INTRODUCTION

Summarization can be defined as a task of producing a concise and fluent summary while preserving key information and overall meaning.

Abstractive Summarization: Abstractive methods select words based on semantic understanding, even those words did not appear in the source documents. It aims at producing important material in a new way.

Input document → understand context → semantics → create own summary.

Extractive Summarization: Extractive methods attempt to summarize articles by selecting a subset of words that retain the most important points.

Input document → sentences similarity → weight sentences → select sentences with higher rank.

SUPERVISED VS UNSUPERVISED METHODS

"A diversity based summarizer with XM -means, which is unsupervised, was **comparable to and even superior** to the supervised approach at some compression rates." The results indicate that it is possible with an unsupervised approach to model subjective judgments by humans, at least as well as a learning (decision tree) based approach.

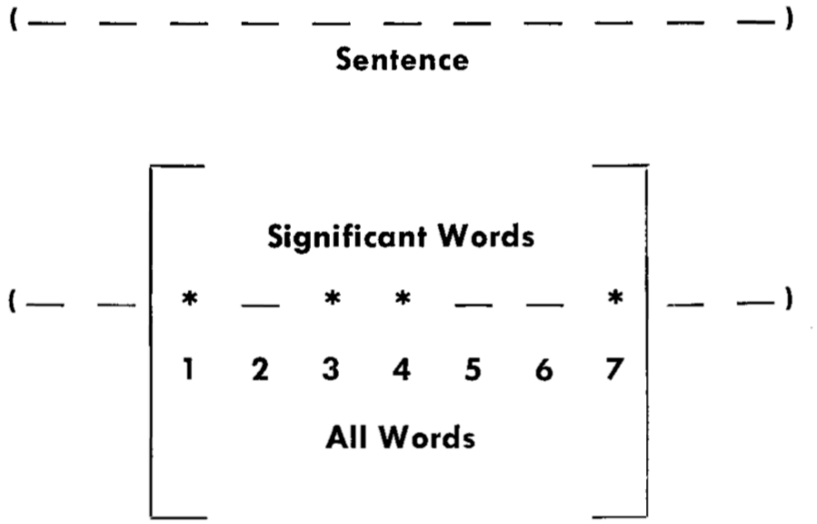
Hence, we explore unsupervised and reinforcement learning methods only. **(*https://ieeexplore.ieee.org/document/989585*)**

TIMELINE

A. Unsupervised

1950: Automatic Creation of Literature Abstracts

- One of the first text summarization techniques.
- Each sentence has a **significance factor** computed from the words contained.
- Words with **greater frequency of occurrenc**e are significant words.
- **Relative positioning** of significant words along with number of significant words determines the significance of the sentence.



Portion of sentence bracketed by and including significant words not more than four non-significant words apart. If eligible, the whole sentence is cited.

Figure 2 **Computation of significance factor.** *The square of the number of bracketed significant words (4) divided by the total number of bracketed words (7) = 2.3.*

http://courses.ischool.berkeley.edu/i256/f06/papers/luhn58.pdf

1960: New Methods in Automatic Extracting

- Apart from frequency of occurrence, this paper uses three new methods to determine the significant words.
- Presence of **cue words, words used in the title appearing in the text, and the location of sentences** are the other three parameters.

http://courses.ischool.berkeley.edu/i256/f06/papers/edmonson69.pdf

These approaches are title and location based.

1960 onwards: Shift from linguistic to statistic and graph based approaches

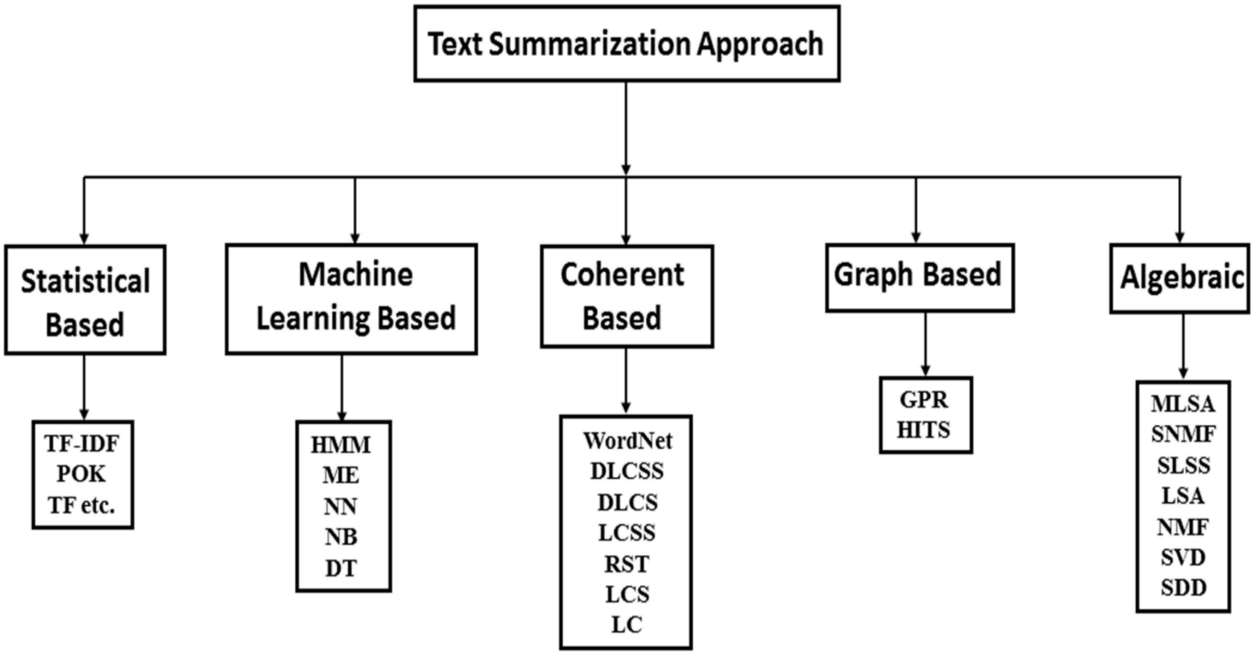


Figure 3: Classification of Text Summarization Approaches

Cue word method

- Weight is assigned to text based on its significance like positive weights "verified, significant, best, this paper" and negative weights like "hardly, impossible".
- The cue phrase method is based on the assumption that such phrases provide a

Tf - Idf method

- The term frequency - inverse document frequency is a numerical statistic which reflects how important a word is to a document. It is often used as a weighting factor in information retrieval and text mining.
- The tf-idf value increases proportionally to the number of times a word appears in the document.
- used by search engines as a central tool in scoring and ranking.

Discourse based approaches (Mann and Thompson)

- Mann and Thompson proposed Rhetorical Structure Theory (RST) in computational linguistics domain to act as a discourse structure.
- RST has two main aspects. Coherent Texts:
  - contain a few number of units, connected together by rhetorical relations.
  - Have some kind of relation between various parts of the text.

Coherence as well as **cohesion** are the two main issues.

Graph based approaches

In a graph, text elements (words or sentences) are represented by nodes and edges connect the related text elements (semantically related) together.

LexRank

- If similarity among two sentences lies above a given limit, then there is a connection between them in the graph.
- After the network is made, important sentences are selected by the system by carrying out a random walk on the graph.

Graphsum

- Add on to Lexrank

- represents correlations among multiple terms by discovering lot of new association rules.

Page Rank and Text Rank Method

TextRank algorithm is a graph based algorithm which is applied in summarization. A graph is constructed by adding a vertex for each sentence in the text. Edges between vertices are established using sentence inter-connections.

https://www.analyticsvidhya.com/blog/2018/11/introduction-text-summarization-textrank-python/
https://link.springer.com/article/10.1007/s10462-016-9475-9
https://medium.com/sciforce/towards-automatic-text-summarization-extractive-methods-e8439cd54715
https://www.researchgate.net/publication/284411761_Comparative_Study_of_Text_Summarization_Methods

B. Reinforcement

RELIS, REFRESH, DRESS, etc.

Further, reinforcement learning is also used for abstractive summarization.

