“I am Kalam”: Analyzing and Generating Kalam’s Answer Patterns

Anannya Uberoi, Sarthika Dhawan, Shreyash Arya
Indraprastha Institute of Information Technology, Delhi
anannya15014@iiitd.ac.in,
sarthika15170@iiitd.ac.in,
shreyash15097@iiitd.ac.in

Abstract

Dr. APJ Abdul Kalam was one of the most vocal and prolific presidents of India, who served as a scientist at Indian Space Research Organization (ISRO). He was popularly called the “Peoples President since he would not only talk to journalists and interviewers but also speak abundantly to students and teachers. His numerous interactions have been recorded in various books such as “Agni Ki Udaan, “Ignited Minds, “India 2020 and so on. Some books are authored by him, whereas others are records of his interviews. Though these books are available to the general public, there is no way to gain a fundamental insight into Kalams views and philosophies. Also, it is not possible to skim through the entire corpus by human readers in order to know how Kalam would answer a particular question or address a given issue. Even if someone has read the entire set of books by Kalam, it is very demanding on their end to extrapolate the information acquired and summarize his views on a particular topic queried. The major problem we are trying to address is to bring the work of an author closer to the general masses through question answering and easy search based on topics and subjects discussed by the author.

1 Introduction

Question answering systems have been a central part of core natural language processing and information retrieval research, and have progressed through the years. There are various ways of building Q&A systems, namely (i) linguistic approaches, wherein the knowledge information is gathered through ontologies or knowledge graphs, and a thorough pre-processing of the query, a template is generated to be matched with an existing answer in the database; (ii) statistic approaches, which employ maximum entropy model and Bayesian classifiers; and (iii) AI based models, which generate answers based on semantics learnt from a training corpus(Sasikumar and L, 2014).

Our problem statement differs from popular chatbots systems in the following ways:

- “I am Kalam” must retrieve answers in the same style of talking as Kalam.
- The system is not a conventional chatbot where there is a flow of conversation. Rather, the data comprises of questions and answers which may be independent of each other.

Recurrent Neural Networks are extensively used to build chatbots with personality. The idea is to train the network on a particular author’s texts and then generate answers that emulate the author’s style. Our work is different from traditional personality based chatbots in the way that we aim to not only retain the author’s style, but also his or her views and motivations in answering a particular query. In addition, the long short term memory (LSTM) layer guarantees better control over the memory mechanism of the network. Sequence 2 Sequence models, with an encoder module encapsulating the query into a finite-dimensional representation, and the decoder module generating a fixed length sentence that best answers the question will be used for the same. The autoencoder approach, hence, seems to be a very promising alternative to the traditional approach.

The research would help in understanding the works of influential personalities better, thereby making them more accessible to the general public. It will save a lot of query time by reducing voluminous information about a query into short and brief answers allowing quick and meaningful access to motivational speeches. It will encourage literary research and research on famous public figures. Results can be optimized if the query dataset is refined. We will be able to get a better insight into the views by influential personalities on a specific topic.

2 Related Work

Majority of work in this field is developed after the intro- duction of sequence to sequence model which paraphrases the dialogues from utterances to response. But the systems developed using this model lacked any personal personality or persona. There is very limited research related to the chatbots having a exact persona of a person. A recent paper published in 2017 tried to tackle this problem using the sequence to sequence models which consists on RNNs. This uses encoder and decoders to map the utterances and responses which is further experimented using the Glove pretrained word vector embedding. Human judgment is used for evaluating the model which gave 50% accuracy i.e. 50% of the human judge believed that the sentence were generated by the real personalities and not the chatbot.

3 Dataset

We have prepared a dataset, "Koffee With Kalam", encompassing Kalam's ideas through tagged quotes on Goodreads, and his interview QnA's published on the web. We plan to extend our dataset by further including books, news articles and audio interviews in the future. Such data can be found freely on the web and most of his work is free for students to peruse and learn from his experiences and dialogue. The official website and Goodreads are good starters to collect his interviews and quotes. Following the initial data extraction, data collected has been pre-processed to create a corpus topic-wise as well as question-wise. Paragraphs, phrases and quotes have been tagged according to their relevance to certain spheres of his ideologies, such as economics, space science, politics, education and so on. These can be fine-grained using optimal keyword labels which can be decided by scanning the corpus itself. The data contains a total of 505 questions and answers.

