Research Article

Question Classification for Efficient QA System

K. P. Moholkar [1], S.H. Patil [2], J. Naveenkumar[3]

[1]Research Scholar, Bharti Vidyapeeth Deemed to be University, Pune/ India
[2]Bharti Vidyapeeth Deemed to be University, Pune/ India
[3]Bharti Vidyapeeth Deemed to be University, Pune/ India

Article History: Received: 10 November 2020; Revised: 12 January 2021; Accepted: 27 January 2021; Published online: 05 April 2021

Abstract: Natural Language Processing (NLP), a subfield of Artificial Intelligence (AI), supports the machine to understand and manipulate the human languages in different sectors. Subsequently, the Question and answering scheme using Machine learning is a challengeable task. For an efficient QA system, understanding the category of a question plays a pivot role in extracting suitable answer. Computers can answer questions requiring single, verifiable answers but fail to answer subjective question demanding deeper understanding of question. Subjective questions can take different forms entailing deeper, multidimensional understanding of context. Identifying the intent of the question helps to extract expected answer from a given passage. Pretrained language models (LMs) have demonstrated excellent results on many language tasks. The paper proposes model of deep learning architecture in hierarchical pattern to learn the semantic of question and extracting appropriate answer. The proposed method converts the given context to fine grained embedding to capture semantic and positional representation, identifies user intent and employs a encoder model to concentrate on answer span. The proposed methods show a remarkable improvement over existing system

Keywords: Question answering systems, Natural language processing, Information retrieval, Transformers, LSTM

1. Introduction

With available knowledge, a question answer system provides answer to question. Early, systems dealing with question answer were restricted to domains with limited ability. Recent developments have shown the emergence of powerful platforms for answering human questions automatically using databases or natural language document collection. (Chali, 2011, Dwivedi 2013,Ansari 2106,Lende 2016). QA systems have been categorized on basis of domains (Athenikos 2010, Kolomiyets 2011) and paradigms (A. Kalyanpur, 2012). The necessity of such systems has increased with need of short precise answers identifying the intent of question. ALEXA, IBM Watson (Kara 2012) has shown potential of such systems. A question answer system is composed of a triplet <c, q, a> where given a ‘q’ a question for expected ‘c’ context, extract ‘a’ the answer. The context can be anything from passage, curated knowledge graph, a document which is extracted by a search engine or hybrid corpora provided by user. The standard of answer depends understanding the context and the intent of question. The question can range from ‘wh’ questions like who, when, where, why, how, counting, yes/no, decision making, factoid, descriptive and likewise. A question can also have a timeline to infer the answer. Similarly, the answers vary from single word, short answer, yes/no or descriptive.

In recent days, deep learning model has been applied successfully for various applications. The QA selection process is done using machine learning algorithms such as SVM or deep learning algorithms such as CNN, RNN (kara,2012). The conventional method of question system has data retrieval and handcrafted rules (Weise 2017). Author (Moholkar,2019) has proposed a hybrid model for question answer system using ANN and BiLSTM. The accomplishment of such systems is limited due to handcrafted rule and imbalance nature of data. Question processing is a two-step process. The first step analyses the construction of user question and in second step transforming it to meaningfully query. The nature of answer depends on given context.

