International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 03 | Mar 2019 www.irjet.net p-ISSN: 2395-0072

Factoid Question and Answering System

Monika. GB[1], Nivetha. S[2], Mrs. K. Rejini[3]

[1,2]Student, Department of Computer Science, Anand Institute of Higher Technology, Kazhipattur, India
[3]Assistant Professor, Department of Computer Science, Anand Institute of Higher Technology, Kazhipattur, India

---

Abstract - *Natural language processing (NLP) is a form of artificial intelligence (AI) which is used to process the natural language data such as text, image, video, and audio. NLP act as a tool for computer to understand and analyze the real-time data in human language. The applications of NLP are machine translation, information extraction, text summarization, and question answering. Question answering (QA) is a well-researched problem in NLP. QA system is similar to an information retrieval system. In the QA system, the user will state a query to the system then the query will be processed by NLP methods to retrieve the answer. Neural network plays a major role in training the QA system. The neural network is a model consists of a series of an algorithm to determine the relationship among the dataset by mimic the working of the human brain. Tensorflow (TF) is one of the frameworks to train the neural network in an efficient way. It automatically calculates the gradient by expressing numerical computations as a graph. TF is trained with the large dataset to find the similarity between question and answer. Our system answers the factoid questions over paragraphs using neural networks along with Tensorflow framework. In order to justify the retrieved answer reasoning is used. The reasoning is the process of analyzing data in a logical way to make decisions. In QA system reasoning plays an important role in extracting the answers with better accuracy.*

***Key Words*: Factoid, Natural Language processing, Tensorflow, Neural Networks, QA system.**

1. INTRODUCTION

Natural language is a largely used tool for thinking and communicating our ideas to others. Various people from various regions will speak various languages. So the communication between two different language peoples become tough without a translator. In this technological era computer has the capability of performing complex tasks easily. Such a complex task is natural language processing (NLP), which is a tool to process natural language. Natural language processing is a field at the intersection of computer science, artificial intelligence (AI) and linguistics. It is the application of computational techniques to synthesis and analysis the natural language data such as text, speech, image, etc. The major goal of NLP is to make computers to understand the natural language in order to perform tasks like question answering, robotics, language translation and knowledge representation. Question answering is a computer science discipline within the fields of information retrieval and natural language processing, which leads to building the systems that automatically answer the questions posed by the humans in natural language. Question answering system will allow user to ask question in natural language and retrieves answer in a human understandable format. QA system needs more training to produce better results. Neural networks is a model which can train the QA system with large dataset for efficient model. Neural networks (NN) are computational model based on the structure and function of human brain. It is used to receiving, processing and transmitting the information in computer. The NN consists of neurons and connection between neurons. The neurons and connections are represented as nodes and edges respectively.

1.1 Objective

The main objective of our project is to design, implement and to evaluate QA system. This system specifically deals with text based queries, creating and evaluating new techniques. The system will analyze the text based queries and provides perfect solution.

1.2 Scope

The scope of the project is to perform machine translation, Information extraction, Text summarization and Extracting the answers with better accuracy.

2. EXISTING SYSTEM

START is abbreviated as Syntactic Analysis using Reversible Transformation was developed by Boris Katz at MIT's in AI lab. Currently, the system is undergoing further deployment by Info lab group, led by Boris Katz. It was connected to the World Wide Web in 1993 and it will answer millions of question from users around the world. It is a software system designed to answer the questions posed in natural language. This mechanism handles all variety of media including text, image, audio, video, etc. START parses the incoming questions matches the query created from the parse tree against its knowledge base and present the application information segment to the user. It displays the answer along with the evidence where it fetches the answer. "Natural language annotation" is the key technique used by START that helps to connect information seekers to information sources. The technique employs natural language sentences and phrases – annotations – as descriptions of content that are associated with information segments at various granularities. An information segment is

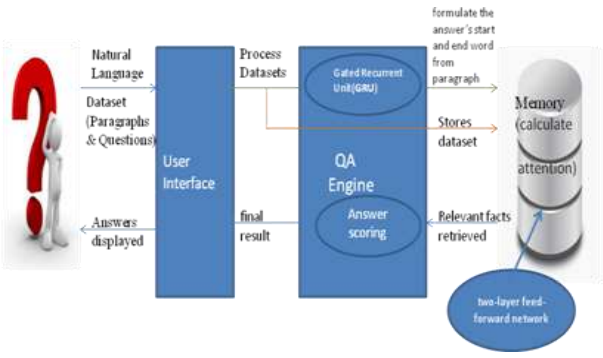
retrieved when its annotation matches an input question. The natural language processing component consists of two modules which share the same grammar. In the understanding module, the grammar used to analyses the English text and create a knowledge base to encode the text information. In generating module the answer relevant to the input question is generated from the knowledge base in user understandable format using the same grammar. These two modules use multimedia information access by putting the power in sentence level natural language processing.