SUMMARY OF PAPERS

Mallika - basic algorithms summarization

https://blog.floydhub.com/gentle-introduction-to-text-summarization-in-machine-learning/

https://towardsdatascience.com/understand-text-summarization-and-create-your-own-summarizer-in-python-b26a9f09fc70

Hetvi - graph based approaches

- “TextRank: Bringing Order into Texts”
- https://ieeexplore.ieee.org/abstract/document/4410421 (Extraction-Based Single-Document Summarization Using Random Indexing)

- https://www.sciencedirect.com/science/article/pii/S0885230814000722#bib0030 (Random Indexing and Modified Random Indexing based approach for extractive text summarization)
- click here and here and here (paper unavailable, to be released soon)

Bhumika - statistical approaches

- https://ieeexplore.ieee.org/document/7019360
- https://www.aclweb.org/anthology/X98-1025.pdf
- Multi-document summarization - https://www.aclweb.org/anthology/W00-0405.pdf

Sakshi and Sharut - RL based approaches

https://www.aclweb.org/anthology/I17-2033.pdf (Embedding features)
https://arxiv.org/pdf/1703.10931.pdf (DRESS)
https://arxiv.org/pdf/1802.08636.pdf (REFRESH)
https://arxiv.org/pdf/1907.12894.pdf (RELIS)

Niladri Sir's papers -

Fuzzy Sets - https://www.tandfonline.com/doi/full/10.1080/02564602.2018.1516521
Genetic Algorithms - https://ieeexplore.ieee.org/iel5/6389599/6407845/06407852.pdf

Random Indexing and Modified Random Indexing (covered above)
Random Indexing and Neural networks - https://pdfs.semanticscholar.org/5db3/69c80c4fad8d30f2bc2c216e6a03df4c2cc8.pdf

Datasets for testing -

https://www.kaggle.com/pariza/bbc-news-summary/data

https://www.kaggle.com/sunnysai12345/news-summary

Keep adding all the relevant links to the document itself :)

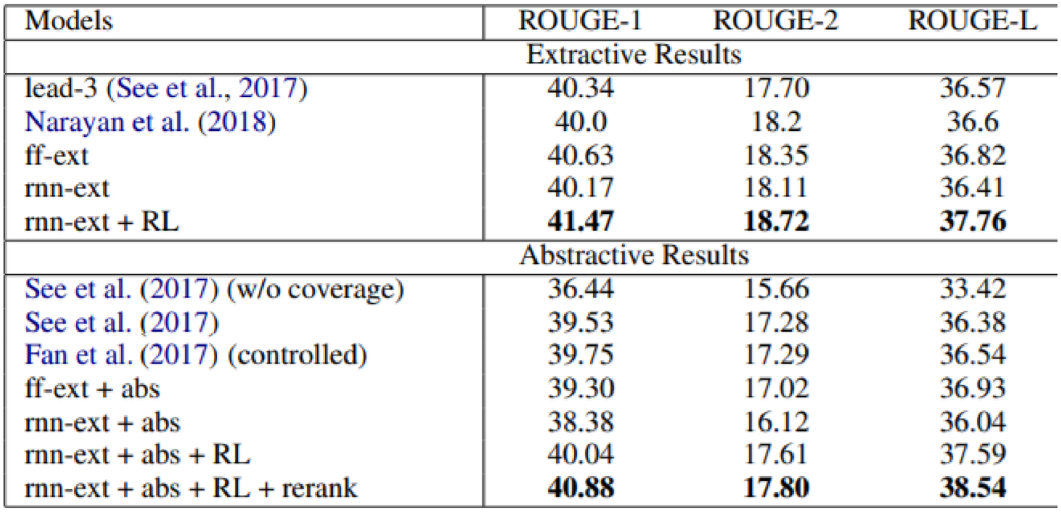
Literature Review during Internships

Fast Abstractive Summarization with Reinforce-Selected Sentence Rewriting (Chen and Bansal, 2018)

Short summary of paper:

Similar to how humans summarize long documents, their accurate and fast summarization model first selects salient sentences and then rewrites them abstractively (i.e., compresses and paraphrases) to generate a concise overall summary. Approximations to extractor are as seen in the results table, they compare feed forward neural net, and recurrent neural net extractor. The extractor sentences are extracted greedily - during training maximization of the ROUGE score with respect to corresponding ground truth sentence is used as the similarity measure. For the abstractor training, they create training pairs by taking each summary sentence and pairing it with its extracted document sentence, trained to minimize cross-entropy loss. If the extractor chooses a good sentence, after the abstractor rewrites it the ROUGE match would be high and thus the action is encouraged. If a bad sentence is chosen, though the abstractor still produces a compressed version of it, the summary would not match the ground truth and the low ROUGE score discourages this action. In the RL training phase, they add another set of trainable parameters which result in a stop action - this is to indicate how many sentences are to be extracted. Abstractive models can be more concise by performing generation from scratch, but they suffer from slow and inaccurate encoding of very long documents, since the attention model looks at all encoded words for decoding each generated summary word. By first operating at the sentence-level and then the word-level, they enable parallel decoding of the neural generative model. This results in substantially faster (10-20x) inference speed as well as 4x faster training convergence than previous long-paragraph encoder-decoder models.

Results on CNN/DM dataset:



Models	ROUGE-1	ROUGE-2	ROUGE-L
Extractive Results			
lead-3 (See et al., 2017)	40.34	17.70	36.57
Narayan et al. (2018)	40.0	18.2	36.6
ff-ext	40.63	18.35	36.82
rnn-ext	40.17	18.11	36.41
rnn-ext + RL	**41.47**	**18.72**	**37.76**
Abstractive Results			
See et al. (2017) (w/o coverage)	36.44	15.66	33.42
See et al. (2017)	39.53	17.28	36.38
Fan et al. (2017) (controlled)	39.75	17.29	36.54
ff-ext + abs	39.30	17.02	36.93
rnn-ext + abs	38.38	16.12	36.04
rnn-ext + abs + RL	40.04	17.61	37.59
rnn-ext + abs + RL + rerank	**40.88**	**17.80**	**38.54**

Fine-tune BERT for Extractive Summarization

Abstract:

BERT , a pre-trained Transformer model, has achieved ground-breaking performance on multiple NLP tasks. This paper describes BERTSUM, a simple variant of BERT, for extractive summarization. The models described in the paper perform single document summarisation. This paper tests the model on CNN/Dailymail dataset. Experimentation on two large-scale datasets and found that BERTSUM with inter-sentence Transformer layers can achieve the best performance.

Result:

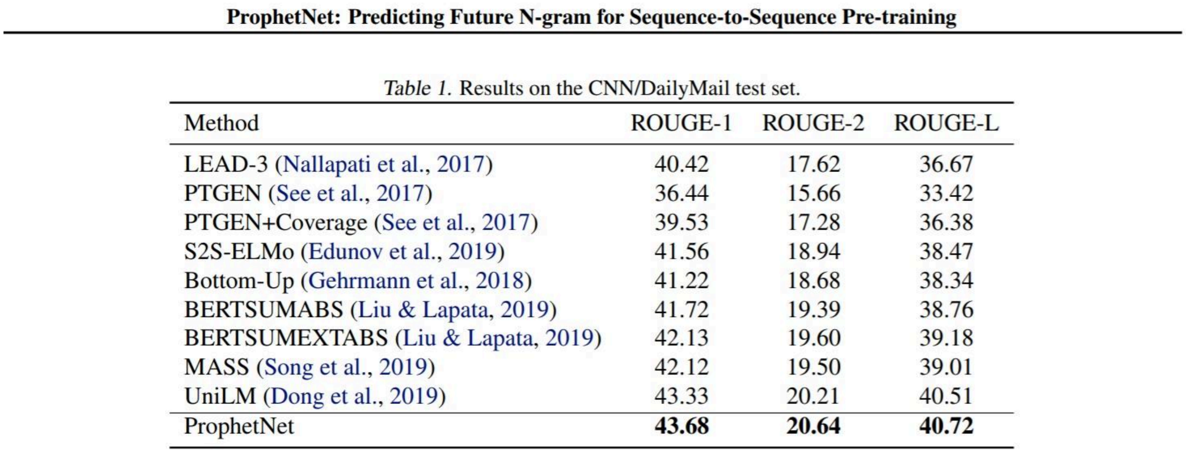
Models	ROUGE-1	ROUGE-2	ROUGE-L
Transformer Baseline	40.9	18.02	37.17
BERTSUM+Classifier	43.23	20.22	39.60
BERTSUM+Transformer	43.25	20.24	**39.63**
BERTSUM+LSTM	43.22	20.17	39.59

ProphetNet: Predicting Future N-gram for Sequence-to-Sequence Pre-training

Abstract:
In this paper, a new sequence-to-sequence pre-training model called ProphetNet is introduced which is a novel self-supervised objective named future n-gram prediction and the proposed n-stream self-attention mechanism. Instead of the optimization of one-step ahead prediction in traditional sequence-to-sequence model, the ProphetNet is optimized by n-step ahead prediction which predicts the next n tokens simultaneously based on previous context tokens at each time step. The future n-gram prediction explicitly encourages the model to plan for the future tokens and prevent overfitting on strong local correlations. Experiments on CNN/DailyMail have been done about summarization and question generation tasks. Experimental results show that

ProphetNet achieves new state-of-the-art results on all these datasets compared to the models using the same scale pre-training corpus.

Result:



ProphetNet: Predicting Future N-gram for Sequence-to-Sequence Pre-training

Table 1. Results on the CNN/DailyMail test set.

