“I am Kalam” - Reliving Kalam’s Words

Anannya Uberoi, Shreyash Arya, Sarthika Dhawan

Abstract— **Analyzing answer pattern of APJ Abdul Kalam and responding to a query following his answering pattern.**

I. INTRODUCTION

A. Challenge (What is the challenge you are trying to address?)

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.

The major challenge we are trying to address here is that though these books are available to the general public, there is no way to gain a fundamental insight into APJ Abdul 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.

B. Solution (How does the innovation work? What are its components?)

The innovation works primarily on Machine Learning and Natural Language Processing. We would be preparing a dataset of Kalams ideas through various material collected from the articles, blog, books, videos, audios etc. 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 will be pre-processed to create a corpus topic-wise as well as question-wise. Paragraphs, phrases and quotes will be 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.

Information Retrieval techniques will be used to match a user query, such as Kalams views on the education system in India or What would Kalam say to decide between science and commerce streams?, with a relevant document in the corpus. The word Kalam can later be omitted once the development of a Kalam-based question answering and summarizing system is complete. Summarization could help users gain access to succinct, to the point answers which could potentially help their study of the author. The query would initially single out a set of documents which relate to the query. The matching can be achieved using similarity metrics. This can be achieved using Natural Language Processing techniques such as abstractive summarization and paraphrasing. Absxtractive summarization on the retrieved text will be reported as the answer to the query.

C. Impact (What are the potential benefits of your innovation?)

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.

II. CURRENT PROGRESS

A. Dataset

The dataset for the quotes (from www.goodreads.com), interview question-answers and books has been collected from various open sources using the BeautifulSoup web scrapper. The data is stored in dictionary format for the easy retrieval. Data processing has been kept minimal including removal of HTML tags and lower casing to preserve the original speech structure.

For the current progress, we are using the quotes and interview question-answers as the books parsing is a tricky job. The quotes and interview questions have been tagged using Rake-NLTK for tags similarity matching and questions have been converted to word vectors for sentence similarity matching using Word Mover’s Distance (WMD).

B. Methodology

The user inputs a query or a keyword. The query input is passed through the tagging function which gives the possible tags that the query contains using the Rake-NLTK.

1) Jaccard Similarity of tags: Jaccard similarity is used to find the maximum number of intersecting tags between the query and the dataset. The Jaccard similarity is found using the below formula which gives score between 0 and 1:

$$JaccardSimilarity = (A \cap B)/(A \cup B) \tag{1}$$

2) Sentence Similarity: Sentence similarity is calculated using Gensim's package Word Mover's Distance (WMD) trained on Stanford Glove word vectors. Word Mover's Distance tries to find 'the dissimilarity score between two text documents as the minimum amount of distance that the embedded words of one document need to "travel" to reach the embedded words of another document.' The WMD score is normalized (to give score between 0 and 1) by dividing the score with the unique number of tokens in both query and dataset after removing the stopwords.

Quotes and answers for the questions with the highest score is returned. Also, combined normalized score of Jaccard and Sentence similarity is used to give the final score.

C. 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 B, and if you have many words from the human references appearing in the system results you will have high ROUGE.

We created a test set by manually modifying the questions and doing 80-20 train-test split. For the BLEU and ROUGE score, we were able to achieve 90% accurate results but there can be possibility that the changes performed manually in the test dataset is not that accurate and hence, over fitting is possible.

III. RELATED WORK

Majority of work in this field is developed after the introduction 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 chat-bots having a exact persona of a person. A recent paper published in 2017[1] 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 pre-trained 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.

IV. FUTURE WORK

Incorporating sequence to sequence model to improve overall accuracy and increasing the dataset by incorporating the books dataset.

REFERENCES

[1] Nguyen et al. ”A Neural Chatbot with Personality.”

[2] Li, Jiwei, et al. ”A persona-based neural conversation model.”arXiv preprint arXiv:1603.06155 (2016).

[3] Lin, Chin-Yew. ”ROUGE: A package for automatic evaluation of summaries.” Text summarization branches out: Proceedings of the ACL-04 workshop. Vol. 8. 2004.

[4] Papineni, Kishore, et al. ”BLEU: a method for automatic evaluation of machine translation.” Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002.