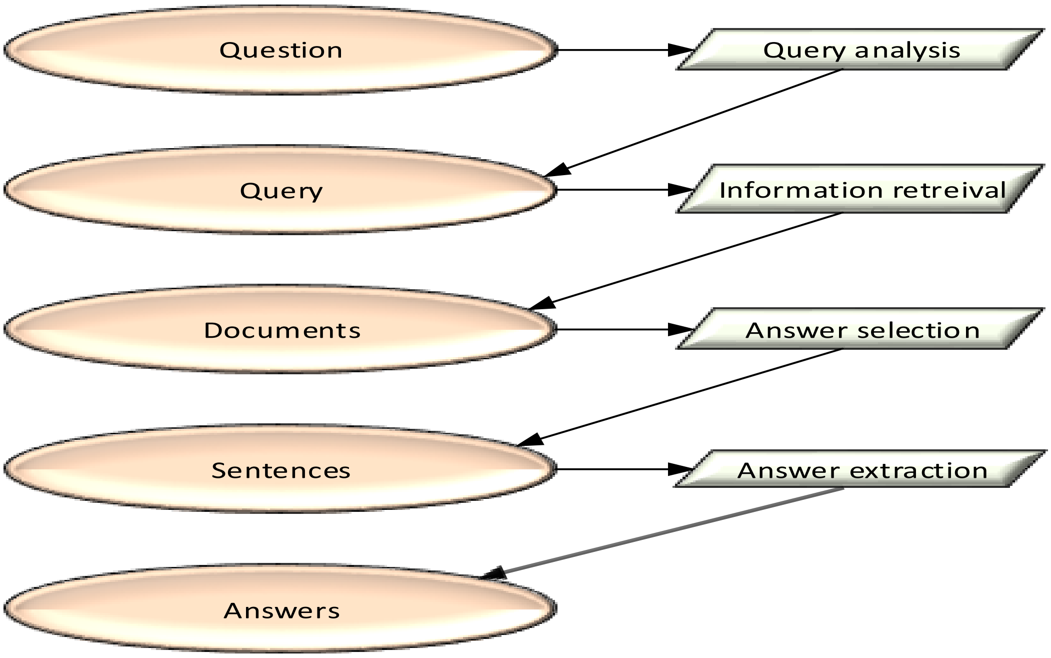


Figure 1. Pipeline architecture for QA model

1.1. Problem Statement

The aim behind this research is to separate the different question types based on semantic classes depending on given context. We assume that questions are answered based on given context considering the timeline. Answering question like “How many balls are there?” or “In which direction did Sam go? requires the system to understand the location of information and perform some operation to reach to conclusion. These operations might include addition, subtraction, comparison or deduction to reach to conclusion. The major contribution of the paper is as follows:1. Dataset available till date have different question answer pairs with context but type of questions is not identified. This paper proposes a strategy to identify question category. 2. The paper proposes stack transformer model by a fine tuning pre trained model and ensemble approach for identifying question intent given the context and question and extracts answers.

1.2. Literature Review

Wiese et al.2017, introduced a neural framework to recognize answer ranges in the highest quality level pieces gave by BioASQ for question noting (Task B) rivalry. Their framework was first pretrained on the SQUAD, a 100 000-tidbit question noting dataset created from Wikipedia abstracts by publicly supporting.Then, they fine-tuned their system on BioASQ 5b factoid and list subset achieving state of the art performance on the factoid questions. She (Preena,2019) propose a memory network using long short-term memory (LSTM) model for text-based question answering. The proposed model uses a score function to match each pair of question sentence to predict the expected answer. In paper(Kembhavi et al,2016) proposed QA baseline machine comprehension models. These models which were based on BiDAF proposed by (Seo et al ,2014), Memory net (Weston et al,2018) gave modest results for true/false and MCQ type question answer. With the advancements in field of NLP and large-scale models like ELMo proposed by (Peters et al 2018), ULMift (Howard and Ruder, 2018) large amount of text can now be processed. The advent of BERT (Devlin et al ,2018) and RoBERTa (Liu et al 2019), which were trained on Wikipedia and google book corpus of 10K books, helped to train model to learn different languages make prediction like identifying missing word, generating next word in sequence etc. Transformers have also proved their mettle as soft reasoners (Fabio, 2019), exhibiting capabilities for natural language inference. Furthermore, whilst learning linguistic information, transformers have shown to capture semantic knowledge and general understanding of the world from the training text (Pranav et.al 2019), including a notion of common sense that can be useful in question answering. RNN algorithm required to decode entire sequence in single context vector to reason about previous event. RNN lacked the memory element need to capture meaning of a sentence. LSTM and GRU provided memory units with help of gates to fetch past information. These algorithms failed to capture long range dependencies. Despite of GRUs and LSTM, RNN needed attention mechanism to deal with long range dependencies. RNN suffered from sequential execution of task making it incapable of harnessing the power of GPUs. Our approach is the first to influence the language understanding and reasoning capabilities of existing transformer language models for recognizing the intent, the knowledge and reasoning required by the question.

2. Method

The proposed QA system can be modelled as: $Q = \{q_1, q_2, \dots \dots, q_n\}$, $C = \{s_1, s_2, \dots \dots, s_n\}$, $A = \{a_1, a_2, \dots \dots, a_n\}$where Q is a set of questions asked on context C, to extract answer A. Context C is a set of sentences containing answer snippets. A is set of answers varying from one word to long answers. Each question is classified into one of 20 different classes Q ε {i1, i2…….,120}. We have modelled the QA task in two phases. Phase I deals with question intent classification. We apply fine tune BERT transformers (Liu 2019) to question intent, treating the task as question classification. Bidirectional Encoder Representations from Transformers (BERT) provides a pre-trained deep neural model over very large corpus of unannotated texts. The model is extended to QA scenario by finetuning the entire architecture. Phase II deals with identifying appropriate model for answer extraction depending on question intent.