Method	ROUGE-1	ROUGE-2	ROUGE-L
LEAD-3 (Nallapati et al., 2017)	40.42	17.62	36.67
PTGEN (See et al., 2017)	36.44	15.66	33.42
PTGEN+Coverage (See et al., 2017)	39.53	17.28	36.38
S2S-ELMo (Edunov et al., 2019)	41.56	18.94	38.47
Bottom-Up (Gehrmann et al., 2018)	41.22	18.68	38.34
BERTSUMABS (Liu & Lapata, 2019)	41.72	19.39	38.76
BERTSUMEXTABS (Liu & Lapata, 2019)	42.13	19.60	39.18
MASS (Song et al., 2019)	42.12	19.50	39.01
UniLM (Dong et al., 2019)	43.33	20.21	40.51
ProphetNet	**43.68**	**20.64**	**40.72**

Neural Document Summarization by Jointly Learning to Score and Select Sentences (6th July, 2018)

Link to paper - https://arxiv.org/pdf/1807.02305.pdf

They have developed a neural network framework (NeuSum) for extractive document summarization which jointly learns to score and select sentences from the data. Initially an encoder is used to generate representation of the sentences and then these sentences are picked one by one to generate the output summary. The approach integrates the selection strategy into the scoring model, which directly predicts the relative importance of a sentence given previously selected sentences. The goal of training is to learn a scoring function f(S) which can be used to find the best summary during testing by maximising the gain after selecting each sentence. They tested it on the CNN/Daily Mail dataset and showed that the proposed framework significantly outperforms the state-of-the-art extractive summarization models.

This was similar to document summarization using MMR (the techniques of jointly scoring and selecting were the same).

Full length ROUGE-F1 evaluation (%) on CNN/Daily Mail test set -

Models	**Rouge-1**	**Rouge-2**	**Rouge-3**
LEAD3	40.24	17.70	36.45
TextRank	40.20	17.56	36.44
CRSum	40.52	18.08	36.81
PGN	39.53	17.28	36.38
NN-SE	41.13	18.59	37.40
SummaRuNNer	39.6	16.2	35.3
NeuSum	**41.59**	**19.01**	**37.98**

SummaRuNNer: A Recurrent Neural Network based Sequence Model for Extractive Summarization of Documents (14th November, 2016)

They proposed a Recurrent Neural Network(RNN) based sequence model for extractive summarization. The major distinguishing feature is that it can be trained end-to-end using abstractive summaries. Each sentence is visited sequentially in the original document order and a logistic layer makes a binary decision (taking into account previous decisions made) in terms of whether or not it should be included in the summary. This takes salience and novelty of the sentences into account with respect to the originally generated summary.

For training data sets with ground truth in the form of sentence-level binary labels for each document, representing their probability of occurrence (importance) in the summary are needed. For this they employ a method to convert abstractive summaries to extractive labels, picking sentences which maximise the rouge score with respect to the gold summaries. For the experiments, they used the CNN/DailyMail corpus originally constructed for the task of passage-based question answering, and re-purposed for the task of document summarization.

SummaRuNNer trained using extractive labels outperforms other state-of-the-art models.

Models	**Rouge-1**	**Rouge-2**	**Rouge-3**
LEAD3	21.9	7.2	11.6
LReg(500)	18.5	6.9	10.2
Cheng et al’16	22.7	8.5	12.5
SummaRuNNer	**26.2**	**10.8**	**14.4**

Pre-Internship compilation

Introduction and basic approach to text summarisation

Mallika

Definition and requirement

- Text summarization is the technique for generating a concise and precise and fluent summary of voluminous texts while focusing on the sections that convey useful information, and without losing the overall meaning.

- Automatic text summarization aims to transform lengthy documents into shortened versions which is difficult and costly to do manually.

- In modern times with our busy schedules, we prefer to read the summary of those articles before we decide to jump in for reading the entire article. Reading a summary help us to identify the interest area, gives a brief context of the story.

- Also, most of the information available to us is redundant, insignificant, and may not convey the intended meaning. Therefore, using automatic text summarizers capable of extracting useful information that leaves out inessential and insignificant data is becoming vital.

Roadmap

1. We take the input article

2. Split the article into sentences

3. Text processing (Remove stop words)

4. Tokenization (taking a text or set of text and breaking it up into its individual words)

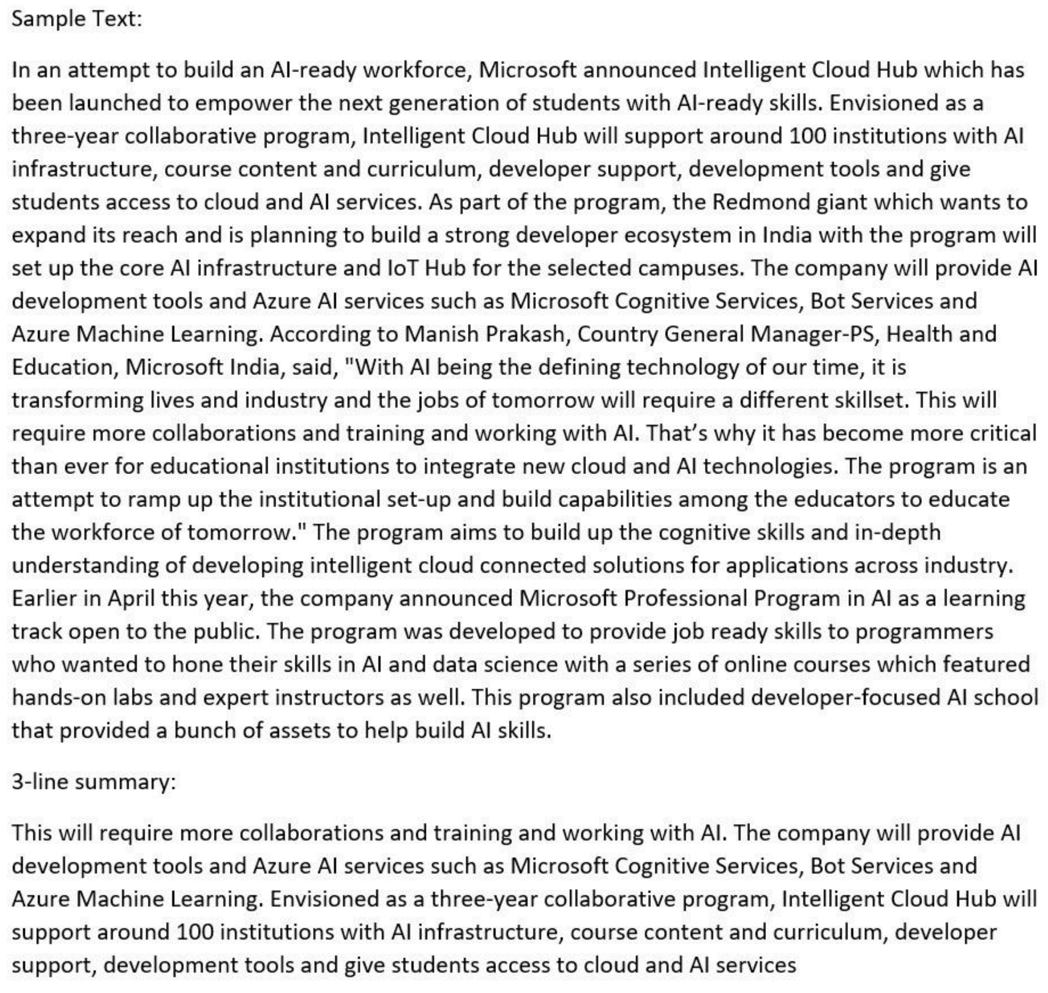
5. Calculate weighted occurrence frequency of words

6. Rank the lines accordingly

7. Pick the sentence with the most weight.

8. A summary is generated.

Sample testing:



Sample Text:

In an attempt to build an AI-ready workforce, Microsoft announced Intelligent Cloud Hub which has been launched to empower the next generation of students with AI-ready skills. Envisioned as a three-year collaborative program, Intelligent Cloud Hub will support around 100 institutions with AI infrastructure, course content and curriculum, developer support, development tools and give students access to cloud and AI services. As part of the program, the Redmond giant which wants to expand its reach and is planning to build a strong developer ecosystem in India with the program will set up the core AI infrastructure and IoT Hub for the selected campuses. The company will provide AI development tools and Azure AI services such as Microsoft Cognitive Services, Bot Services and Azure Machine Learning. According to Manish Prakash, Country General Manager-PS, Health and Education, Microsoft India, said, "With AI being the defining technology of our time, it is transforming lives and industry and the jobs of tomorrow will require a different skillset. This will require more collaborations and training and working with AI. That’s why it has become more critical than ever for educational institutions to integrate new cloud and AI technologies. The program is an attempt to ramp up the institutional set-up and build capabilities among the educators to educate the workforce of tomorrow." The program aims to build up the cognitive skills and in-depth understanding of developing intelligent cloud connected solutions for applications across industry. Earlier in April this year, the company announced Microsoft Professional Program in AI as a learning track open to the public. The program was developed to provide job ready skills to programmers who wanted to hone their skills in AI and data science with a series of online courses which featured hands-on labs and expert instructors as well. This program also included developer-focused AI school that provided a bunch of assets to help build AI skills.