3. PROPOSED SYSTEM

The proposed system is used to answer the questions in a closed domain. The user can provide a document and question as input and QA engine process the paragraph and extract the facts relevant to the question by analyzing the input paragraph and display the answer. The answer extraction and training become a time-consuming process by using Tensorflow framework. The Tensorflow (TF) is a deep learning framework which computes the gradient function automatically. It computes the complex computation easily. TF represents the complex computations as a graph like structure. The reasoning is the process of making a logical decision by analyzing the large dataset. In the field of information technology, reasoning systems are reserved systems for taking complex decisions for problems. The application of reasoning systems is robotics, natural language processing, intrusion detection, predictive analytics, and complex event processing.

4. SYSTEM ARCHITECTURE

The diagram represents the architecture diagram of our project. It also provides the overall representation of working of the project.



4.1 DATASET PROCESSING

The input module is the first module of our system where a dynamic memory network uses to come up with its answer. The input is a dataset (text document) which consist of the context (paragraph) and a question relevant to that. It can carry more than one paragraph and question. We classify the question based on the training dataset using Support Vector Machine (SVM).It is typically used for the classification task. SVM algorithm finds a hyper plane that optimally divided the classes. It is best used with a non-linear solver.

4.2 TRAINING QA ENGINE

The module consist of a simple pass over the gated recurrent unit (GRU) to gather pieces of evidence. Each piece of evidence, or *fact*, corresponds to a single sentence in the context, and is represented by the output at that time step. This requires a bit of non-Tensorflow pre-processing so we can gather the locations of the ends of sentences and pass that into Tensorflow for use in later modules. Gather_nd is a useful tool to select the corresponding sentence by tensor transformations such as joining and joining.

4.3 DATASTORAGE AND RETRIVAL

We calculate attention in this model by constructing similarity measures between each fact, our current memory, and the original question. We pass the results through a two-layer feed-forward network to get an **attention** constant for each fact. We then modify the memory by doing a weighted pass with a GRU over the input facts. To calculate the closest word for question we create a "score" for each word, which indicates the final result distance from the word is called **answer scoring**.

4.4 DATASTORAGE AND RETRIVAL

The final module is the answer module, which regresses from the question and episodic memory modules' outputs using a fully connected layer to a "final result" word vector, and the word in the context that is closest in distance to that result is our final output (to guarantee the result is an actual word). To calculate the closest word for the question we create a "score" for each word, which indicates the final result distance from the word is called answer scoring.

4.5 MANIFESTING ANSWER

Relevant facts are retrieved from previous module and framed as a accurate answer that will be presented to the user. This module regresses from the previous module's outputs using a fully connected layer to a "final result" word vector. The word in the context that is closest in distance to that result is our final output.

5. ALGORITHM

5.1 SUPPORT VENDOR MACHINE(SVM)

An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. In addition to performing linear classification, SVMs can efficiently perform a non-linear classification, implicitly mapping their inputs into high-dimensional feature spaces. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM

training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier.

6. CONCLUSION

The proposed system would be a stepping stone into an intelligent query handling program .In next phases it could not just respond but self-learn to improve itself thereby increasing the quality of service , reducing human load, increase in productivity and increasing number of satisfied Students .

REFERENCES

C. Manning, H. Schütze, "Foundations of Statistical Natural Language Processing", MIT Press, Cambridge, 1999.

[1] Dipanjan Das, Desai Chen, Andre F. T. Martins, Nathan Schneider, and Noah A Smith, "Framesemantic parsing". Computational Linguistics, vol. 40, no. 1, pp. 9–56, December

[2] Karl Moritz Hermann, Dipanjan Das, Jason Weston, and KuzmanGanchev. "Semantic, Frame identification with distributed word representations," in Proceedings of ACL, pp.1448-1458, June 22-27, 2014.

[3] H. Garis, C. Shuo, B. Goertzel, L. Ruiting, "A world survey of artificial brain projects, Part I, Largescale brain simulations", Neurocomputing, vol. 74, no. 1–3, pp. 3–29, August 2010..

B. Goertzel, R. Lian, I. Arel, H. Garis, S. Chen, "A world survey of artificial brain projects, Part II, biologically inspired cognitive architectures", Neurocomputing, vol. 74, no. 1–3, pp. 30–49, August 2010.

[4] J.Hummel, K. Holyoak, "A symbolic-connectionist theory of relational inference and generalization", Psychological Review, vol. 110, no. 2, pp. 220–264, April 2003.

M. Saito, M. Hagiwara, "Natural language processing neural network for analogical inference", International Joint Conference on Neural Networks (IJCNN), IEEE, pp. 2886–2892, 18-23 July 2010.

[5] YoshuaBengio, R´ejeanDucharme, Pascal Vincent and Christian Jauvin,"A neural probabilistic language model," JMLR, vol. 3, pp.1137–1155 March 2003.