Question and Answer Model on Textual Data Using Deep Learning

[1]Mohammad Aalam, [2]Mohammad Suaib ,[3]MohdHaroon

[1]M.Tech Scholar, [2]Associate Professor,
[1]Department of Computer Science & Engineering, Integral University Lucknow, UP, India.

Abstract : Question answering (QA) is a high-quality research setback in NLP (Natural Language Processing). Regardless of being one of the oldest research areas, QA has an appliance in an extensive range of tasks, for examplereclaiming information as well as extracting entities. In recent times, QA has also been utilized to build up dialog systems and chatbots designed to imitate human conversation. With advances in deep learning, neural network variants have become a dominant structure for many natural language programming(NLP) tasks.In this paper, we aim to design a system that is based on deep learning approach to generate Question and Answering Automatically.We exploit NLP, machine learning and KNN for the generating question and their best answer.

IndexTerms **- Question answering, NLP, Deep Mining, chatbots.**

1.INTRODUCTION

1.1. Deep Learning

Deep learning permits arithmetic models consisting of multiple layers of processing to learn data illustration at manifold levels of abstraction. These methods have greatly improved the latest technology of speech recognition, visual recognition, object discovery and many other areas such as drug discovery and genomics [1]. Deep learning finds out a multifaceted structure in hefty data sets utilizing the backpropagation algorithm to designate how the device changes its internal parameters that are utilized to calculate representation in each layer of representation in the previous layer. Deep metaphysical networks have caused breakthroughs in image, video, speech and sound processing, while repetitive networks have highlighted serial data such as text and speech [2].

Automated learning technology operates numerous aspects of modern society: from onlinesearch to filtering content on social networks to recommendations on e-commerce sites, and they are progressively present in consumer products such as cameras and smartphones. automated learning systems are utilized to recognize objects in images, transfer speech to text, match news items, publications or products with user interests, and recognize relevant search results. Increasingly, these applications utilize a class of techniques called deep learning.

1.2. Question Answering

The answer to the questions has recently received interest from information retrieval, information extraction, automated learning, and natural language processing communities. The answer-to-question system is intended to retrieve the answers to questions rather than the most complete documents or sections, as most information retrieval systems currently do[3].

There are a number of characteristics of the area of answering questions and the design of a new approach that poses particular challenges in the use of automated learning to classify responses and assess confidence. The most important of which are as follows:

- Some candidate answer sets may be equal, while others may be relevant; in the latter case, one of these answers may be relevant[4].
- The importance of different features may vary radically in different questions and question classes (for example, puzzles and language translation), and there may be little training data available for some categories of questions.
- Some features are more or less valuable at different stages of arrangement. For example, after the initial classification, some features that have little impact on the initial order may be disproportionately important in making a better distinction between high-ranking candidates.
- Features are very heterogeneous. It is derived from a variety of distinct algorithms developed independently. Therefore, many features are not normalized in any way, and their distribution may vary significantly depending on the characteristics of the question. Attribute values are often missing and some occur in the training group.
- There is a big grade imbalance. The system may find a large number of candidate responses using text search [5]. However, few of these answers are correct.

- Challenges in deploying automated learning In a flexible development environment, automated learning to assess trust is vital to Watson's development from its early beginnings when it had only two advantages. As part of the Watson Method [4], experiments are conducted to measure the effect of adding or modifying any component. To evaluate the effect correctly, you must retrain forms with changes included. The confidence estimation framework was used in more than 7,000 experiments.This is a very different environment for machine learning deployment compared to traditional settings where modules are trained on fixed feature sets and allow for extensive manual tuning. As part of our work, the framework should be as ready and automated as possible, leading the framework's facilities to lose value, eliminate disparate features, and use powerful machine learning techniques without a sensitive parameter identification.

2.LITERATURE REVIEW

Rich linguistic connotations have been applied to modern QA matching models by Yi et al. [6]. These models correspond to the semantic relationships of the adjacent words in the QA pairs using a combination of lexical-semantic resources such as WordNet with distribution representations to capture semantic similarity. This policy results in a series of features of the sentence pairs, which are then entered into a traditional workbook. A different version of this idea can also be found in Severyn and Moschitti [7], which used SVM with tree beads to automatically learn the features of shallow analysis trees rather than relying on external resources and sacrificing semantic information for the simplicity of the model. The authors combined these two approaches by proposing a clearly rich model without the need to architect features or annotated external resources.

Yi et al. [8] Models built to question the individual relationship with a trigonometric knowledge base. In the same direction, Borders et al. [8] [9] utilized a type of Siamese network to learn to ask questions and answer pairs in a common space. Ayer et al. [10] worked on a contest quiz job, a task to answer questions requiring the identification of an entity as described in a series of sentences. They formed the semantic composition with the recombinant neural network. These two tasks differ from the work presented here that the choice of answers may require a multiple-question question-answer to answer sentences that also contain several concepts and relationships.

3. PROPOSED METHODOLOGY

The block diagram of the proposed approach is depicted in figure 1. The proposed approach is as follows:

Step 1: We create a trained dataset of questions and words.

