Extractive Summarization

Team - NLP4Fun
Members - Nikhil Pinnaparaju (20161118)
Ramkishore S (20161092)

Introduction

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.

This project goes over Extractive Summarization and the various ways in which we approach it.

Methods

Baselines

1. Logistic Regression
2. Naive Bayes

Graph Based Models

1. TextRank

Neural Summarization Models

1. Encoder-Decoder
2. Bidirectional GRU with Pooling

Baselines

Logistic Regression

Uses a logistic function to draw a linear relation between some features like no. of verbs, no. of stop words etc. and the probability of the sentence being in the summary.

Logistic Regression

Features used	**Weights learned**	**Inference**
No. of pronouns	-0.031	Pronouns indicate dependency on other sentences
No. of verbs	0.035	Verbs indicate importance of sentence. More verbs -> more events
No. of Named entities	0.229	Most important feature. Signifies both importance and Independence
length of sentence	-0.039	Sentences occuring early in the document are better
Position of sentence in document	-0.039	Sentences occuring early in the document are better
No. of stop words	-0.088	More stop words, less importance

Results

Measure	Percentage
Precision	38%
Recall	85.6%
F1 score	52.8%

Sample output

Predicted summary:

- if selected , a post from their page will be acted out during the event , which began wednesday morning at a @entity21 theater and is streaming live on @entity13 's facebook page
- " @entity11 , " happening live until 9 a.m. thursday , is taking the magical , mundane and sometimes mystifying world of @entity2 posts to the stage in a 24 - hour performance
- " actors (from improv to @entity54) , singers , musicians , poets , sculptors , puppeteers and balloon artists all have been among those taking the stage so far

Actual summary:

- " @entity11 , " happening live until 9 a.m. thursday , is taking the magical , mundane and sometimes mystifying world of @entity2 posts to the stage in a 24 - hour performance
- sponsored by online security company @entity13 , the experimental project lets @entity2 users volunteer their profiles
- if selected , a post from their page will be acted out during the event , which began wednesday morning at a @entity21 theater and is streaming live on @entity13 's facebook page

Naive Bayes

A naive bayes classifier to classify sentences based on some features like no. of verbs, length of sentence etc.

Naive Bayes

Measure	Percentage
Accuracy	51.56%

Sample output

Predicted summary:

- if selected , a post from their page will be acted out during the event , which began wednesday morning at a @entity21 theater and is streaming live on @entity13 's facebook page
- " @entity11 , " happening live until 9 a.m. thursday , is taking the magical , mundane and sometimes mystifying world of @entity2 posts to the stage in a 24 - hour performance
- sponsored by online security company @entity13 , the experimental project lets @entity2 users volunteer their profiles

Actual summary:

- " @entity11 , " happening live until 9 a.m. thursday , is taking the magical , mundane and sometimes mystifying world of @entity2 posts to the stage in a 24 - hour performance
- sponsored by online security company @entity13 , the experimental project lets @entity2 users volunteer their profiles
- if selected , a post from their page will be acted out during the event , which began wednesday morning at a @entity21 theater and is streaming live on @entity13 's facebook page

Graph Based Models

Graph Based Models

Compared to traditional supervised learning based models, Graph Based Models follow an unsupervised approach and can be applied to any body of text. This makes them very widely useful and valuable.

Graph based solutions tend to focus upon identifying the key words and key phrases and then selecting sentences from the passage based on scores obtained from previous stages.