3.1 Protocol

The dataset follows a 70% train, 15% validation and 15% test protocol. Results are reported on the test set.

4 Algorithm

4.1 Pre-Processing

Questions and answers are pre-processed wherein only alphanumeric characters are white listed for the two. This is done so that no junk characters, punctuation marks and numbers proceed through the neural network, so as to keep the outputs clean. Since our corpus can have extremely lengthy questions and answers, the data is filtered to a set maximum length. Next, the questions and answers are tokenized into sequences of words. Each word is mapped to an index. Words with relative probability less than 0.2 are filtered out as UNK (unknown words). The sentences are padded to a uniform maximum length of question and maximum length of answers, to make the lengths of all questions and answers equal. The length limits, word to index mappings, index to word mappings and UNK words are stored in a metadata file.

4.2 Sequence to Sequence Model

The Seq2Seq model can be seen as a combination of two sub-models. The first sub-model is called the encoder, and the second sub-model is called the decoder. The encoder takes in raw input text data (the question), and returns a representation of this input data. The returned representation, which captures the context or the semantic summary, is passed to the decoder, which reconstructs the input. In our case, the decoder returns a completely different reconstruction, which is the answer. The output is returned one word at a time, looking at the previous word and the surrounding context, for all the time steps.

In Seq2Seq models, padding is used to handle the variable-length sequence problems. Additionally, mask is multiplied by the calculated loss (a vector), so that the padding does not affect the loss. We use a cross entropy loss function, and Adam Optimizer with learning rate 0.01.

4.3 Implementation Details

The neural network is trained and validated over the "Koffee With Kalam" datasets. Embeddings of fixed dimension 1024 are created. The final data

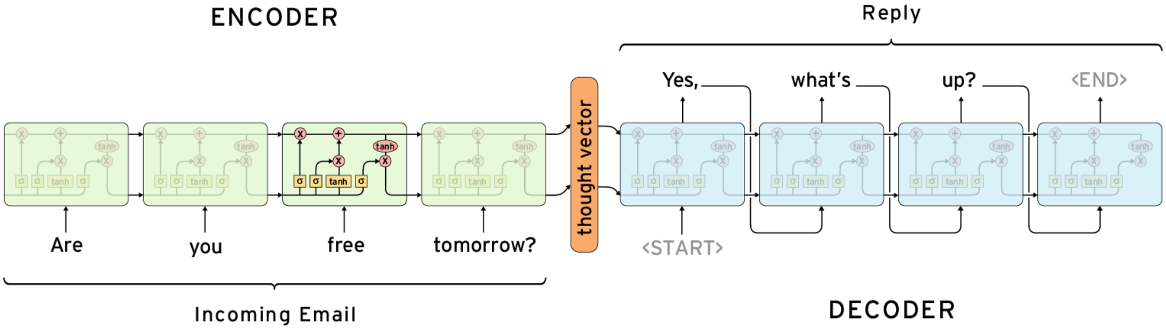

Figure 1: Sequence to Sequence Model. Source: Wiktionary.

passed to the Sequence to Sequence model is of the following format.

```
   input_seqs : ['how', 'are',
'you', '<PAD_ID'>]

   decode_seqs : ['<START_ID>',
'I', 'am', 'fine', '<PAD_ID'>]

   target_seqs : ['I', 'am',
'fine', '<END_ID>', '<PAD_ID'>]

   target_mask : [1, 1, 1, 1, 0]
```

4.4 Information Retrieval and Information Extraction Approach

For extracting the similar information from the dataset based on query question, sentence vectorization is used. The query sentence is converted into vector using the sent2vec (Matteo Pagliardini and 2, 2017) which embeds the semantic features of the sentences. This is an extension to word2vec considers words in the vector space and cluster the words nearby having similar semantic meaning. The questions in the dataset are also converted into vectors following the same approach and then the query question vector is compared with all question vectors in the dataset. Top 3 matched questions from the dataset (based on the cosine similarity between the vectors) are considered for further information extraction. The answers of top 3 matched questions is fed into extractive summarizer.

