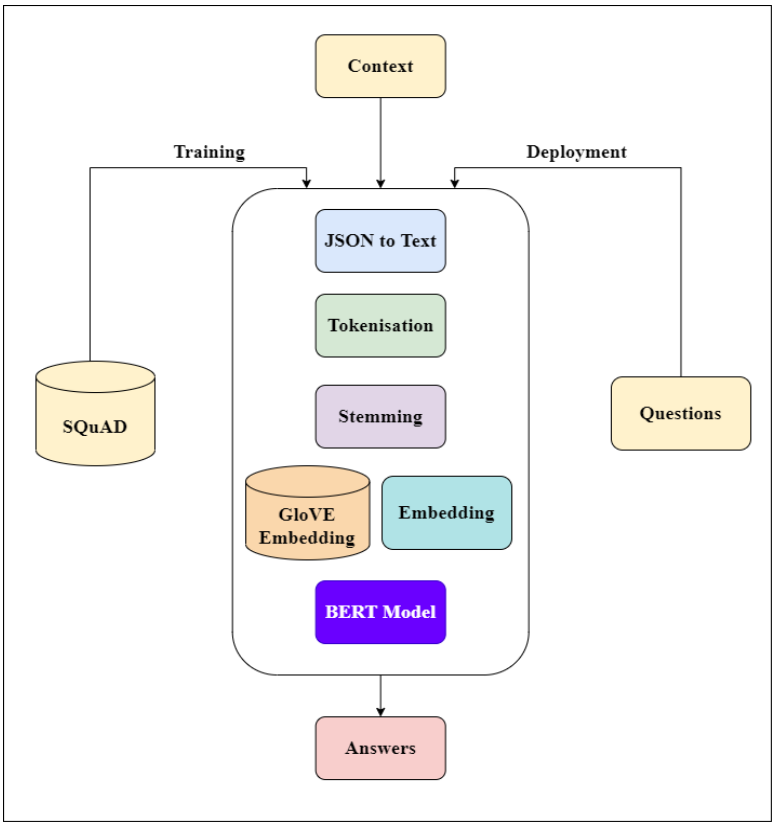

Fig. 1. Architecture of proposed system

```
def json_to_text(js_data):
    data = list()
    max_count = 10000
    max_seq = 500
    max_question = 100
    max_target = 100
    for sample in js_data['data']:
        for text in sample['paragraphs']:
            context = text['context']
            context_list = [word.lower() for word in nltk.word_tokenize(context) if is_stop_word(word)]
            if len(context_list) > max_seq:
                continue
            qs = text['qas']
            for q in qs:
                question = q['question']
                question_list = [word.lower() for word in nltk.word_tokenize(question) if is_stop_word(word)]
                if len(question_list) > max_question:
                    continue
                answers = q['answers']
                for answer in answers:
                    ans = answer['text']
                    answer_list = [word.lower() for word in nltk.word_tokenize(ans) if is_stop_word(word)]
                    if len(answer_list) > max_target:
                        continue

                    if len(data) < max_count:
                        data.append((context, question, ans))

            if len(data) >= max_count:
                break

            break
    return data
```

```
data = json_to_text(qa_data)
data
```

```
of Queen Victoria in 1837.',
  'When did the palace become the London residence for the monarchs?',
  '1837'),
 ('An incandescent light bulb, incandescent lamp or incandescent light globe is an electric light with a wire filament heated to a high temperature, by passing an electri
until it glows with visible light (incandescence). The hot filament is protected from oxidation with a glass or quartz bulb that is filled with inert gas or evacuated. Ir
evaporation is prevented by a chemical process that redeposits metal vapor onto the filament, extending its life. The light bulb is supplied with electric current by feed
wires embedded in the glass. Most bulbs are used in a socket which provides mechanical support and electrical connections.',
  'What type of energy makes an incandescent light bulb glow?',
  'electric current'),
 ('An incandescent light bulb, incandescent lamp or incandescent light globe is an electric light with a wire filament heated to a high temperature, by passing an electri
until it glows with visible light (incandescence). The hot filament is protected from oxidation with a glass or quartz bulb that is filled with inert gas or evacuated. Ir
evaporation is prevented by a chemical process that redeposits metal vapor onto the filament, extending its life. The light bulb is supplied with electric current by feed
wires embedded in the glass. Most bulbs are used in a socket which provides mechanical support and electrical connections.',
```

