Natural Language Question Answering

Akhila Jetty
ajetty@umass.edu

Neeharika Karanam
nkaranam@umass.edu

Nikitha Masanagi
nmasanagi@umass.edu

Roshitha Bezawada
rbezawada@umass.edu

1 Introduction:

Question Answering is one of the most challenging tasks in the natural language processing domain [1]. The most important components in the question answering are the capability to understand the question and the context in which the question has been generated. It has been considered quite challenging as it is dynamic in nature [1] has led to the application of the data-driven methods of question answering. The main idea is to allow the data to be given more importance than the methods in the question answering as there are many text repositories which are available [2].

One of the initially used and the most prominent methods for question answering systems is the rule-based approach. These kinds of systems use the rules which have been devised from the grammatical semantics which will help determine the correct answer for a particular solution. They are usually handcrafted and the heuristic, which relies on the lexical and the semantic hints [3]. These rules help to exploit the predefined patterns which will classify the different answer types based on the questions. The grammatical rules which represent the context in the form of decision trees which helps in finding the path that will lead to the correct answer [4].

Question answering is very helpful as they allow the users to ask a question which is based on the various facts or stories that the system tries to use which are in context to the supporting stories or articles and answer the questions in such a way that it gives back a context sensitive and meaningful answer than some related keywords. There are a lot of problems in the domains of Natural Language Processing and Artificial Intelligence which are aligned as the question answering problems [5]. For instance, the text summarization can be reframed as a question answering task where the user is asking the model for "What is the summary of the article?", which will provide the summary of the entire article. Our aim is to design a system which can be basically used to help solve the problem.

We use Natural Language Processing to help resolve the problem of Question Answering after a deep analysis and review on the various models for the Question Answering System to work. One of the main methods we have considered is the use of the Transformers like DistilBERT model which is already pre-trained on the Masked LM and the Next Sentence Prediction. We can also make minor changes in the model which might produce a good result in the end.

The BiDAF model is used in the earlier stages but ended up switching to using the transformers instead of using the various deep learning models like LSTM, CNN etc., is because the transformers are much more suitable for the question answering tasks as they are easy to develop, and we can achieve good results in a limited data. It is believed that instead of training a new model from scratch every single time, the base layers of the trained network along with the generalized features which acts the backbone can be utilized in various other networks that perform a different task. The pre-training of the DistilBERT- based model is done on other tasks other than the question answering system but the fine-tuning is performed on the SQuAD dataset [5].

2 Background/Related Work:

Question Answering (QA) has been a difficult task in natural language processing and understanding [1]. The fundamental component of QA requires the understanding the meaning of the question as well as the context of the question, before generating the answer. Because of the dynamic nature of natural languages, the task of QA has been quite challenging.

The initial question answering systems were rule-based approaches, that is, logical representations of decision trees. Later, due to the increase in data in the form of text repositories and web data, statistical approaches gained momentum. Models like Support vector machine (SVM) classifiers, Bayesian classifiers, Maximum entropy models are some techniques that have been used for question classification purpose.[8]

With the introduction of machine learning to the QA domain algorithms that can learn to understand linguistic features without explicitly being told to. Using statistical methods cemented the path for this approach that system uses to analyse an annotated corpus or training set and then build a knowledge base. [9]

Many natural language processing (NLP) models have been used for this task.[10] One of the major model proposed uses POS tagging[11] and tf-idf concept [5] to match the question with with questions already present on Yahoo Answers. Although, the major flaw arised when the query might not be present on the internet.

Several Deep Neural networks models have been used to solve NLP tasks.These models have majorly used Recurrent Neural networks like LSTMS,s and GRU,s for text classification,summarization.[11]One of the major breakthrough is Memory networks model (Weston et al.)

[12] which proposed the use of memory in the system in order to effectively answer the questions.