BERT Models pre-trained on large corpus of text have to be fine-tuned on the Q&A task using Q&A datasets. When a question is posed, Candidate passages is extracted based on the scored based on relevance. The top N passages will be input into the model to generate the potential answers (for every passage) along with the confidence scores. The question category is identified and suitable model is selected to generate answers. A [CLS] token is embedded toward the start of the initial sentence and a [SEP] token is embedded toward the finish of each sentence.

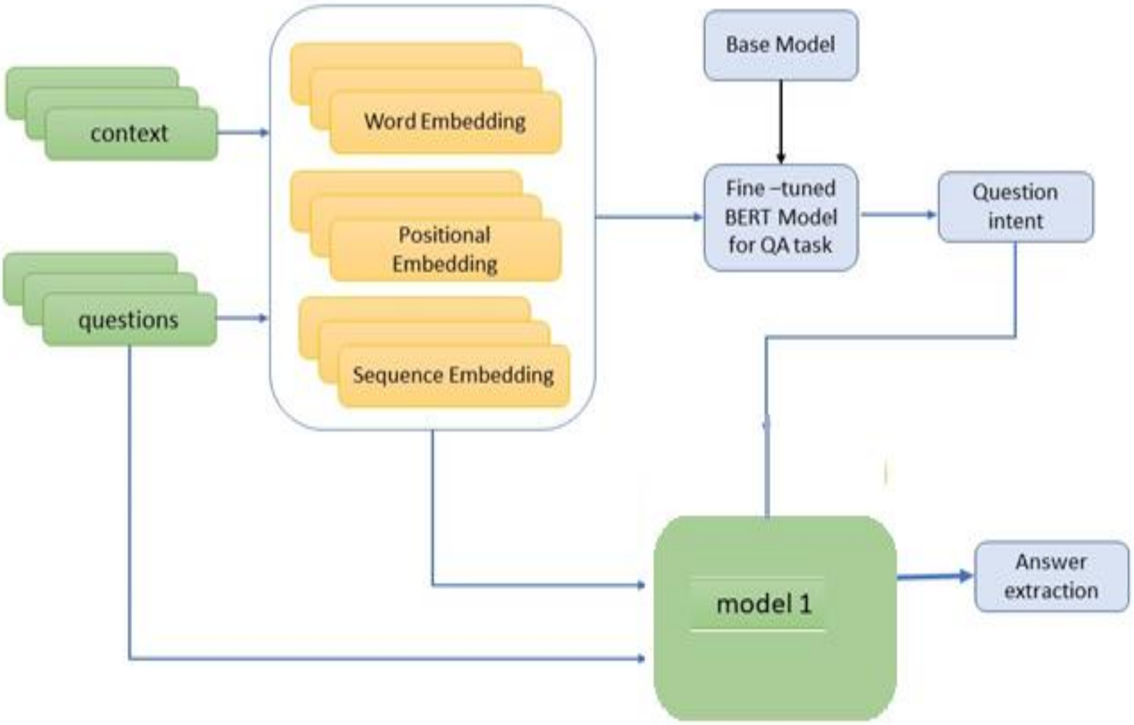


Figure 2. Proposed architecture

Given a question q ∈ Q with answer Ai and background knowledge C, we pass the following sentences S to the transformer:

$$seq(C, Q * Ai) = [CLS]\ C\ [SEP] Q * Ai\ [SEP] \dots \dots . (1)$$

where $QAi = [q, ai]$. Similarly, the correlation between question q, context c and answer the previous method, s is obtained as:

$$seq(q, Cs) = [CLS]\ q\ [SEP]\ Cs\ [SEP] \dots \dots \dots . \dots \dots (2)$$

The input is feed to a word embedding layer to create a vector representation of every word. The location of word is an important information required to identify the semantic of a sentence. The vector representation is then clubbed with positional information.