3-line summary:

This will require more collaborations and training and working with AI. The company will provide AI development tools and Azure AI services such as Microsoft Cognitive Services, Bot Services and Azure Machine Learning. Envisioned as a three-year collaborative program, Intelligent Cloud Hub will support around 100 institutions with AI infrastructure, course content and curriculum, developer support, development tools and give students access to cloud and AI services

Source: sample material provided in the blogs

HETVI

Text Summarization: Graph based approaches

A. TextRank: Bringing Order into Texts (Rada Mihalcea and Paul Tarau)

- They propose a graph-based ranking model for text processing, more specifically two unsupervised methods for keyword and sentence extraction.
- **Motivation:** Graph-based ranking algorithms continue to be used very successfully in citation analysis, social networks, and the analysis of the link-structure of the internet. This is because a graph-based ranking algorithm is a way of deciding on the importance of a vertex within a graph, by taking into account *global information* recursively computed from the entire graph, rather than relying only on local vertex-specific information. Thus, we can also use graphs to extract info from the document as a whole for text oriented ranking applications.
- **Basic idea behind the algorithm:**
  - 1. Identify text units that best define the task at hand, and add them as vertices in the graph.
  - 2. Identify relations that connect such text units, and use these relations to draw edges between vertices in the graph. Edges can be directed or undirected, weighted or unweighted.
  - 3. Starting from arbitrary values assigned to each node in the graph, the computation iterates until convergence below a given threshold is achieved
  - 4. Sort vertices based on their final score. Use the values attached to each vertex for ranking/selection decisions
- They introduced a new formula for graph-based ranking that takes into account edge weights when computing the score associated with a vertex in the graph (d is a damping factor that can be set between 0 and 1, which has the role of integrating the probability of going from a given vertex to another random vertex in the graph into their model - it is usually set to 0.85 (Brin and Page 1988) & that’s what is used in this paper too)

$$WS(V_i) = (1-d) + d * \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j)$$

- **Application to sentence extraction for automatic summarization:** *identifying sequences that are more “representative” for the given text*
    - A vertex is added to the graph for each sentence in the text
    - The edge of the graph is a measure of a “similarity” relation between them, where “similarity” is measured as a function of their content overlap (“recommendation” - when one sentence links you to another)
    - This can be determined simply as the number of common tokens between the lexical representations of the two sentences, or it can be run through syntactic filters, which only count words of a certain syntactic category, e.g. all open class words, nouns and verbs, etc. Moreover, to avoid promoting long sentences, they are using a normalization factor, and divide the content overlap with the length of each sentence (similarity formula in paper)
    - Other sentence similarity measures, such as string kernels, cosine similarity, longest common subsequence, etc. can also be used instead.
    - Once the text is represented as a weighted graph, we calculate scores for each node using the proposed formula and simply extract the top ‘x’ sentences.
- **Evaluation:** Using ROUGE on a single-document summarization task, using 567 news articles provided during the Document Understanding Evaluations 2002 (DUC, 2002).
- **Novelty:**
    - TextRank relies only on the given text to derive an extractive summary, which represents a summarization model closer to what humans are doing when producing an abstract for a given document which is unlike supervised methods.
    - Goes beyond the sentence “connectivity” in a text
    - Easily adaptable:
        - It gives a ranking over all sentences in a text – which means that it can be easily adapted to extracting both very short summaries or longer more explicative summaries
        - Does not require deep linguistic knowledge, nor domain or language specific annotated corpora, which makes it highly portable to other domains, genres, or languages
        - Although the TextRank applications described in this paper rely on an algorithm derived from Google’s PageRank (Brin and Page, 1998), other graph-based ranking algorithms such as e.g. HITS (Kleinberg, 1999) or Positional Function (Herings et al., 2001) can be easily integrated into the TextRank model

Details about my implementation:
Eg. on BBC dataset:

My extractive summary- Similarity as defined by paper: (tech article 1)
1.The Kyrgyz Republic , a small , mountainous state of the former Soviet republic , is using invisible ink and ultraviolet readers in the country 's elections as part of a drive to prevent multiple voting .

2. In an effort to live up to its reputation in the 1990s as `` an island of democracy '' , the Kyrgyz President , Askar Akaev , pushed through the law requiring the use of ink during the upcoming Parliamentary and Presidential elections .
3. The use of ink is only one part of a general effort to show commitment towards more open elections - the German Embassy , the Soros Foundation and the Kyrgyz government have all contributed to purchase transparent ballot boxes .
4. The other common type of ink in elections is indelible visible ink - but as the elections in Afghanistan showed , improper use of this type of ink can cause additional problems .
5. The use of `` invisible '' ink is not without its own problems . In most elections , numerous rumors have spread about it .

My extractive summary- Cosine Similarity: (tech article 1)
1. In an effort to live up to its reputation in the 1990s as `` an island of democracy '' , the Kyrgyz President , Askar Akaev , pushed through the law requiring the use of ink during the upcoming Parliamentary and Presidential elections .
2. This type of ink has been used in many elections in the world , in countries as varied as Serbia , South Africa , Indonesia and Turkey .
3. The other common type of ink in elections is indelible visible ink - but as the elections in Afghanistan showed , improper use of this type of ink can cause additional problems .
4. The use of `` invisible '' ink is not without its own problems . In most elections , numerous rumors have spread about it .
5. However , in reality , the ink is very effective at getting under the cuticle of the thumb and difficult to wash off .

B. Extraction-Based Single-Document Summarization Using Random Indexing

In this paper, random indexing has been used to compute the semantic similarity scores of sentences and graph-based ranking algorithms have been employed to produce an extract of the given text.

Word Space: The complete vocabulary of any text (containing n words) can be represented in an n-dimensional space in which each word occupies a specific point in space and has a vector associated with it defining its meaning. This is the word space. It’s constructed in accordance with two hypotheses - Proximity, and Distribution. **To aid us in our task, we use context vectors assigned to each word (basically a sum of its environments), and a similarity measure to assess how close in meaning is one word with another.**

The problem: The dimension *n* used to define the word space corresponding to a text document is equal to the number of unique words in the document, which increases as text size increases. The other problem is data sparseness. The majority of cells in the co-occurrence matrix constructed corresponding to the document will be zero. The solution to this predicament is to reduce the high dimensionality of the vectors. This paper uses random indexing to tackle this. It’s advantages are - it is an incremental method, which means that the context vectors can be used for similarity computations even after just a few examples have been encountered, it uses fixed dimensionality so new data does not increase the dimensionality of the vectors, and uses implicit dimension reduction, since the fixed dimensionality is much lower than the number of words in the data.

RI algo:

1. Each word in the text is assigned a unique and randomly generated vector called the *index vector*. The index vectors are sparse and high dimensional and ternary (i.e. 1, −1, 0). Each word is also assigned an initially empty *context vector* which has the same dimensionality (r) as the index vector.
2. The context vectors are then accumulated by advancing through the text one word taken at a time, and adding the context's index vector to the focus word's context vector. When the entire data has been processed, the r-dimensional context vectors are effectively the sum of the words' contexts.

Semantic vectors: avg. term vector: {1\over n}\sum_{i=1}^{n}\vec{\rm x}_{i}=\vec{{\rm x}}_{mean}. Once all the context vectors have been accumulated, semantic vectors for the sentences are computed. A mean vector was calculated from the context vectors of all the words in the text. This vector was subtracted from the context vectors of the word appearing in the sentence, the resultants were summed up and averaged to compute the semantic vector of the sentence. Subtraction by mean is done to remove bias.

Graph for summarization: The graph is undirected, and completely connected. Nodes of the graph are sentences, edge weights are given by sentence similarity measure. A 1-dimensional vector was allocated to each sentence, with a length equal to the number of index words and each element referring to an index word. The value of that element was determined by multiplying the number of times the index word occurred in the sentence by its weight. An average document vector was calculated by averaging the context vectors of all its sentences. Cosine similarity of each of the sentences with the document vector was calculated. The value thus obtained was assigned to the respective node and will be called the node weight. Then, using the page rank algorithm, rank each node and create an extractive summary by considering top ‘k’ sentences.