In our proposed work, transformers are being used to make better predictions and metrics. These are very strong and powerful neural network architectures which solely rely on the attention mechanism which is trainable and also helps in identifying the complex dependencies between the various elements for each of the given input sequence. A few of them include LUKE [13] which will use the deep contextualized Entity Representations with the help of entity-aware Self-attention, XLNet [14]which will help us integrate the ideas from the Transformer-XL, the state-of-the-art autoregressive model and then comes the most popular transformer called the BERT.

3 Task Definition:

We define the task of the question answering system using the SQuAD dataset and we primarily start-off by using the recurrent-based BiDAF model. We then use the DistilBERT model which is pre-trained on the Masked LM and the Next Sentence Prediction and compare the results of both the models using the evaluation metrics like Exact Match and F1 score. In the ideal question answering system we should be able to get better results and meaningful answers when we use the DistilBERT model when there is a lot of diversity in the questions and the various answer types. Moreover, we want to implement a custom QuestionAnswering Head using the DistilBERT as the backbone to fine-tune and expect better results. One challenging task is identifying the nature of the question.

4 Datasets:

We will be using the Stanford Question Answering Dataset (SQuAD) which is a reading comprehension dataset, consisting of various questions which are posed by the crowdworkers on a few sets of articles which are gathered from Wikipedia. The creators of the dataset have sampled 536 from the top 1,0000 articles on Wikipedia. Amongst these sampled articles 23,215 individual paragraphs have been extracted. The dataset is split in the form of training, development and testing in the ratio of 80%, 10% and 10% respectively. In this every question will have an answer in the form of a segment of text or in the form of span from the reading passage which is corresponding to it. SQuAD dataset is quite big, and it is very challenging therefore it requires a lot of reasoning when compared to the other existing datasets consisting of reading comprehension. One of the most famous dataset is the cloze dataset in which the model is asked to predict a missing word from the passage. These datasets are a bit too similar, but SQuAD mainly focuses on the task of question answering and tests the ability of the model to read a particular passage of text and answer various questions about it.

One of the main reasons that SQuAD is so good is because the dataset is very big unlike the other datasets, the SQuAD dataset in quite challenging and the SQuAD requires a lot of intensive reasoning making it the better in the aspects of evaluating the model and understanding its capabilities. It has a few key properties which have been very well researched by the creators of the dataset:

- Various categories of answers – There are various categories in which the answers are partitioned like “date”, “person”, “location”, “other numeric”, “adjective phrase”, “verb phrase”, “clause”, “other entity” and “other”.

- Intensive reasoning required – The questions were sampled by the creators from the development set, and they were manually labeled questions into various categories of reasoning required to help answer them.

- Syntactic divergence – The creators have measured the syntactic divergence between the question and the answer to measure the difficulty of a particular question. This is a metric that will help us in evaluating the number of edits which are needed to transform a particular question into the form of an answer.

The SQuAD dataset has 87,599 train samples and 10,570 validation samples. The SQuAD dataset has the following fields:

- Id: string
- Title: string
- Context: string
- Question: string
- Answers: Dictionary

Example of the train data is as below:

{ "title": "University_of_Notre_Dame",
"paragraphs": [{ "context": "Architecturally, the school has a Catholic character. Atop the Main Building's gold dome is a golden statue of the Virgin Mary. Immediately in front of the Main Building and facing it, is a copper statue of Christ with arms upraised with the legend \"Venite Ad Me Omnes \". Next to the Main Building is the Basilica of the Sacred Heart. Immediately behind the basilica is the Grotto, a Marian place of prayer and reflection. It is a replica of the grotto at Lourdes, France where the Virgin Mary reputedly appeared to Saint Bernadette Soubirous in 1858. At the end of the main drive (and in a direct line that connects through 3 statues and the Gold Dome), is a simple, modern stone statue of Mary.",
[{ "answers":
[{ "answer_start": 515,
"text": "Saint Bernadette Soubirous" }],
"question": "To whom did the Virgin Mary allegedly appear in 1858 in Lourdes France?",
"id": "5733be284776f41900661182" }]
}

5 Method:

5.1 Baselines:

The Baseline model BiDAF is a hierarchical multi-stage architecture model introduced in 2016 by University of Washington. In order to answer a question, the model checks the accompanying text that has the needed information by which the question will be answered. We can name the accompanying text as Context.We can say that BiDAF is a

- Closed domain model : This Baseline BiDAF model doesn't rely on pre-existing knowledge. It requires a context to answer a query.
- Extractive model : The model returns a substring of context relevant to the query.
- popular deep learning question and answer model.