$$PE_{(pos,21)} = \sin\left(\frac{pos}{10000^{(2i/d_{model})}}\right) \ldots\ldots\ldots.\ldots\ldots(3)$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{(\frac{2i}{d_{model}})}}\right) \ldots\ldots\ldots\ldots\ldots.(4)$$

Thinking about the straight properties, cos work was utilized for odd record and sine work was utilized for even list of information vector. The subsequent vectors were added to create positional data. BERT gives these sentences inserting and contextualized embeddings every word in a sentence. The attention mechanism is designed as

$c = \sum_j \alpha_j h_j \;\ldots\ldots..(5)$ where $\sum \alpha_j = 1$

Where α is a vector representation of input. In the proposed technique, consideration and combination are directed evenly and vertically across layers at various degrees of granularity among question and passage. First encode the inquiry and passage with fine-grained language embeddings, to all the more likely catch the separate portrayals at semantic level. The base model is ready after training BERT with initial configuration. The pre-training stage acquires an expressive and robust language model, where only the encoder is used and applies masked language model to construct original sentence. In proposed model, a transformer T will produce one vector for each token in s, including [CLS], whose vector we denote as T[CLS] (s), which we use as a pooled representation of the whole sequence. BERT pretrained uncased model consists of 12 Transformer blocks and 12 self-attention heads by taking an input of a sequence of 512 tokens and outputs the representations of the sequence with 110M parameters. Transformer includes an encoder that reads the text input and a decoder that produces a prediction for the targeted task.

$$MultiHead(Q,K,V) = Concat\;(head_1, \ldots\ldots.head_h)W^o \;\ldots\ldots\ldots\ldots.(6)$$

$where\ head_i = Attention\ (QW_i^Q\ \ ,KW_i^K, VW_i^V)$ where the projections are parameter matrices $W_i^Q \in \mathbb{R}^{d_{model}*d_k}$ $,W_i^K \in \mathbb{R}^{d_{model}*d_v}$ $,\ W^O \in \mathbb{R}^{hd_v*d_{model}}$

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \ldots\ldots..(7)$$

At that point it proposes a multi-granularity combination way to deal with completely intertwine data from both worldwide and went to portrayals. Structure the portrayal the inquiry goal is anticipated. A fine tuning of pretrained parameters is done. The model at that point piles up profound learning designs to give particular comprehension at each degree of the report order. Let the encoded context and question be$C' = \{\,c_1, c_2, \ldots\ldots., c_{3\}}$ and $Q' = \{q_{1,}q_{2,\ldots\ldots,}q_n\}$. Let the target labels $Y = \{y_1, y_2, \ldots\ldots, y_n\}$ where $y_i \epsilon$ 1,2,3, .….10 .The probability $\mathrm{P}(y_j|c_iq_i)$ of class y_i given sample c_iq_i is a multinominal distribution. The output of BERT converts the vector H_([CLS]) into conditional probability over predefined categories. This output forms the input to dense network followed by a softmax layer. The model predicts the question intent as

$\mathrm{P}(Y = y_j|X = c_iq_i\ ,w) = \exp(w_k^T c_iq_i)\ /\ \sum_{k'}^{K} \exp\ (w_k^T c_iq_i)$……… (8)

where K =10, w, the trainable parameter. The output is predicted using the largest

$$Y_c = argmax(Py_i|H_{[CLS]}, w) \ldots\ldots\ldots\ldots\ldots\ldots\ldots..\ldots.(9)$$

In stage II, the anticipated expectation is utilized to discover reasonable response to the offered conversation starter. Base BERT model backings entries of greatest length 512 in particular. As the section arrives at the most extreme breaking point, a couple of inquiries may not be tended to precisely. To defeat this restriction, the section is distanced into covering bunches through a delimiter '\n'. Each part is questioned to discover the appropriate response. At last, it acquaints a progressive consideration network with centres around the appropriate response range logically with staggered delicate arrangement.