Step 2: Our algorithm read the paragraph and find out the keywords. Which are available in our words dataset.

Step 3: Apply KNN (the nearest neighbor algorithm) based on a trained data set in a paragraph and create optional questions. K-Nearest Neighbors (kNN) identifies the most frequently asked training questions for each test question, then uses well-known labels for similar training questions to predict the classification of the test question. The similarity between two questions can be calculated as a number of interrelated features, as the inverse ofEuclidean distance between the vectors of the feature.

Step 4: For the specific query q, neighbors closest to K are found among the training questions and use the nearest K neighbors classes for the weight category candidates. The degree of similarity of each question close to the test question is used as a weight for the neighborhood question categories. If many of the closest K neighbors share a row, the weights for each neighborhood of that class will be added together, and the resulting weighted amount is used as a probability level for that category for the test question. By sorting the scores of the candidate categories, an ordered list is obtained for the test question.

$$score(q, c_i) = \sum_{q_j \in KNN(q)} sim(q, q_j)\delta(q_j, c_i)$$

Here KKN(q) indicates the set of K-nearest neighbors of test question q. $\delta(q_j,c_i)$ is the classification for q_j with respect to class c_i, that is,

$$\delta(q_j, c_i) = \begin{cases} 1 & q_j \in c_i \\ 0 & q_j \notin c_i \end{cases}$$

For test question q, it should be assigned the class that has the highest resulting weighted sum.

The architecture of the proposed approach is shown in figure 2.