Results

Word Embedding	**Rouge-1**	**Rouge-2**	**Rouge-L**
Random Indexing	**24.39**	**6.98**	**14.47**
Word2Vec	**20.18**	**4.94**	**13.08**

FastText	**20.24**	**5.00**	**13.18**
BERT	**12.82**	**2.59**	**10.25**
SBERt	**20.70**	**6.32**	**13.63**
GloVe	**19.39**	**4.74**	**12.58**

Automatic text summarization with statistical and linguistic features using successive thresholds (Bhumika)

“In this paper, we propose an automatic text summarization technique using both linguistic and statistical features using successive threshold for finding the summary i.e important sentences from the given input text document. Here the sentences are selected for summary based on the weight of the sentence. The weight of the sentences is calculated based on the statistical and linguistic features. Our approach assigns scores to the sentences by weighting the features like term frequency, word occurrences, and noun weight, phrases etc. In our approach, the number of sentences present in our summary would be equal to the number of paragraphs present in a text document, which can be achieved by using our successive threshold approach. “

They have introduced 2 new ideas to the general extractive summarization processes previously used -

1. Successive thresholds
2. Additional statistical and linguistic features

Statistical features - used to rank words and hence sentences

1. **Keyword features**
2. Sentences position
3. Term frequency
4. Length of the word
5. Part of speech tag

Linguistic features -

1. Proper noun features
2. Pronoun features

Proposed Method

“The Proposed Method will be processed by the following Algorithm: -

1. Count the total number of paragraphs in the given document.
2. Pre-Process the given document, segmenting the given document into sentences and then segment each sentence into words.
3. Carry out an analysis of stop words, *and* then apply stop word removal and stemming procedure.

4. Assign a Score for each sentence in a document based on features(both linguistic and statistic) as follows:
    1. Assigning a weight for each word based on its level of importance.
    2. Calculate the total weight of each word(Twt) of a sentence by applying all feature's weight Twt(word)=(TermFrequencywt(word)) / Total no of features.
    3. Calculate the total score of a sentence by summing the total weights of a word.
       Score (sentence)= (Twt() + Twt() +
       Twt()+)/(Total no of words)
5. Set the threshold value for selecting important sentences by computing the average weight of all sentences initially.
6. Select the sentences that meet the threshold value only.
7. To achieve both compression and retention:
    1. Check whether there are any sentences that are repeated by computing the similarity among all the selected sentences. Remove the repeated sentences.
    2. Count the total number of sentences n left, if n is equal to the number of paragraphs of a document, and then these are the sentences finally included in the summary.
    3. Else, Re-Compute the threshold value by averaging the score of all the sentences, which met the initial threshold value.
8. Repeat step 7 until our n value will be equal to the number of paragraphs of a given document”

“For evaluating the results of our proposed approach, we used The evaluation measures such as precision, Recall and F-measures. Here for each document, the summary generated by our algorithm as candidate summary (denoted by Scan) and the reference summary(denoted as Sref). We compare the candidate summary and reference summary and compute the precision, recall, and F value as follows: “

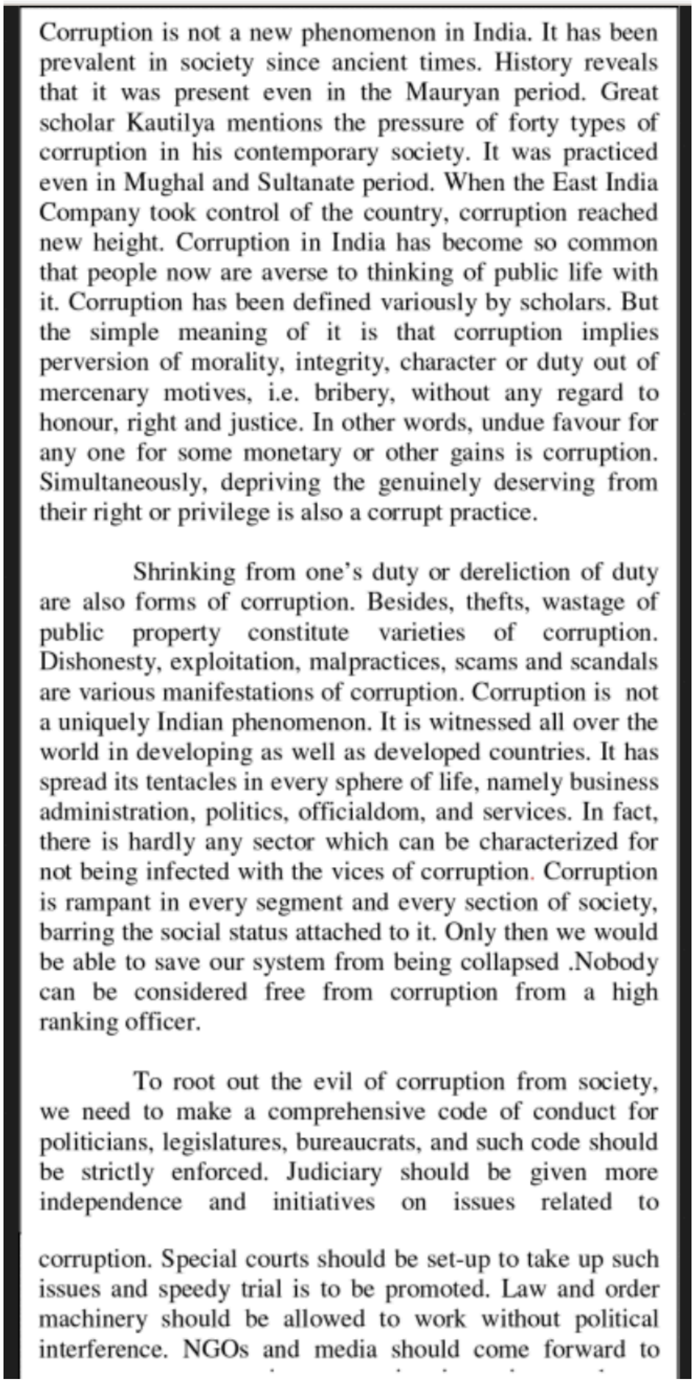
$$\mathrm{P(Precision)= Sref{\cap}Scan / Scan}$$

$$R(Recall)= Sref{\cap}Scan/Sref$$

$$F_{Value} = 2PR/P+R$$

“We have computed the precision, recall and F values for the summaries generated by online summarizer [5] and Word summarizer [9] as reference summary by comparing them with Scan. Finally we compare the P, R, F-values corresponding to our summarizer with these values”

Implementation details -
I tested my results on a sample dataset -
(source - the research paper and article I read)

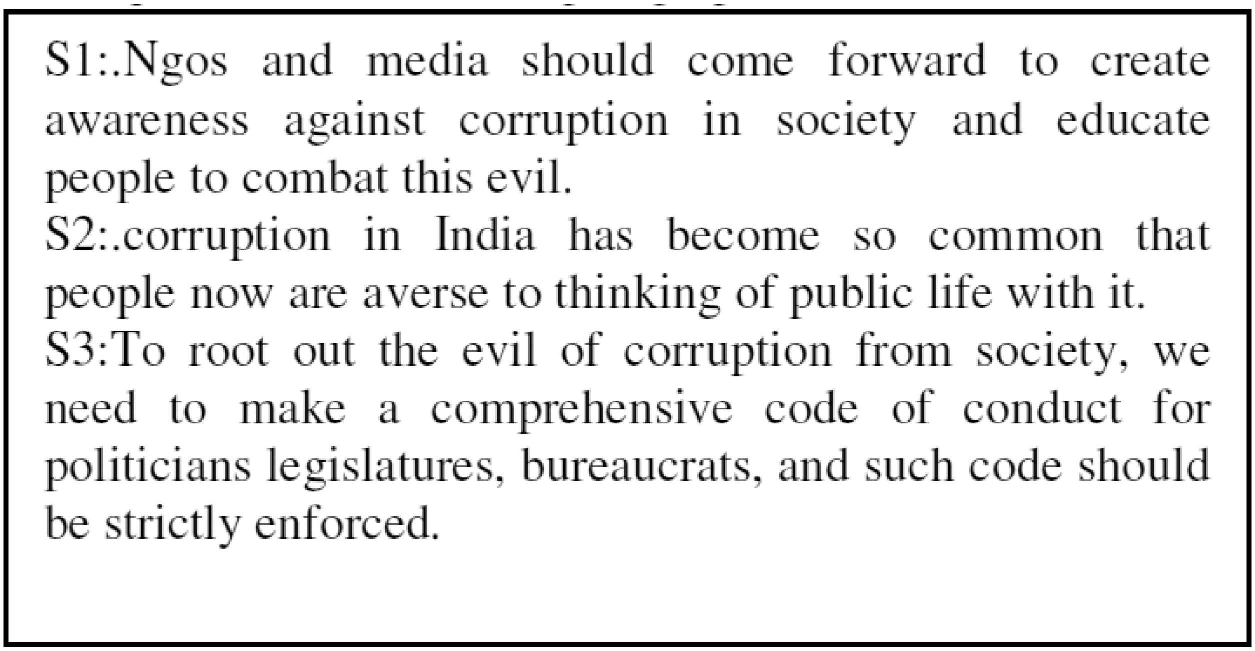


Corruption is not a new phenomenon in India. It has been prevalent in society since ancient times. History reveals that it was present even in the Mauryan period. Great scholar Kautilya mentions the pressure of forty types of corruption in his contemporary society. It was practiced even in Mughal and Sultanate period. When the East India Company took control of the country, corruption reached new height. Corruption in India has become so common that people now are averse to thinking of public life with it. Corruption has been defined variously by scholars. But the simple meaning of it is that corruption implies perversion of morality, integrity, character or duty out of mercenary motives, i.e. bribery, without any regard to honour, right and justice. In other words, undue favour for any one for some monetary or other gains is corruption. Simultaneously, depriving the genuinely deserving from their right or privilege is also a corrupt practice.