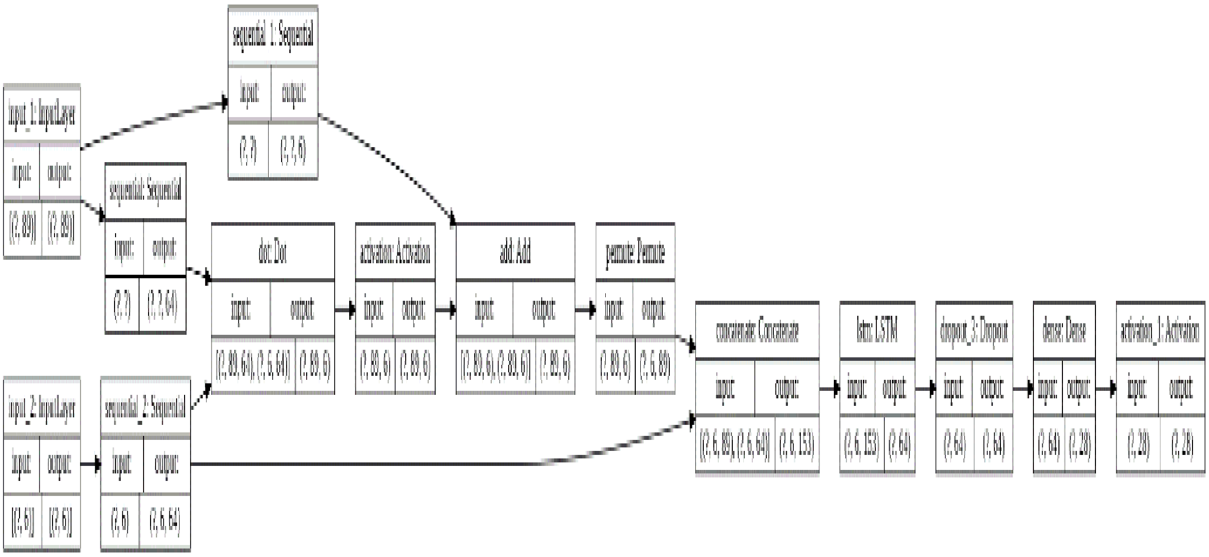


Figure 3. Answer generation model

2.1. Data Collection Tools

A numerous QA dataset consist of abundant documents or passages with substantial length. Every passage would have different questions and answers are chunk of the passage. The goal is to find the correct text span considering the type of question posed. Project of Facebook AI Research named bAbI (Lan 2019) consists of 10K training and testing dataset for automatic text understanding and reasoning. The custom dataset for proposed work is build by extracting task types from bAbI. The custom dataset consists for more than 10k+ context, question answer pair with category of questions. The custom dataset at present has 10 categories of question viz factoid QA with one, two, three supporting facts, two/three object relation, yes/no, counting, list, conjunction, indefinite knowledge, time manipulation, basic deduction and induction, path finding, abbreviation, description, motivation and positional reasoning. The customized dataset contains 200 samples for each type of question. A total of 3800 records are considered for elementary stage. Sample questions in dataset are

Context: “Fred is either in the school or the park. Mary went back to the office. Is Mary in the office?
Answer: yes
Question type: indefinite knowledge
Context: Sandra and Mary went back to the office. Daniel and Sandra went to the bedroom. Where is Sandra?
Answer: bedroom
Question intent: Conjunction
Context: Bill travelled to the office. Bill picked up the football there. Bill went to the bedroom. Bill gave the football to Fred.
Question: What did Bill give to Fred?
Answer: football
Question intent: 3 arguments
Context: This morning Fred went to the kitchen. Fred journeyed to the bedroom yesterday. Mary travelled to the bedroom this morning. Yesterday Mary went to the cinema.
Question: Where was Mary before the bedroom?
Answer: Cinema
Question intent: time reasoning
Context: Sumit is tired.
Question: Where will Sumit go?
Answer: bedroom
Question intent: entity motivation

The question investigation uncovers that intent of question describes the type of answer expected. Answer to be extracted Depends on question type. The difference in question need problem-specific attention. The anomalies in the question acts as a hindrance in the formation of robust rules for the classification of questions into their correct type thereby affecting the answer quality.

2.2. Experimental Evaluation

The proposed model is evaluated on vali 10K babi dataset dealing with 10 different tasks. The model employs Tesla T4 Nvidia GPU. The ratio of train-validation dataset is 90-10 and separate file is created for testing the model. Maximum length of passage is considered 200 per context with overlapping. Training batch size is set to 8 and validation to 4. The model is trained for 25 epochs with learning rate of 1e-05. AdamW is the optimizer user with learning rate of 3e-5. We calculate weights of each class to manage the imbalance. The categorical entropy loss function is used to track model loss. The training 98% and test accuracy is 95% for intent classification task. The table 1 shows the training loss for different settings:

Table 1. Parameter tuning for customized BERT model for question intent detection

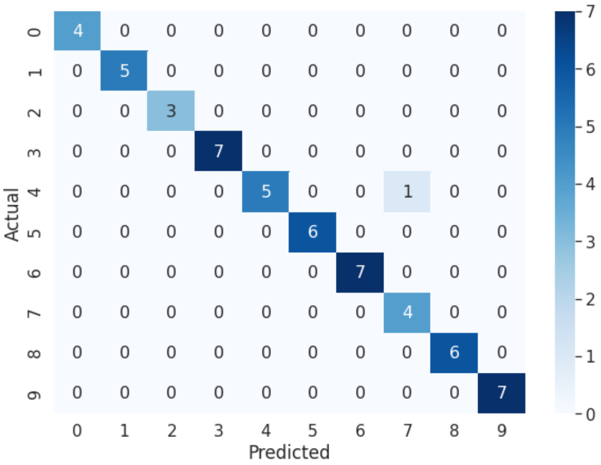
Epoch	**Learning rate**	**Batch size**	**Training loss**	**Validation loss**
5	3e-05	8	0.133	0.223
5	5e-5	8	0.193	0.217
5	2e-5	8	0.042	0.249
5	3e-05	16	0.128	0.231
5	5e-5	16	0.172	0.213
5	2e-5	16	0.039	0.238
5	1e-05	16	0.011	0.347

For experimentation it was observed that validation loss increased after 2nd epoch for learning rate of 5e-5 and batch size of 8. The training loss is minimum for a learning rate of 2e-5 and batch size of 8.Table 2 presents the test data score for the fine tuned BERT model.

Table 2. Test data result for question intent classification

```
          precision    recall  f1-score   support
       0       1.00      1.00      1.00         4
       1       1.00      1.00      1.00         5
       2       1.00      1.00      1.00         3
       3       1.00      1.00      1.00         7
       4       1.00      0.83      0.91         6
       5       1.00      1.00      1.00         6
       6       1.00      1.00      1.00         7
       7       0.80      1.00      0.89         4
       8       1.00      1.00      1.00         6
       9       1.00      1.00      1.00         7

    accuracy                           0.98        55
   macro avg       0.98      0.98      0.98        55
weighted avg       0.99      0.98      0.98        55
```



Actual \ Predicted	0	1	2	3	4	5	6	7	8	9
0	4	0	0	0	0	0	0	0	0	0
1	0	5	0	0	0	0	0	0	0	0
2	0	0	3	0	0	0	0	0	0	0
3	0	0	0	7	0	0	0	0	0	0
4	0	0	0	0	5	0	0	1	0	0
5	0	0	0	0	0	6	0	0	0	0
6	0	0	0	0	0	0	7	0	0	0
7	0	0	0	0	0	0	0	4	0	0
8	0	0	0	0	0	0	0	0	6	0
9	0	0	0	0	0	0	0	0	0	7

Figure 4. Confusion matrix for classification task

While testing for answer generation, it is observed that BERT model does not perform well. The graphs in figure 6 shows the performance of system for QA task. Testing BERT on different task passage, we came to conclusion that proposed encoder model outperforms BERT in case of quality of answer generated.

The following graph illustrates the efficiency of our model over BERT and ALBERT models.

