Extractive Summarization

7th October 2018

NIKHIL PINNAPARAJU (20161118) &
RAMKISHORE S (20161092)

OVERVIEW and UNDERSTANDING

Automatic summarization is the process of shortening a text document with software, in order to create a summary with the major points of the original document.

The two approaches to automatic summarization, Extractive Summarization, and Abstractive Summarization.

Extractive methods work by selecting a subset of existing words, phrases, or sentences in the original text to form the summary, whereas, abstractive methods build an internal semantic representation and then use natural language generation techniques to create a summary that is closer to what a human might express. Such a summary might include verbal innovations.

The aim of this project is to be able to build complex systems that provide good results for the task of automatic summarization.

OUR RESEARCH

With respect to Extractive summarization, we have a few papers and have understood the basics w.r.t the field. The problem of extractive summarization seems to be tackled in two ways currently. One is through learning algorithms and the Second is through Graph-Based methods. With this in mind, we have selected our baselines.

BASELINES

We aim to implement two baselines for the task. One following a learning approach and the other following a graph-based approach.

1. Linear Regression - Attempt to identify the weights for each feature. (We have already thought of what all features we could use to represent the text and how to transform the text to feature vectors with normalization).
2. Naive Bayes - Implementing a Bayesian model that, when given a text unit can identify whether it would be a Summary Unit (In Summary) or a Non-Summary unit (not in summary), thereby modeling it as a classification problem.

FURTHUR GOALS

1. TextRank - Inspired by Pagerank algorithm used by Google Search to rank websites in their search engine results. TextRank is a graph-based ranking algorithm for NLP. For keyphrase extraction, it builds a graph using some set of text units as vertices and then runs the algorithm on a graph. This method allows us to select sentences or phrases for the summary so as in incorporate diversity as well as a ranking system for how likely it is for the text unit to be a part of the summary.
2. Neural Networks - We would also like to explore neural approaches to the task of summarization and see how well they would perform compared to our other implementations.

DATASETS

We will be using the CNN / Daily Mail dataset (Herman et al. 2015) for evaluating summarization. The dataset contains online news articles (781 tokens on average) along with corresponding multi-sentence summaries (3.75 sentences or 56 tokens on average).

MILESTONES

MILESTONE 1 - 7th October

Initial Scope Document

MiLESTONE 2 - 18th October

Completed Baselines

Project Report

The Code in Github Repository