Shrinking from one's duty or dereliction of duty are also forms of corruption. Besides, thefts, wastage of public property constitute varieties of corruption. Dishonesty, exploitation, malpractices, scams and scandals are various manifestations of corruption. Corruption is not a uniquely Indian phenomenon. It is witnessed all over the world in developing as well as developed countries. It has spread its tentacles in every sphere of life, namely business administration, politics, officialdom, and services. In fact, there is hardly any sector which can be characterized for not being infected with the vices of corruption. Corruption is rampant in every segment and every section of society, barring the social status attached to it. Only then we would be able to save our system from being collapsed .Nobody can be considered free from corruption from a high ranking officer.

To root out the evil of corruption from society, we need to make a comprehensive code of conduct for politicians, legislatures, bureaucrats, and such code should be strictly enforced. Judiciary should be given more independence and initiatives on issues related to corruption. Special courts should be set-up to take up such issues and speedy trial is to be promoted. Law and order machinery should be allowed to work without political interference. NGOs and media should come forward to

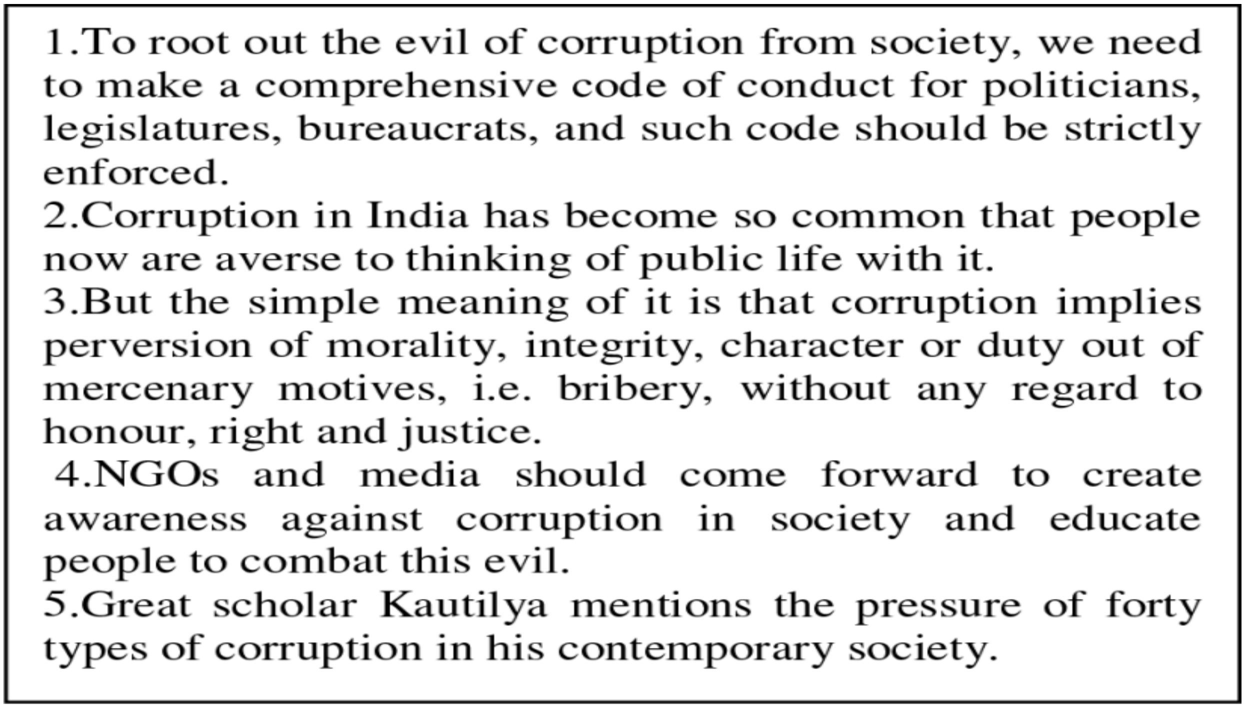
My extractive summary -
1st iteration -> threshold value - 0.57
7 sentences were picked
But the number of paragraphs is only 3
2nd iteration -> threshold value - 0.70
3 sentences were picked -



S1:.Ngos and media should come forward to create awareness against corruption in society and educate people to combat this evil.
S2:.corruption in India has become so common that people now are averse to thinking of public life with it.
S3:To root out the evil of corruption from society, we need to make a comprehensive code of conduct for politicians legislatures, bureaucrats, and such code should be strictly enforced.

Now the number of sentences == number of paragraphs so the algorithm stops.

On comparison with other online document summarizer -



1.To root out the evil of corruption from society, we need to make a comprehensive code of conduct for politicians, legislatures, bureaucrats, and such code should be strictly enforced.
2.Corruption in India has become so common that people now are averse to thinking of public life with it.
3.But the simple meaning of it is that corruption implies perversion of morality, integrity, character or duty out of mercenary motives, i.e. bribery, without any regard to honour, right and justice.
4.NGOs and media should come forward to create awareness against corruption in society and educate people to combat this evil.
5.Great scholar Kautilya mentions the pressure of forty types of corruption in his contemporary society.

Multi-Document Summarization By Sentence Extraction

Multi-document summarization differs from single in that the issues of compression, speed, redundancy and passage selection are critical in the formation of useful summaries.Ideally, multi-document summaries should contain the key shared relevant information among all the documents only once, plus other information unique to some of the individual documents that are directly relevant to the user's query.
The main things we need to factor are -

1. Anti-redundancy methods
2. Articles may contain a temporal dimension
3. Compression ratio - (size of summary w.r.t document size)
4. Co-referencing problem

This model builds up on the single document summarization model - MMR - Maximal marginal relevance.
‘’Most modern IR search engines produce a ranked list of retrieved documents ordered by declining relevance to the user's query. However, there is no known way to directly measure new-and-relevant information. A first approximation to measuring relevant novelty is to measure relevance and novelty independently and provide a linear combination as the metric. We call the linear combination "marginal relevance" -- i.e. a document has high marginal relevance if it is both relevant to the query and contains minimal similarity to previously selected documents.’’
On the basis of their definition of MR our aim is to maximise it i.e Maximal MR = MMR
Formula used to compute MMR -

MMR(C,Q,R,S)=Argmax[X*Sim 1 (Di,Q)-(1-X)Max(Sim2(Di,Dj))]
Di - is in R\S
Dj - is in S
0<=x<=1

MMR computes incrementally the standard relevance-ranked list when the parameter X=1, and computes a maximal diversity ranking among the documents in R when X=0.
Sim1 and sim2 - are similarity metrics for eg cosine similarity.

Procedure -

1. Segment the documents into passages, and index them using inverted indices (as used by the IR engine). Passages may be phrases, sentences, n- sentence chunks, or paragraphs.

2. Identify the passages relevant to the query using cosine similarity with a threshold below which the passages are discarded.
3. Apply the MMR-MD metric. Depending on the desired length of the summary, select a number of passages to compute passage redundancy using the cosine similarity metric and use the passage similarity scoring as a method of clustering passages.
4. Reassemble the selected passages into a summary document using a summary-cohesion criteria.

Our aim is to first computer R using X = 1 in the MMR defined above(i.e. select sentences based on their similarity and relevance to the query) then re-compute the summary using MMR re-ranking by taking X = 0.7 and X = 0.3
Also the passages are taken to be sentences.
Set the compression ratio = 0.2%

Tested on sample CNN data sets.

Code sample to understand major functions -
(hyperparameters)
Set lambda = 0.5
Number of words = 200
Size of best query = 10

```
//
TF_IDF_w = tf_idf(sentences)

   query = bestquery(sentences,TF_IDF_w, 10)

   bestsent = bestscoringsent(sentences, query)

   #vary length and lambda
   summary = selectsent(sentences, bestsent, query, 200, 0.5)
   final_summary = " "
   for sent in summary:
      final_summary = final_summary + sent.getoriginalwords() + " ";
//
```

Summary generated for multiple documents -
****

```
“My main reaction,” he said, “was it’s a lot easier to talk to beautiful
women in a bar when you’re working on a hit show. That was too late for
Mr. Fernandez. That’s how I feel. ” In addition, Mr. Leahy is the music
```

supervisor for “Survivor’s Remorse” on Starz. “That perception really sticks,” he said. Devotees of the “Harry Potter” movies were saddened by the death of Alan Rickman, who played the deliciously dour professor Severus Snape in that blockbuster franchise but whose career, on both stage and screen, was far richer than many of Snape’s younger fans may have known. “That shame that was on my shoulders went off,” he said. Just how is Hillary Kerr, the founder of a   digital media company in Los Angeles? Retired detectives are skeptical that community relations alone can drive down crime in the city’s last the busiest precincts. But of the dozens of films on which he worked, it was for “Bambi” that Mr. Wong was —   belatedly —   most renowned. ” Because Mr. Wong’s father had previously lived in the United States as Look Get, he was able to clear Immigration quickly.