Steps in BiDAF model :

1. Tokenization: The query(T) and the context(J) contents are tokenized and broken down into their constituent parts.

2. Embedding: In BiDAF, Embedding in performed in different levels on granularity:

   (a) Word Level Embedding: This is the first embedding layer where we substitute the resultant words from the above step using pre-trained glove embeddings into vectors containing numbers then passed on to the character level embedding layer. Glove means Global vectors for word representation.Glove is an unsupervised learning algorithm where word-word co-occurrence statistics from the dataset is used to get linear substructures in the word vector space. Mathematical operations are performed on these vectors which capture both syntax and semantics of the words. We can encounter some words not found in the huge Glove corpus called the out of vocabulary words.Glove assigns some random values to such inputs.

   (b) Character Level Embedding: The out of vocabulary words are handled by using Convolutional neural networks which work like a feature extractor for each word and numeric vector representation for each word is found by observing character level constitution. The output is similar to the output in the above step. The outputs obtained from Word level embedding and character level embedding are then vertically concatenated and sent to a highway network.

   (c) Highway network: A Highway network is a series of feed-forward layers with a gating mechanism. The relative contribution from words resulting from word level embedding and character level embedding are adjusted as while for out of context words, the Glove gives a random output and the character level embedding layer gives a better output. We are increasing the relative importance of the out of context words from CNN.

   (d) Contextual Embedding: The output from the highway network is passed to the final embedding layer called the Contextual embedding layer implemented with a bidirectional LSTM composed of forward and backward LSTM sequences. When we take Homophonic words, they should not be treated the same as the words that might spell the same but do not have the same meaning when used in a context. Contextual information of the words is not taken into account assigning the same numeric vector representation to actually different words, confusing the model.A word should also be understood in a contextual way deriving meaning from its surroundings. The output representation from bidirectional LSTM incorporates the contextual meaning of a word.

3. The attention Layer: Attention was introduced in 2016 and got popular very fast. A similarity matrix is generated by application of a comparison function to each column in context H and query matrix U. A cell in the similarity matrix represents the similarity between a particular context word and a query word. Context to query and query to context attention is computed for both the directions from the similarity matrix. The output of the layer are the query-aware vector representations of the context words . The context matrix, the context to query attention matrix and the query to context attention matrix for merged into a single matrix by a concatenation function.

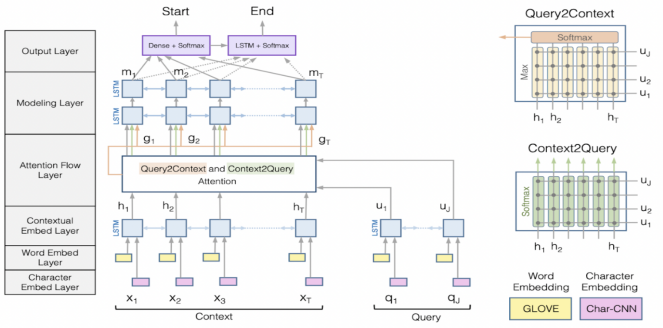


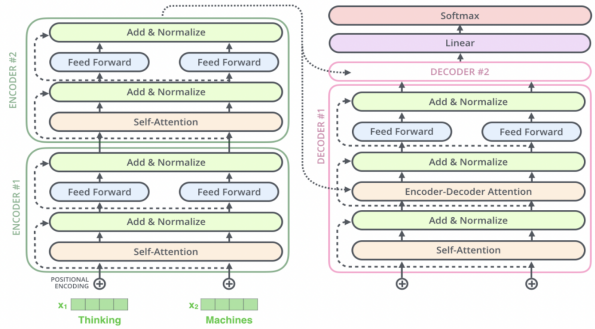
Figure 2: BiDAF architecture

5.2 Proposed Architecture:

- BERT(Bidirectional Encoder Representations from Transformers):