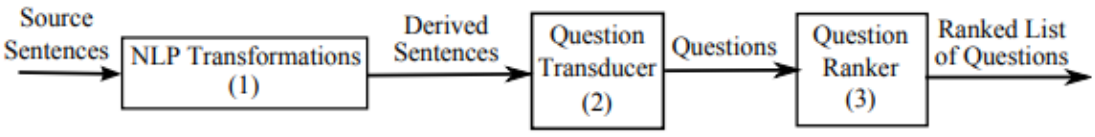

Figure 1:Block diagram of the proposed approach

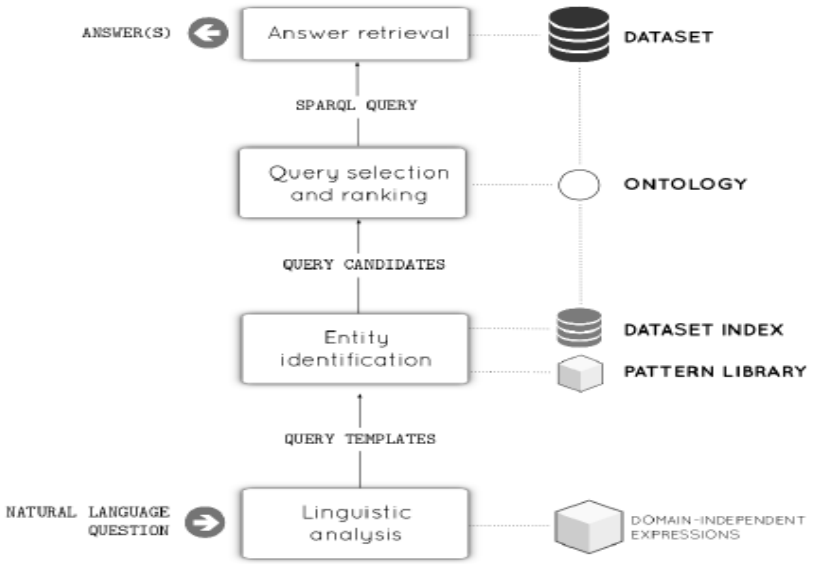

Figure 2: Architecture of the proposed approach

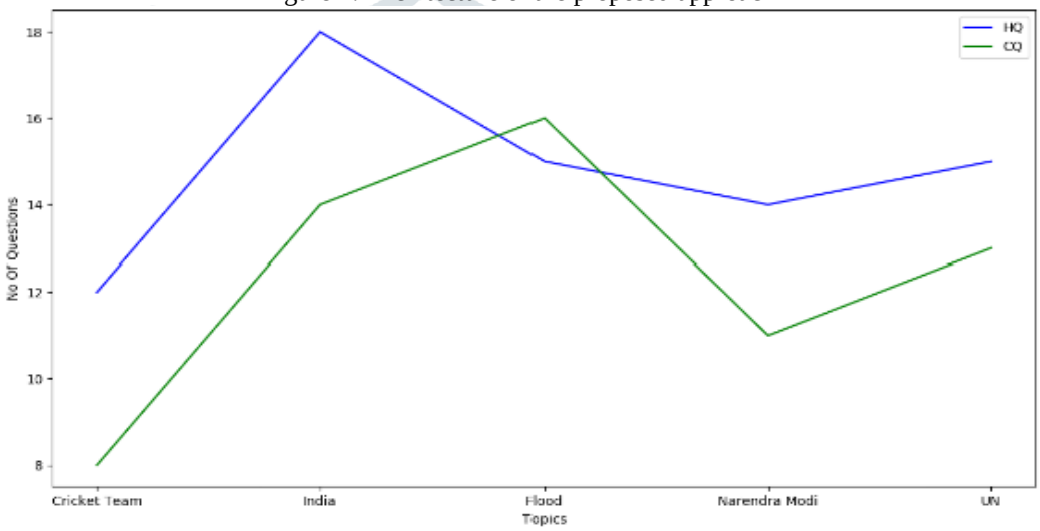

Figure 3: Comparison ofthe number of questions generated by human and machine

S. N.	Question classes	Question processing	Question context	Data source and QA	Answer formulation	Real-time QA	Interactive QA	Answer extraction
1	Long	Interrogative	universal	Relevant data	Simple extraction	Simple question	Ambiguous question	Specific type information
2	Short	Assertive	Previous knowledge	Exhaustic	Partial answer	Complex question	Clear doubt	Data source
3	Factoid	Wh-type	location	Base information	Extracted from various sources	Wait for hour	Boring question ignore	complexity
4	Not factoid	Semantic	proceeding	web	combination	Get answer	Optional	Expectation of user
5	Deeper fact	Exclamatory	person	Database	Result as natural	Jeopardy	place	Extraction of time or date
6	Simple fact	Imperative	Rarely asked question	Structured	Generate answer	Developed architecture	Answer question	Extraction of name
7	Deeper understanding	Understand question	interested	Queries	natural	Watson	Nonambiguous	Extraction of place

Table 1: Categorization of question classes and answer extraction

4. EXPERIMENTAL EVALUATION AND ANALYSIS

In the running of the algorithm,we take a few paragraphs in different fields like sports, health, education, etc. First of all, we check how many questions are generated manually. After that,we process this paragraph in my machine learning algorithm and it generates questions. The comparison between a number of questions generated by human and machine is depicted in figure 3.

Since this system gives better accuracy as compared to manually question generation. When we give any paragraph like sports then the machine gives 83% accuracy.

5. CONCLUSION

Originally, question answering had a strong focus on textual data sources to find answers, relying mostly on information retrieval techniques. In the early 1970s, questions began to be addressed in the integration of structured data and the development of natural language interfaces for databases. At the present time, as knowledge grows in the associated open data cloud, interest in answering the question about structured data quickly re-takes interest. In this paper, we designed a system that is based on deep learning approach to generate Question and Answering Automatically. We exploited NLP, machine learning and KNN for the generating question and their best answer.

REFERENCES

[1] Eraslan G, Avsec Ž, Gagneur J, Theis FJ. Deep learning: new computational modeling techniques for genomics. Nature Reviews Genetics. 10:1, 2019.
[2] Shobhit Srivastava, MohdHaroon, AbhishekBajaj :"Web document information extraction using class attribute approach", 2013 4th International Conference on Computer and Communication Technology (ICCCT), Pages 17-22. Publication date2013/9/20.
[3] R Khan, M Haroon, MS Husain, "Different technique of load balancing in distributed system": A review paper, 2015 Global Conference on Communication Technologies (GCCT), Pages 371-375, Publication date 2015/4/23
[4] Mohammad Haroon, Mohd Husain, "Interest Attentive Dynamic Load Balancing in distributed systems", 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom), Pages 1116-1120, Publication date2015/3/11
[5] Pouyanfar, S., Sadiq, S., Yan, Y., Tian, H., Tao, Y., Reyes, M.P., Shyu, M.L., Chen, S.C. and Iyengar, S.S., A survey on deep learning: Algorithms, techniques, and applications. ACM Computing Surveys (CSUR), 51(5), p.92, 2018.
[6] J. Chu-Carroll, J. Fan, B. K. Boguraev, D. Carmel, D. Sheinwald, and C. Welty, Finding needles in the haystack: Search andcandidate generation, IBM J. Res. & Dev., vol. 56, no. 3/4, Paper 6, pp. 6:1–6:12, 2012.
[7] D. A. Ferrucci, Introduction to 'This is Watson', IBM J. Res. & Dev., vol. 56, no. 3/4, Paper 1, pp. 1:1–1:15, 2012.
[8] Wen-tau Yih, Ming-Wei Chang, Christopher Meek, and AndrzejPastusiak, Question answeringusing enhanced lexical semantic models. In Proceedings of ACL, 2013.
[9] Wen-tau Yih, Xiaodong He, and ChristopherMeek, Semantic parsing for single-relation question-answering. In Proceedings of ACL, 2014.
[10] AliakseiSeveryn and Alessandro Moschitti, Automatic feature engineering for answer selectionand extraction. In EMNLP, 2013.
[11] Antoine Bordes, Sumit Chopra, and JasonWeston, Question answering with subgraphembeddings, In Proceedings of EMNLP, 2014.
[12] Antoine Bordes, Jason Weston, and Nicolas Usunier, Open question answering with weakly-supervised embedding models. In Proceedings of ECML, 2014.
[13] MohitIyyer, Jordan Boyd-Graber, Leonardo Claudino, Richard Socher, and Hal Daum´e III. Aneural network for factoid questionsanswering over paragraphs. In Proceedings of EMNLP,2014.