****

For single document -
(hyperparameters)
Set lambda = 0.5
Number of words = 200
Size of best query = 10

****

“My main reaction,” he said, “was it’s a lot easier to talk to beautiful women in a bar when you’re working on a hit show. That was too late for Mr. Fernandez. That’s how I feel. ” In addition, Mr. Leahy is the music supervisor for “Survivor’s Remorse” on Starz. “That perception really sticks,” he said. Devotees of the “Harry Potter” movies were saddened by the death of Alan Rickman, who played the deliciously dour professor Severus Snape in that blockbuster franchise but whose career, on both stage and screen, was far richer than many of Snape’s younger fans may have known. “That shame that was on my shoulders went off,” he said. Just how is Hillary Kerr, the founder of a   digital media company in Los Angeles? Retired detectives are skeptical that community relations alone can drive down crime in the city’s last the busiest precincts. But of the dozens of films on which he worked, it was for “Bambi” that Mr. Wong was —   belatedly —   most renowned. ” Because Mr. Wong’s father had previously lived in the United States as Look Get, he was able to clear Immigration quickly.

****

(hyperparameters)

Set lambda = 0.7 (increased relevance)
Number of words = 100
Size of best query = 10

```
********************
“My main reaction,” he said, “was it’s a lot easier to talk to beautiful
women in a bar when you’re working on a hit show. “But he was more than
that,” Mr. Canemaker explained. ” In addition, Mr. Leahy is the music
supervisor for “Survivor’s Remorse” on Starz. “That shame that was on my
shoulders went off,” he said. Just how is Hillary Kerr, the founder of a
digital media company in Los Angeles? That’s how I feel. Retired
detectives are skeptical that community relations alone can drive down
crime in the city’s last the busiest precincts.

************************
```

Set lambda = 0.3 (increased similarity)
Number of words = 100
Size of best query = 10

```
*********************************

“My main reaction,” he said, “was it’s a lot easier to talk to beautiful
women in a bar when you’re working on a hit show. That was too late for
Mr. Fernandez. That’s how I feel. ” In addition, Mr. Leahy is the music
supervisor for “Survivor’s Remorse” on Starz. “He created an art direction
that had really never been seen before in animation. Devotees of the
“Harry Potter” movies were saddened by the death of Alan Rickman, who
played the deliciously dour professor Severus Snape in that blockbuster
franchise but whose career, on both stage and screen, was far richer than
many of Snape’s younger fans may have known.

*****************************
```

Generally, we take a combination of the summaries generated when lambda = 0.3 and lambda = 0.7.

Best results were obtained for lambda = 0.5
For multi-documents if the documents are related, more concrete summaries are generated.

K-means Clustering -

Comparison of different word embeddings -

1. Word2vec - imported from Gensim, trained on the first 15,000 articles from the dataset. Summary generated -

```
      She was a brave young Dutch student and a gentile who risked
her life to save Jews from death camps in the early 1940s, in one
instance shooting a Nazi stooge before he could seize three little
children she had been hiding. The activity quieted briefly, but it
returned after the young men rented a room in a woman’s apartment
upstairs. That smiling skinny man pedaling his bicycle among the
honking cabs in a blue French worker’s jacket with a camera slung
around his neck  —   what a picture! “I haven’t got one single phone
call that’s putting me in the right direction here,” said Sergeant
LoPuzzo, the head of the precinct’s detective squad, one day this
summer as he worked on an answer to an email inquiry from a murder
victim’s aunt about why the killer had not been caught. And that,
said Dr. Michael Schwartz, an obesity and diabetes researcher who is
a professor of medicine at the University of Washington, is “new and
important. He said a recent decision to bring gang, narcotics and
vice detectives under his command made it easier to shift personnel.
Mr. Fernandez and his wife went so far as to give officers keys to
the building door, so they could get in whenever they wanted, showed
them the videos and offered them   access to his camera so they
could see what was happening in the hallway. She can tell you what
song was playing five years ago on the jukebox at the bar where she
somewhat randomly met the man who became her husband. Still
determined to live something like a normal life, they started a
family. Because a huge number of municipal documents, including
birth and immigration records, were destroyed, many newly arrived
Chinese capitalized on the loss, maintaining that they had been born
in San Francisco before the fire. And before they left for their
Hawaiian honeymoon, Mr. Leahy completed his first important act as
husband. “This looks like it’s art directed,” Eva Chen, the head of
fashion partnerships for Instagram, said as she took it all in. ”
The Bronx’s struggles extend into prosecutions. In Sacramento, where
he joined his father, a schoolteacher Americanized “Tai Yow” to
“Tyrus,” and he was known as Tyrus Wong ever after. “My main
reaction,” he said, “was it’s a lot easier to talk to beautiful
women in a bar when you’re working on a hit show.
```

1. Using fasttext -

    Their first child, José Jr. had a heart defect. Young men scan court paperwork in prison, looking for the names of people who turned on them. “This is a subset of the most successful” dieters, he said. The bride and groom planned accordingly. But analysts in the region said the North’s launchings of rockets to put satellites into orbit in recent years showed that the country had cleared some key technological hurdles. ’” Invoking the exquisite landscape paintings of the Song dynasty (A. D. 960 — 1279) he rendered in watercolors and pastels a series of nature scenes that were moody, lyrical and atmospheric — at once lush and spare — with backgrounds subtly suggested by a stroke or two of the brush. ” Mr. Cahill left the show’s stage in Hollywood and flew directly to New York to start a triumphal tour of the talk shows, chatting with Jay Leno, Regis Philbin and Joy Behar. The band’s lead singer is Jack Antonoff, who is Ms. Dunham’s boyfriend. One flung a racial epithet at him. Three of the precinct’s 16 detectives are carrying more than 400 cases each this year, and many others have loads in the high 300s, even though the department advises 150 in violent precincts. Like the film’s title character, the artist, Tyrus Wong, weathered irrevocable separation from his mother — and, in the hope of making a life in America, incarceration, isolation and rigorous interrogation — all when he was still a child. From 1936 to 1938, Mr. Wong was an artist for the Works Progress Administration, creating paintings for libraries and other public spaces. Only when that challenge is solved, they say, can progress truly be made against obesity. The queen, who ascended to the throne in 1952, became the world’s monarch following the death of King Bhumibol Adulyadej of Thailand in October. “Your music has changed my life!” she told Mr. Leahy.

2. Using BERT -

    But analysts in the region said the North’s launchings of rockets to put satellites into orbit in recent years showed that the country had cleared some key technological hurdles. ’” Invoking the exquisite landscape paintings of the Song dynasty (A. D. 960 — 1279) he rendered in watercolors and pastels a series of nature scenes that were moody, lyrical and atmospheric — at once lush and spare — with backgrounds subtly suggested by a stroke or two of the brush. ” Mr. Cahill left the show’s stage in Hollywood and flew directly to New York to start a triumphal tour of the talk shows, chatting with Jay Leno, Regis Philbin and Joy Behar. He said a recent decision to bring gang, narcotics and vice detectives under his command made it easier to shift personnel. Mr. Fernandez and his wife went so far as to give officers keys to the building door, so they could get in whenever they wanted, showed them the videos and offered them access to his camera so they could see what was happening in the hallway. She can tell you what song was playing five years ago on the jukebox at the bar where she somewhat randomly met the man who became her husband. Still determined to live something like a normal life, they started a family. Because a huge number of municipal documents, including birth and immigration records, were destroyed, many newly arrived Chinese capitalized on the loss, maintaining that they had been born in San Francisco before the fire.

1. Using Glove -

   Still determined to live something like a normal life, they started a family. Because a huge number of municipal documents, including birth and immigration records, were destroyed, many newly arrived Chinese capitalized on the loss, maintaining that they had been born in San Francisco before the fire. And before they left for their Hawaiian honeymoon, Mr. Leahy completed his first important act as husband. “This looks like it’s art directed,” Eva Chen, the head of fashion partnerships for Instagram, said as she took it all in. ” The Bronx’s struggles extend into prosecutions. Invoking the exquisite landscape paintings of the Song dynasty (A. D. 960 — 1279) he rendered in watercolors and pastels a series of nature scenes that were moody, lyrical and atmospheric — at once lush and spare — with backgrounds subtly suggested by a stroke or two of the brush. ” Mr. Cahill left the show’s stage in Hollywood and flew directly to New York to start a triumphal tour of the talk shows, chatting with Jay Leno, Regis Philbin and Joy Behar. He said a recent decision to bring gang, narcotics and vice detectives under his command made it easier to shift personnel.