  The key innovation of the BERT model is applying bidirectional training of transformers to language modeling. BERT produces a state-of-the-art language model. At the highest level, its architecture is a trained stack of Transformer Encoders. Although we are using bi-directional LSTM in BIDAF, that captures forward and backward context but that is outperformed by the transformers in BERT. BERT gives better performance

when compared to BIDAF because it includes transformers that use self attention, whereas in BIDAF, we are using only attention, that is how much attention we have to pay to other words in the sequence. Self attention layer in BERT helps the encoder look at other words in the input sentence as it encodes a specific word. The entire input sequence is read at once, so it would be more correct to define it non-directional. BERT has two layers. a self-attention layer and a feed-forward neural network layer. The self-attention layer receives embeddings that are concatenated with positional encoding vectors in order to account for the order of input words. In contrast, the feed-forward layer is agnostic to dependencies between words, which means that operations can be executed in parallel in the feed-forward layer, leading to the impressive performance speedups associated with the Transformer. BERT model is pre-trained on 2 tasks. One of the tasks is Masked Language Model, where a percentage of the input tokens are masked and then predicted. The other task on which the BERT is pre-trained is Next Sentence Prediction, in order to learn the relationship between two sentences, which is not directly modeled by language modeling.



3: Transformers architecture

- DISTILBERT

As a part of an advancement, we will be updating the model to use DistilBERT which is a small, fast, cheap and light transformer model trained by distilling BERT Base. Compared to the BERT base uncased model, DistilBERT has 40% less parameters and runs 60% faster while preserving 95% of the BERT's performance. Knowledge Distillation could be seen as a transfer learning technique, which is a form of compression from a huge high precision model to a smaller one without losing too much in generalization. In DistilBERT the pooler and the token-type embeddings were removed. Reinforcement learning helps us to reduce the dimensions of a huge model and also reduce the training time by transferring knowledge between two models. Our smaller, faster and lighter model is cheaper to pretrain and we demonstrate its capabilities for on-device computations in a proof-of-concept experiment and a comparative on-device study. It uses a single linear layer, with no activation function, to reduce the 768 output dimension that comes from the DistilBERT backbone to 2.

5.3 Evaluation:

The performance of the models is measured by using two metrics.Given a paragraph and a question regarding it, the model provides us with an answer. The answer given by the model is a text selected from the paragraph itself. These scores are computed on individual question-answer pairs.

Exact match: The metric measures the percentage of predictions that match any one of the ground truth values in an exact way. This metric is simple but is fairly strict as well. For each pair of the question and answer, if the characters of prediction output from the model matches exactly with the characters of the true answers , the EM score would be zero. Let us take an example to describe the strictness of this metric. Example: Let us suppose we got an answer 'cat' for some question asked to the model. And the true answer is 'Black cat'. Then the EM score for this example would be zero. We can also say that exact match is a binary measure (true/false).

F1 score: In general F1 score is the harmonic mean of precision and recall. We always desire true positive and true negative in our model but we might also get false positive and false negative.Accuracy tells us the percentage of correctly classified data instances over total amount of data instances. Accuracy is generally used a lot but it is not a good measure when we have a non-balanced dataset.Precision also known as positive predictive value gives us the measure of the only relevant data instances. Precision the number of true positives divided by the number of true positives plus the number of false positives. Recall also known as sensitivity or true positive rate is a measure of the ability of a model to find all the relevant cases within a data instances.It is the number of true positives divided by the number of true positives plus the number of false negatives. In a good model , we want both values of precision and recall to be one. For this reason we use the F1 metric which takes both recall and precision into account. F1 score would be high when both the precision and recall are high . In this case , precision is the fraction of number of shared words to the total number of words in the model prediction and recall is the fraction of number of shared words to the total number of words in the ground truth. And F1 score is the harmonic mean of both of these values.