Fig. 2. Snippet of implemented system

B. Data Pre-processing

1) Tokenisation: Tokenization is the process of converting a piece of text into small units known as tokens. A token can be a word, a word fragment, or just characters such as punctuation. We can train our model better using it because by analysing the words in the text, we can easily interpret the meaning of the text because the entire data is better processed individually.
2) Stemming: Stemming is a technique for removing affixes from words in order to extract their base form. We utilise stemming mechanisms to store the meaning and usage of only the stems, not the entire words. By using stemming, there is reduction in size of the index and dataset while increasing retrieval accuracy.
3) Embedding: Word embedding is a technique where textual tokens are converted into a numerical representation in the form of a vector. The vectors attempt to capture various aspects of that word in relation to the overall text. For textual data word embedding in our task, we use GloVe embedding. GloVe (Global Vectors for Word Representation) is an unsupervised learning algorithm that aggregates global word co-occurrence matrices from a given corpus of data to generate word embeddings. GloVe word embedding derives the relationship between words primarily from statistics.
4) Model training and hypertuning: After pre-processing, we build an architectural model and train it to extract answers to queries from a given context. We used a variety of models to train our model, including BiLSTM and RNN models. The best results were obtained using the BERT model (Bidirectional Encoder Representations from Transformers), which is a transformer Encoder stack that is specifically tuned for the SQuAD dataset. We fine-tuned the BiLSTM model further on the CORD 19 dataset, but for deployment, we use transfer learning via the BERT model due to its low computational cost and time.

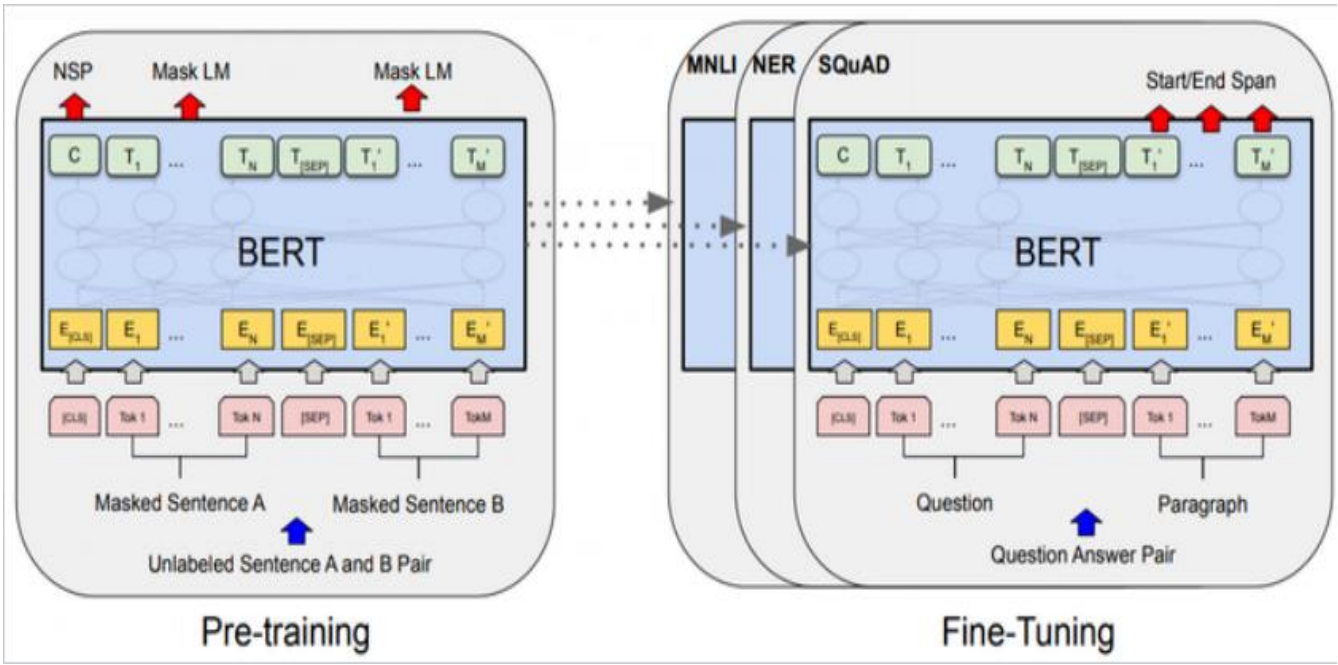

Fig. 3. Pre-training and fine-tuning in proposed system

IV. IMPLEMENTATION

To deploy the tuned BERT model in an interactive server, we used streamlit library, a powerful front end library, and a custom Python script. In the script we used torch, numpy and other necessary libraries to deploy the model and call stremlt over it. We took the user's input in the form of contextual data and the question. The data is then tokenized, cleaned, embedded, and sent to the model, which uses context awareness to display the answer by performing processing on a local server. Using a company's private dataset and GPU servers, this entire application could be expanded into a custom nlp application that will help with the automation of their services.

