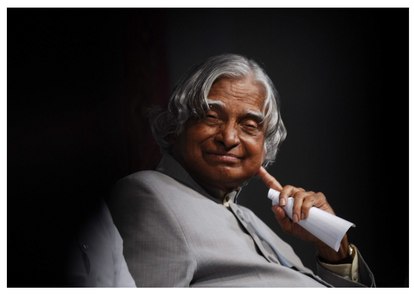
"I am Kalam" - Analyzing and Generating Kalam's Answer Patterns

Anannya Uberoi (2015014), Shreyash Arya (2015097), Sarthika Dhawan (2015170)
Advisor: Dr. Tanmoy Chakraborty
Indraprastha Institute Of Information Technology, Delhi, India

Motivation



- **Dr. APJ Abdul Kalam** - one of the most vocal and prolific Presidents of India, scientist at Indian Space Research Organization (ISRO).
- Popularly called the **People's' President** - would not only talk to journalists and interviewers, but also abundantly to students and teachers.
- *How to bring Kalam's works closer to the general public?*

Problem Statement

- Though Kalam's books are available, no way to gain a fundamental insight into his views and philosophies.
- Not feasible 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.
- Very demanding on their end to extrapolate the information acquired and summarize his views on a particular topic queried.

Goal: *To bring the work of an author closer to the general masses through question answering based on topics and subjects discussed by the author.*

Background

Differs from popular chatbots systems in two ways:

- "I am Kalam" must retrieve answers in the same style of talking as Kalam.
- No flow of conversation. Rather, questions and answers which may be independent of each other.

Data Collection

- **"Koffee With Kalam"** - Kalam's ideas through tagged quotes on Goodreads, and his interview QnA's published on the web.
- Preprocessed to create corpus:
  - Topic-wise
  - Question-wise
- Paragraphs, phrases and quotes tagged according to relevance to economics, space science, politics, education and so on.
- Total set of 505 questions and answers.

Proposed Pipeline

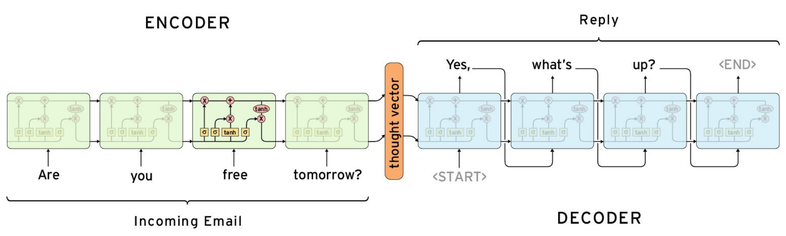
Pre Processing
- White-listing of alphanumeric characters, filtering of data to a set maximum length, tokenization of Q&A, word-index mapping.

User Query
- User enters query - could be either:
  - Question
  - Topic

Seq2Seq - Encoder
- Encoder returns representation of query.
- Captures the context or the semantic summary.

Seq2Seq - Decoder
- Decoder reconstructs the encoder output, which is the answer.

Implementation



- Padding is used to handle variable-length sequence problems.
- Mask is multiplied by the calculated loss so that the padding does not affect the loss. We use a cross entropy loss function, and
- Adam Optimizer with learning rate 0.01.
- Embeddings of fixed dimension 1024 are created.

Data passed to Seq2Seq Model:

```
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]
```

Information Retrieval Approach

- Query sentence converted to sent2vec.
- Questions in the database also converted to sent2vec.
- Query question vector compared with the questions in the database.
- Top-3 matched queries, based on cosine similarity, selected.
- Answers of top-3 matched queries fed to the extractive summarizer.
- Extractive summary returned.

Results

- BLEU-2 and Rogue-2 scores reported.
- 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.

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

Seq2Seq Samples

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

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

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

IR Approach Samples

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 reusable spacecraft. Space labs are laboratories in space aboard the space stations or space vehicle in which experiments in zero gravity are conducted.

Conclusion

- Successfully produce results which approximate Kalam's writing style.
- Scope of improvement in understanding the semantics of the question.
- Since Kalam's quotes have been quite wordy, it is difficult to capture the central meaning of his quotes through simple natural language models.

Future Work

- Expand the dataset.
- Try bi-directional Seq2Seq models.
- Try CNNs for contextual features.

Contact

Shreyash Arya
IIIT-Delhi
Email: shreyash15097@iiitd.ac.in

Report available at: https://www.overleaf.com/read/xgjhjyqymvpr

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.