6 Schedule and Tools:

6.1 Schedule:

The group has decided to work collectively on each of the subtasks.

- Understanding and exploring the contents of Dataset (3 weeks)
- Understanding and implementation of baseline BiDAF Model (2-3 weeks)
- Understanding and implementing DistilBERT Model(2 weeks)

- Advancements on DistilBERT(2 weeks)
- Work on final report (1 week)

6.2 Tools:

We will start off using model using BIDAF architecture on SQuAD dataset which is trained specifically for QA Task in PyTorch framework. We will be using Google Colab for computation. Inbuilt pre-trained models like BERT will also be used in this project.

7 References:

[1] L. Kodra and E. Kajo, "Question Answering Systems: A Review on Present Developments, Challenges and Trends", International Journal of Advanced Computer Science and Applications, vol. 8, no. 9, 2017 [Online]. Available:https://thesai.org/Downloads/Volume8No9/Paper_31- Question_Answering_Systems_A_Review_on_Present_Development s.pdf. [Accessed: 22- May- 2018].

[2] E. Brill, J. Lin, M. Banko, S. Dumais and A. Ng, "Data-Intensive Question Answering", Trec.nist.gov, 2018. [Online]. Available: https://trec.nist.gov/pubs/trec10/papers/Trec2001Notebook.AskMSRF inal.pdf. [Accessed: 22- May- 2018].

[3] H. Madabushi and M. Lee, "High Accuracy Rule-based Question Classification using Question Syntax and Semantics", Aclweb.org, 2018. [Online]. Available: http://www.aclweb.org/anthology/C16- 1116. [Accessed: 23- May- 2018].

[4] E. Riloff and M. Thelen, "A Rule-based Question Answering System for Reading Comprehension Tests", 2018. [Online]. Available: https://pdfs.semanticscholar.org/4454/06b0d88ae965fa587cf5c167374 ff1bbc09a.pdf. [Accessed: 23- May- 2018].

[5] Trait Larson, Johnson (Heng) Gong, Josh Daniel "Providing a Simple Question Answering System by Mapping Questions to Questions.", Technical report, Department of Computer Science, Stanford University, 2006

[6] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. "DistilBERT, a distilled version of BERT: smaller, faster, cheap and lighter". In: CoRR abs/1910.01108 (2019). arXiv: 1910 . 01108. URL: http://arxiv.org/abs/1910.01108

[7] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. "SQuAD: 100,000+ Questions for Machine Comprehension of Text". In: (2016). URL: https://nlp.stanford.edu/pubs/rajpurkar2016squad.pdf

[8]. K. Dwivedia and V. Singh, "Research and reviews in question answering system," in Proceedings of International Conference on Computational Intelligence: Modeling Techniques and Applications,2013, pp. 417 – 424

[9] Senevirathne, K. U., N. S. Attanayake, A. W. M. H. Dhananjanie, W. A. S. U. Weragoda, A. Nugaliyadde, and S. Thelijjagoda. "Conditional Random Fields based named entity recognition for sinhala." In 2015 IEEE 10th International Conference on Industrial and Information Systems (ICIIS), pp. 302-307. IEEE, 2015

[10] Leon Derczynski, Alan Ritter, Sam Clark, Kalina Bontcheva,(2013) "Twitter Part-of-Speech Tagging for All: Overcoming Sparse and Noisy Data " Proceedings of Recent Advances in Natural Language Processing, pages 198–206

[11] J. Ramos(2003) "Using TF-IDF to Determine Word Relevance in Document Queries"Technical report, Department of Computer Science, Rutgers University, 2003

[12] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez, and K. Kochut. 2017. Text Summarization Techniques: A Brief Survey. ArXiv e-prints (2017). arXiv:1707.02268

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention is all you need. In Advances in Neural Information Processing Systems". In: (2017). url: http://papers.nips.cc/paper/7181-attention-is-all-you- need.pdf.

[14] Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. LUKE: Deep Contextualized Entity Representations with Entity-aware Self-attention. 2020. arXiv: 2010.01057 [cs.CL].