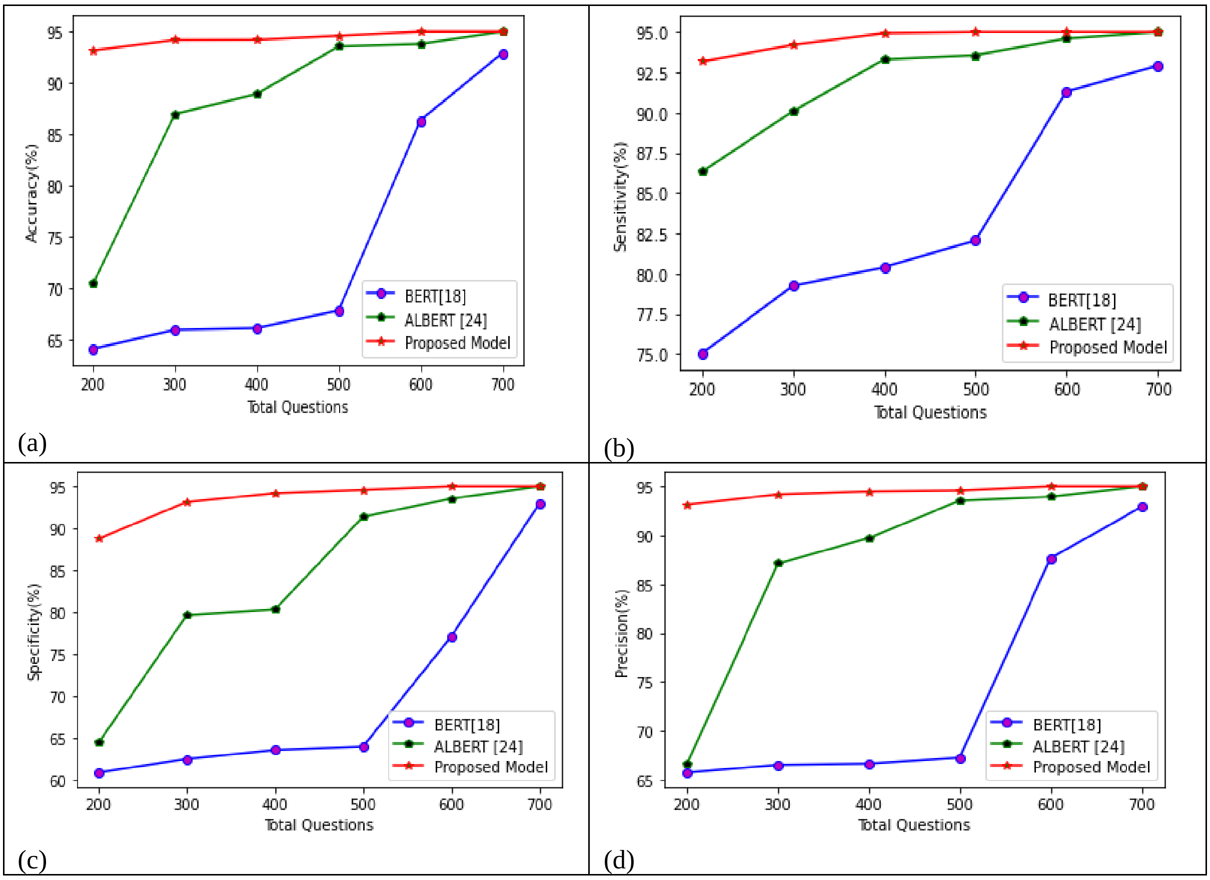


Figure 5. (a) Accuracy (b) Sensitivity (c) Specificity (d) Precision graphs for training data

The table 3 demonstrates the performance of the model for different question intent tasks. It was observed that model performs is high for question intents which consider single or two supporting facts, or two or three arguments . The performance of model is satisfactory for counting or listing operations but fail to perform on yes no or question requiring indefinite knowledge base for making conclusion. The overall performance of model is 83.58% which is a fair performance considering variety of task.

Table 3. Demonstrates the Performance of the Model for Different Question Intent Tasks

Task	**Accuracy %**
single_supporting_fact	95
two_supporting_facts	92
three-supporting-facts	90.8
two-arg-relations	92
three-arg-relations	96
yes-no-questions	60
Counting	84
lists-sets	84
simple-negation	63
indefinite-knowledge	49
Average performance	**83.58**

3. Discussion and Conclusion

This paper recommends a methodology to solve QA model, where question intent plays a main role for precise answer extraction. The question intent is mined for enhancing the ability of QA system. The hybrid BERT RNN approach which could handle both formative and summative questions. Finally, the results were validated by analysing the efficiency of the proposed model. From the outcomes, the answer generated by proposed model are of better quality than BERT, ALBERT models. The presented model also performs better than BERT and ALBRT model for all 10 tasks with a 83.58% average accuracy across all given tasks. In further, we further decide to fine tune model for yes/no or true false answer and questions requiring indefinite knowledge base without compromising on other task performance. We plan to analyse the capacity of model by increasing the task and size of passage in future.