1. Extraction of Keywords
2. Obtaining the Key Sentences
3. Generation of summary
4. Evaluation of the Model Built

TextRank

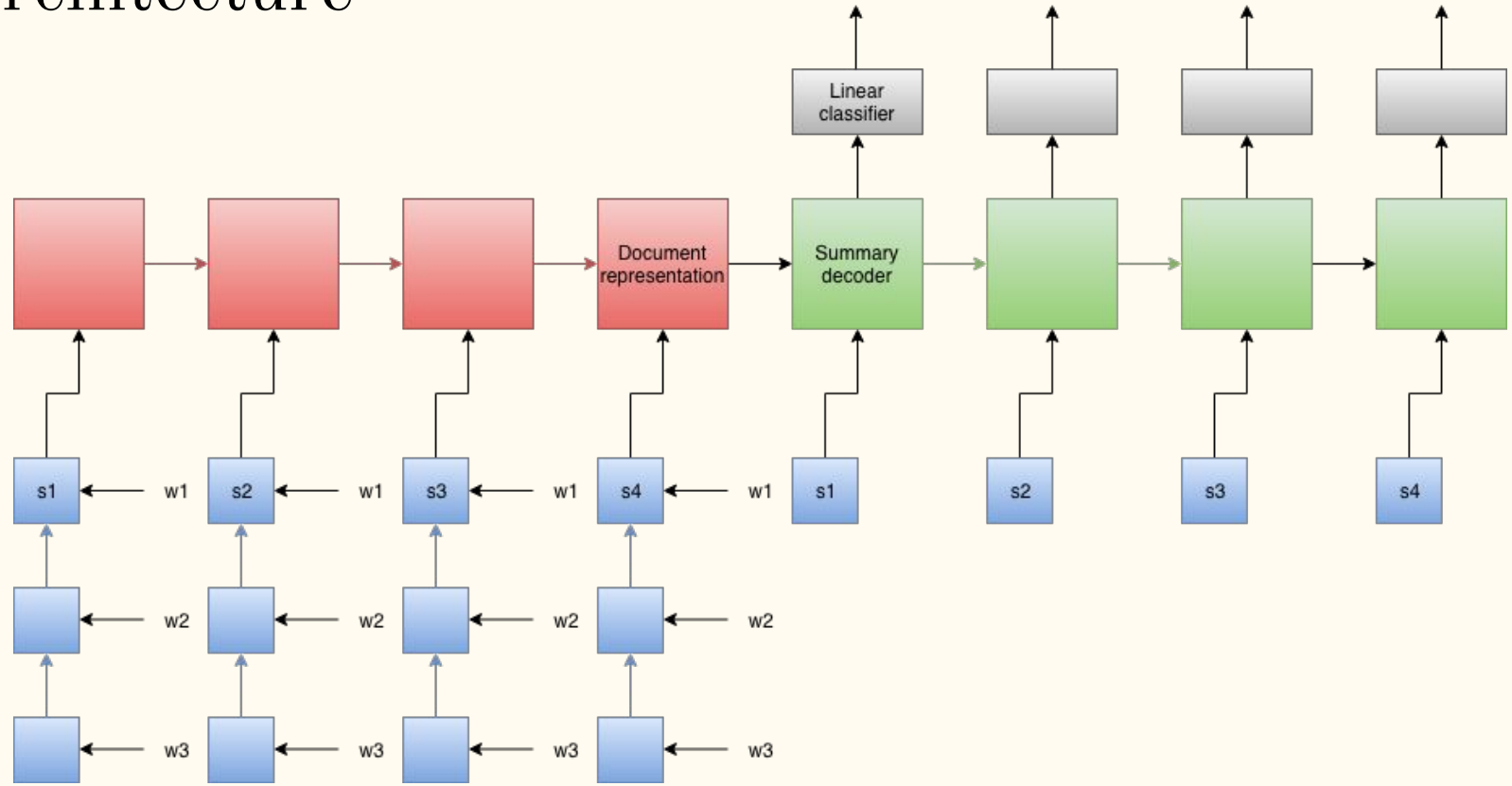
Edges are created based on word co-occurrence in this application of TextRank. Two vertices(words here) are connected by an edge if the unigrams appear within a window of size N in the original text. N is typically around 2 to 10.

> For example:- Thus, "*natural*" and "*language*" might be linked in a text about NLP. "Natural" and "processing" would also be linked because they would both appear in the same string of N words.

These edges build on the notion of "text cohesion" and the idea that words that appear near each other are likely related in a meaningful way.

Neural Architecture
Encoder-Decoder

Architecture



Results

ROUGE	PERCENTAGE	STATE OF THE ART
R1	25.26%	~30%
R2	8.7%	~10%
RL	22.4%	~25%

Sample output

Predicted summary:

- @entity1 , @entity2 amid growing scrutiny over whether a 73 - year - old volunteer deputy who killed a suspect during a sting operation was qualified to be policing the streets , a new report raises a troubling allegation
- claims that the volunteer deputy 's records had been falsified emerged " almost immediately " from multiple sources after @entity15 killed @entity23 on april 2 , reporter @entity18 said
- but the orders apparently started years ago , before @entity23 ' death , " back when (@entity15) was trying to get on as a deputy , " reporter @entity31 told @entity3 's " @entity35

Actual summary:

- @entity153 in @entity154 says @entity15 never trained with them
- " he met every requirement , and all he did was give of himself , " his attorney says @entity11 newspaper : three supervisors who refused to sign forged records on @entity15 were reassigned

THANK YOU