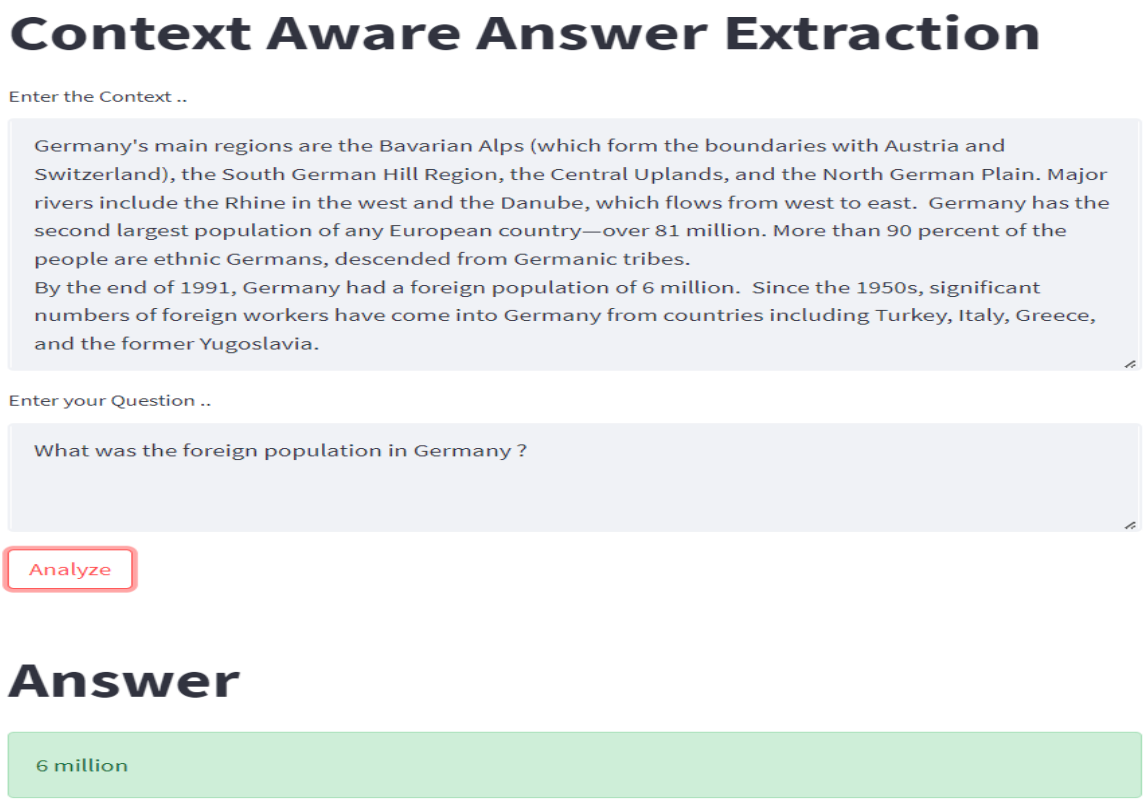

Fig. 4. Implemented system

V. CONCLUSION

Question Answering is a computer science discipline that focuses on building systems that automatically answer questions posed by humans in natural language. We used various NLP datasets in this project to train an NLP model for developing a context aware question answering system. In order to help with the task, we trained multiple models using BERT and LSTM architecture and performed a variety of processing tasks. The final model was then deployed to a front end application, which takes context data and questions and, after processing and utilising the deployed model, provides answers based on its learning.

REFERENCES

[1] Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. J Am Med Inform Assoc. 2011 Sep-Oct;18(5):544-51. doi: 10.1136/amiajnl-2011-000464. PMID: 21846786; PMCID: PMC3168328.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, & Kristina Toutanova (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. CoRR, abs/1810.04805.

[3] Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT Rediscovers the Classical NLP Pipeline. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

[4] Z. Gao, A. Feng, X. Song and X. Wu, "Target-Dependent Sentiment Classification With BERT," in IEEE Access, vol. 7, pp. 154290-154299, 2019, doi: 10.1109/ACCESS.2019.2946594

[5] T. Shao, Y. Guo, H. Chen and Z. Hao, "Transformer-Based Neural Network for Answer Selection in Question Answering," in IEEE Access, vol. 7, pp. 26146-26156, 2019, doi: 10.1109/ACCESS.2019.2900753.

[6] Aliaksei Severyn and Alessandro Moschitti. Automatic Feature Engineering for Answer Selection and Extraction. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 458–467, Seattle, Washington, USA. Association for Computational Linguistics.

[7] Mateusz Malinowski, Marcus Rohrbach, & Mario Fritz (2016). Ask Your Neurons: A Deep Learning Approach to Visual Question Answering. CoRR, abs/1605.02697.

[8] K. Lei, Y. Deng, B. Zhang and Y. Shen, "Open Domain Question Answering with Character-Level Deep Learning Models," 2017 10th International Symposium on Computational Intelligence and Design (ISCID), 2017, pp. 30-33, doi: 10.1109/ISCID.2017.58.

[9] Cai, R. A Stacked BiLSTM Neural Network Based on Coattention Mechanism for Question Answering. Computational Intelligence and Neuroscience, 2019, 9543490.

[10] L. Zhang and L. Ma, "Coattention based BiLSTM for answer selection," 2017 IEEE International Conference on Information and Automation (ICIA), 2017, pp. 1005-1011, doi: 10.1109/ICInfA.2017.8079049.