References

1. Y. Chali, S.A. Hasan, S.R. Joty (2011) Improving graph-based random walks for complex question answering using syntactic, shallow semantic and extended string subsequence kernels Inf. Process. Manage., 47 (6), pp. 843-855, 10.1016/j.ipm.2010.10.002
2. S.K. Dwivedi, V. Singh (2013), *Research and reviews in question answering system*, Procedia Technol., 10, pp. 417-424, 10.1016/j.protcy.2013.12.378
3. A. Ansari, M. Maknojia, A. Shaikh (2016), *Intelligent question answering system based on artificial neural network* 2016 IEEE International Conference on Engineering and Technology (ICETECH), IEEE, pp. 758-763, 10.1109/icetech.2016.7569350
4. S.P. Lende, M. Raghuwanshi (2016) *Question answering system on education acts using nlp techniques* World Conference on Futuristic Trends in Research and Innovation for Social Welfare (Start-up Conclave), IEEE (2016), pp. 1-6, 10.1109/startup.2016.7583963
5. S.J. Athenikos, H. Han (2010) , *Biomedical question answering: a survey*, Computer Methods Programs Biomed., 99 (1) , pp. 1-24, 10.1016/j.cmpb.2009.10.003
6. O. Kolomiyets, M.-F. Moens(2011) *A survey on question answering technology from an information retrieval perspective* Inf. Sci., 181 (24) , pp. 5412-5434,
7. Kalyanpur, B.K. Boguraev, S. Patwardhan, J.W. Murdock, A. Lally, C. Welty, J.M. Prager, B. Coppola, A. Fokoue-Nkoutche, L. Zhang, *Structured data and inference in deepqa* IBM J. Res. Dev., 56 (3.4) (2012),
8. D.A. Ferrucci Introduction to this is Watson IBM J. Res. Dev., 56 (3.4) (2012), Al-Ayyoub, Mahmoud, et al. "Deep learning for Arabic NLP: A survey." Journal of computational science 26 (2018): 522-531
9. Kara, Soner, et al. (2012) "*An ontology-based retrieval system using semantic indexing."* Information Systems 37.4 : 294-305.
10. Wiese G., Weissenborn D. and Neves M.L. (2017) *Neural Question Answering at Bioasq* 5b. CoRR, abs/1706.08568.
11. M P, Preena and Joseph, Shibily(2019), *Question Answering Using Deep Learning* (September 4, 2019). In proceedings of the International Conference on Systems, Energy & Environment (ICSEE) 2019, GCE Kannur, Kerala, July 2019, Available at SSRN: https://ssrn.com/abstract=3447734 or http://dx.doi.org/10.2139/ssrn.3447734
12. Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi.(2016.) *A diagram is worth a dozen images*. In ECCV.

13. Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. (2017). *Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension*. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 5376–5384
14. Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. (2017). *Bidirectional attention flow for machine comprehension.* CoRR, abs/1611.01603.
15. Jason Weston, Sumit Chopra, and Antoine Bordes. (2014). *Memory networks*. CoRR, abs/1410.3916.
16. Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. (2018). *Deep contextualized word representations.* In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.
17. Jeremy Howard and Sebastian Ruder.(2018). *Finetuned language models for text classification.* CoRR, abs/1801.06146
18. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. (2018). *BERT: pre-training of deep bidirectional transformers for language understanding.* CoRR, abs/1810.04805.
19. Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. (2019). *Roberta: A robustly optimized bert pretraining approach*. ArXiv, abs/1907.11692
20. Moholkar, Kavita, and Suhas Patil. (2019) *"Hybrid CNN-LSTM Model for Answer Identification."* International Journal of Recent Technology and Engineering (IJRTE)ISSN: 2277-3878, Volume-8 Issue-3
21. Peter Clark, Oyvind Tafjord, and Kyle Richardson. (2020) *Transformers as soft reasoners over language.* ArXiv, abs/2002.05867.
22. Fabio Petroni, Tim Rocktaschel, Sebastian Riedel,(2019) ¨ Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. *Language models as knowledge bases?* In Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLPIJCNLP), pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.
23. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang.(2016). Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250.
24. Jason Weston, Antoine Bordes, Sumit Chopra, Tomas Mikolov, Alexander M. Rush, Bart van Merriënboer, *"Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks",* arXiv:1502.05698 [cs.AI].
25. Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel Piyush Sharma, Radu Soricut,(2019) *"ALBERT: A LITE BERT FOR SELF-SUPERVISED LEARNING OF LANGUAGE REPRESENTATIONS",* 30 Oct 2019.
26. Tan et al.(2018), *"Context-Aware Answer Sentence Selection With Hierarchical Gated Recurrent Neural Networks,"* in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 3, pp. 540-549, March 2018.