For extractive summarization, Gensim's inbuilt summarization module is used which implements TextRank, an unsupervised approach based on weighted graphs(Mihalcea and Tarau, 2004). The summarizer first pre-process the data by removing the stop words and then stemming the remaining words. It then creates a complete graph(every node is connected to every other node) with sentences as the nodes and weights on each edge based on the similarity score. TextRank algorithm then scores the edges and the edges with the top score is considered. In the original TextRank algorithm, the score is given on the basis of what percentage words appear in both sentences. Although, the summarizer implements a modified Okapi BM25 function stated in (Federico Barrios, 2016). Average answer length from the dataset (i.e. 50) is considered as the word_count parameter which defines the length of the summary.

5 Results

We tried to evaluate our model using the BLEU and ROGUE score. BLEU measures precision: how much the words (and/or n-grams) in the machine generated summaries appeared in the human reference summaries. ROUGE measures recall: how much the words (and/or n-grams) in the human reference summaries appeared in the machine generated summaries. Naturally - these results are complementing, as is often the case in precision vs recall. If you have many words from the system results appearing in the human references you will have high BLEU, and if you have many words from the human references appearing in the system results you will have high ROUGE. ROGUE score for both the approaches is coming more than BLEU score which shows that the retrieved answers are nearer to the ground truth available,i.e., number of false positives is more than number of false negatives.

Table 1: ROGUE-2 and BLEU-2 scores

Method	ROGUE-2	BLEU-2
IR-IE	0.48	0.42
Seq2Seq	0.37	0.35

5.1 Sample Results: Seq2Seq

Enter Query: what gives you happiness?
> discussing with children and understanding their smile on the faces of one billion people

Enter Query: who is your inspiration?
> prof vikram should internationally competitive products

Enter Query: what do you think about india's youth?
> i listen to and both perseverance dreams smile for the one success

Enter Query: how much should student study?
> not is first towards will favour seeker if are the a hardwork there decade of a for help

Enter Query: how do you manage to allow for india's development in space and technology?
> wherever i have been i have worked hard and sweated i am trying to market the idea of the developed india by 2020

Enter Query: how is the government contributing to education of indian youth?
> hardwork comes first fortune will favour you if you are persistent in hardwork there is a famous saying god help those who help themselves

5.1.1 Sample Results: Information Retrieval and Information Extraction Approach

Enter Query: how is the government contributing to education of indian youth?
> One girl got up and said ”I want to become a political leader.” Everybody laughed, you know. She said, ”to remove corruption.” That is, you should not be frightened and run away from the problem. (Laughs).When I go there, most of them want to join government jobs.

Enter Query: what are you views on space programs?
>We are doing very well in space technology and are capable of the most things. ISRO is doing a good job and we can do almost anything that other space going countries are capable of doing, including re-usable spacecraft. Space labs are laboratories in space aboard the space stations or space vehicle in which experiments in zero gravity are conducted.

Enter Query: how to eradicate child labour?
>It has been found in many states that the Mid Day Meal scheme in schools has helped retain children in class. Definitely we are all working to see that child labour is eliminated in a decade. There are three persons who can help in eradicating corruption in the society.

Enter Query: education and poverty?
>Modern education system should also provide value based education to create enlightened citizens. It is possible by combination of education with value system, religion transforming into spiritual force, economic development in an integrated way will lead to national development. We will definitely get enlightened citizens in a nation of billion people.

6 Conclusion

From our analysis, we are able to successfully produce results which approximate Kalam's writing style. However, there is scope of improvement in understanding the semantics of the question. This remains to be a challenge because of the verbose nature of the questions and ambiguity of the answers. Since Kalam's quotes have been quite wordy, it is difficult to capture the central meaning of his quotes through simple natural language models.

7 Future Work

More robust learning methods in this domain, such as convolutional neural networks and reinforcement learning remain to be explored. Moreover, the dataset is limited, because of which training on deep neural networks has not been considerable. Dataset expansion would help improve performance.

References

Federico Lpez Luis Argerich Rosa Wachenchauzer Federico Barrios. 2016. *Variations of the Similarity Function of TextRank for Automated Summarization.*

Prakhar Gupta Matteo Pagliardini and Martin Jaggi 2. 2017. *Unsupervised Learning of Sentence Embeddings using Compositional n-Gram Features.*

Rada Mihalcea and Paul Tarau. 2004. *TextRank: Bringing Order into Text.*

Unmesh Sasikumar and Sindhu L. 2014. *A Survey of Natural Language Question Answering System*, volume 108. Prentice-Hall.