SAKSHI and SHARUT

TEXT SUMMARISATION (DRESS)

TEXT SUMMARISATION (Embedding Features)

Ranking Sentences for Extractive Summarisation

Previous RL papers

1. First paper using RL for text summarisation (exactractive): Ryang and Abekawa (2012). The authors regard the extractive summarization task as a search problem. In their work, a state is a subset of sentences and actions are transitions from one state to the next state. They only consider the final score of the whole summary as reward and use TD(λ) as RL framework.
2. Rioux et al. (2014) extended this approach by using TD. They employed ROUGE as part of their reward function and used bi-grams instead of tf ∗ idf as features.
3. Henß and Mieskes (2015) introduced Q-learning to text summarization. They suggest RL-based features that describe a sentence in the context of the previously selected sentences and how adding this sentence changes hypothetical summary

Why Reinforcement Learning for text summarisation:

In RL, The system learns the optimal policy that can choose a next action with the most reward value in a given state. That is to say, the system can evaluate the quality of a partial summary and ***determine the sentence to insert in the summary to get the most reward***. It can produce a summary by inserting a sentence one by one with considering the quality of the hypothetical summary.

Why not other RL Approaches:

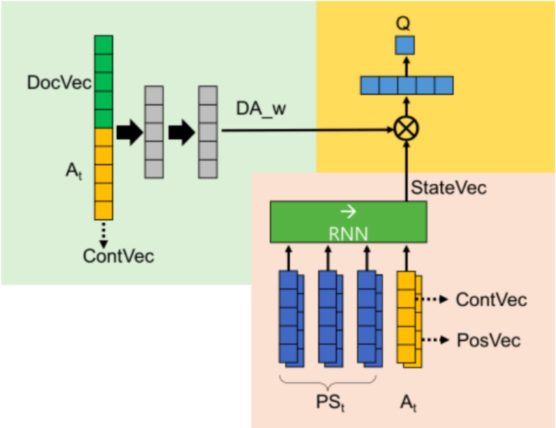
Previous studies mainly exploited handcrafted complex features in RL-based automatic text summarization. However, choosing important features for a task, re-implementing the features for a new domain, and re-generating new features for a new application are very difficult and time consuming jobs. So this approach tries to reduce the burden of hand-crafted features. We use **'Embedding' (Content embeddings vector and position embeddings vector).** This network hence is devised to consider the relevance of a candidate sentence for an entire document as well as the naturalness of a generated summary.

Our Model:

The state denotes a summary which can still be incomplete and the action denotes the addition of a sentence to this summary.

Reward(state(t), state(t+1) = Score(state(t+1), Hd) - Score(state(t),Hd)

Where Score measures the quality of the partial summary (state) by comparing it with the corresponding human reference summary using ROUGE-2.



Module 1 (Upper left)

Input:

The input document is provided as DocVec in which the whole meaning of the document is embedded. The candidate sentence is represented as a sentence embedding vector ContVec.

Output:

Output is DA_w found after applying sigmoid, tanH and finally concatenation.

Module 2 (Bottom Right): represents the relationship between the partial summary and the candidate sentence

Input

Using RNN, Each sentence is represented as the two vectors which contain the content information(ContVec) and the position information(PosVec).

Output

The partial summary, candidate sentences are transformed into the final state vector (StateVec).

Module 3 (Top Right):

Input: DA_w and StateVec

Output: Q value using linear regression i.e. the expected value that can be obtained when a new summary is generated by inserting a candidate sentence into a partial summary.

Features:

1. ContVec is estimated by the average of word embeddings in a sentence
2. The positional information has three views:
    a. An absolute position of the sentence (PosSent) within a paragraph
    b. An absolute position of the paragraph (PosPara) in which the sentence belongs within a section

    c. An absolute position of the section (PosSect) in which the sentence belongs in a document

3. PosVec = element wise sum of {PosSectvec, PosParaVec, PosSentVec}

TEXT SUMMARISATION (REFRESH)

Ranking Sentences for Extractive Summarisation using RL

Introduction -

1. Existing models rely on recurrent neural networks.
2. They are trained using cross- entropy loss in order to maximize the likelihood of the ground-truth labels and do not necessarily *learn to rank* sentences based on their importance due to the absence of a ranking-based objective.
3. There is mis-match between the learning objective and the evaluation criterion (ROUGE) which takes the entire summary into account.

New Objective

Paper conceptualizes extractive summarization as a **sentence ranking** task and proposes an algorithm which **globally optimizes the ROUGE evaluation metric** through a reinforcement learning objective.

During training, the algorithm combines the maximum - likelihood cross-entropy loss with rewards from policy gradient reinforcement learning to directly optimize the evaluation metric.

Dataset used for testing - CNN and Daily Mail

Method

Given a document D consisting of a sequence of sentences $(s1,s2,...,sn)$, an extractive summarizer aims to produce a summary S by selecting m sentences from D (where $m < n$). For each sentence $si \in D$, we predict a label $yi \in \{0,1\}$ (where 1 means that si should be included in the summary) and assign a score $p(yi|si,D,\theta)$ quantifying si's relevance to the summary.

We estimate $p(yi|si,D,\theta)$ using a neural network model and **assemble a summary S by selecting m sentences with top $p(1|si , D, \theta)$ scores.**

Sentence Encoder

Temporal narrow convolution is used by applying a kernel filter *K* of width *h* to a window of *h* words in sentence *s* to produce a new feature.

Document Encoder

It composes a sequence of sentences to obtain a document representation. We use a recurrent neural network with Long Short-Term Memory (LSTM) cells to avoid the vanishing gradient problem.

Sentence Extractor

It sequentially labels each sentence in a document with 1 (relevant for the summary) or 0 (otherwise). It is implemented with another RNN with LSTM cells and a softmax layer.

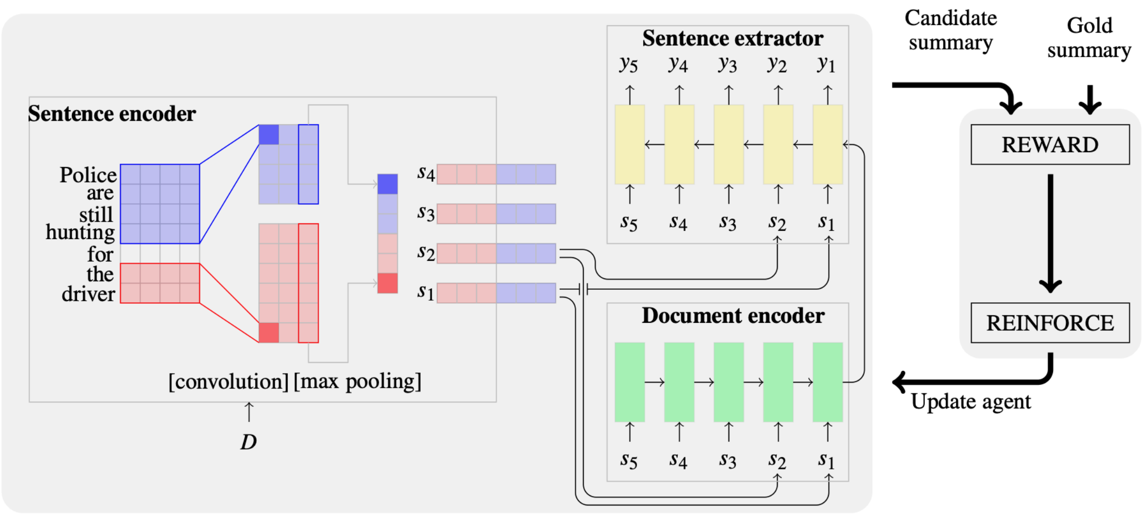


Figure 1: Extractive summarization model with reinforcement learning: a hierarchical encoder-decoder model ranks sentences for their extract-worthiness and a candidate summary is assembled from the top ranked sentences; the REWARD generator compares the candidate against the gold summary to give a reward which is used in the REINFORCE algorithm (Williams, 1992) to update the model.

Cross-entropy training leads to two kinds of discrepancies -

1. Maximum likelihood aims to maximize the likelihood of the ground-truth labels but the model is expected to rank sentences to generate a summary and is evaluated using ROUGE at test time.

2. Document collections for training summarization systems do not naturally contain labels indicating which sentences should be extracted. Instead, sentence labels are extrapolated from abstractive summaries and are not reliable.

Reinforcement Learning Model -

We propose an objective function that combines the maximum-likelihood cross-entropy loss with rewards from policy gradient reinforce- ment learning to globally optimize ROUGE.

The original cross-entropy loss function and the new loss function proposed are mentioned in the paper.

Evaluation Scheme -

When evaluated automatically (in terms of ROUGE), our model out-performs state-of-the-art extractive *and* abstractive systems.

Two human evaluations are -

(a) which type of sum- mary participants prefer (extractive/abstractive).

(b) how much key information from the document is preserved in the summary.

Advantages -

1. A novel application of reinforcement learning
2. Cross-entropy training is not well-suited to the summarization task
3. Large scale user studies following two evaluation paradigms which demonstrate that state-of-the-art abstractive systems lag behind extractive ones when the latter are globally trained.