# Abstract

The field of automated text was pioneered by Hans Peter Luhn a researcher at IBM, with the abundance of information today it is difficult to assess the relevance of a document without reading a large chunk of text. Summaries indicate the most salient points of the document and the overall topic it addresses.

To generate quality summaries the initial writer needs to summarise his document, often they find this to be redundant and do not or an unbiased third-party reviewer who understands the topic reads the document and creates a condensed writeup. With the number of documents available and with the number increasing this is infeasible.

Machine learning offers a way to automate summarization of documents quickly and accurately in large corpora of data.

# Introduction

Machine learning is a type of computational algorithmics where the algorithm improves itself as it is run over the corpus of data, “learning”. When implementing Machine learning for automated text summarization is split into main approaches abstractive and extractive, abstractive focused on the individual words and re-generates texts and content while extractive instead ranks sentences by prominence.

The results of the summarization are split into two types as well query based where a topical query is given and the machine learning algorithm summarizes based on the query, the alternative a generic summaries that simply condense the most important details of the document.

Research in automated summarization has come to the consensus that summarization models work best on corpora with similar topical content in both results topical similarity helps the algorithm learn.

# Papers

## Paper 1

Samidha et al. propose an Extractive Text summarization model using RBM in combination with fuzzy logic to generate a meaningful losses summary of large singular text documents. The model takes several word and sentence features which are used to generate two summaries with a restricted Boltzmann Machine (RBM) and Fuzzy logic these summaries are then combined to make the final summary.

### Models

The feature extraction paradigm is based upon Luhn’s work which states the initial and final 7% of a document are the most meaningful and that longer sentences are more meaningful than shorter sentences.

The text is pre-processed, it is broken down into a n ordered matrix of sentences and a bag of words. Feature extraction processes are then carried out on the tokenized document: Sentence position rank, Sentence length ratio, Numerical token ratio, TF-ISF (Term frequency- Inverse Sentence Frequency),Cosine similarity to centroid (TF-ISF), Bi-Gram/Tri-gram (Calculated using the NLTK libraries), Proper Noun ratio, Thematic Word ratio.

### These features are then combined to form a sentence feature matrix, as a result each sentence will have nine features. The matrix is then normalized by dividing the values by the largest value which is then fed into the RBM.

RBM’s are a variant of Boltzmann Machines (BM) which are stochastic generative networks, in BM’s each note is connected to the other with symmetric connections, In a RBM model there is a clear division of visible and hidden nodes into two layers, where the visible notes are only connected to nodes in the next hidden layer and not parallel nodes. The RBM and BM do not process inputs linearly instead nodes are reset changing their states based on a Boltzmann distribution until the entire network is in equilibrium where the probability distribution has converged across the network.

In the RBM this equilibrium is achieved through forward and backward passes from and to the visible nodes, a randomly generated bias is added to the hidden nodes on the forward and backward pass but it is only added to the visible node on the backward pass.

### For the forward pass, the following equation is used to determine the probability whether a node will activate:

### p(𝑆i |𝑠j ) = σ(∑ sj × wij + bi )

### Where Si is the following node and sj is the preceding node, and the sigmoid equation is:

### 𝜎(𝑥) = 1/ (1+e- 𝑥)

### For the backward pass, a different equation is used to calculate the values for the same sigmoid equation:

### p(𝑠j |𝑆i ) = σ(∑mi=1 sj × wij + bi )

### Through these equations the values of the node’s inputs are predicted this is known as Gibbs Sampling. The differences between the input values and the new values are used to gain the training loss using contrasted difference calculated by the following equations:

### wij new = wij old + (LR × Δw)

### Where LR is the learning rate, Δw the change in weights as given by the difference of the probabilities of the node outputs:

### Δw = (𝑠j ⊗ 𝑝(𝑆i|𝑠j)− 𝑠j′ ⊗ 𝑝(𝑆i ′|𝑠j ′)

Once the network has reached equilibrium then an improved feature matrix is output, the values of these improved features are summed together to give a sentence score. To generate the summary the first sentence of the document is taken then the remaining sentences are ranked by score the upper 50% are taken for the summary and then sorted according to their position in the original document to give the first summary.

The fuzzy logic summary is generated by converting the previous feature scores into percentages these percentages are then sorted using triangular membership functions: HIGH, MEDIUM and LOW. Fuzzy logic IF-THEN rules are applied to this set to de fuzz the sentences into new categories of Important, Average and Unimportant the Important are then used to create the second summary by sorting them into their occurrence in the original document.

To generate the final summary sentences common to both are taken from both, the remaining uncommon sentences are then sorted by their score and the top 50% are taken. These sentences like before are then sorted into their original position in the document.

### Experiments

### Two main experimental paths were taken a control and the proposed method, the control would the RBM generated summary to act as a baseline to the results gained from the proposed method if it improved upon the RBM summary. To evaluate the results the ROUGE metric was used, using the precision, recall and F measure to compare the two summaries.

### Evaluation

### Ten documents were summarized, and the evaluation results were compared between the two from the precision, recall and F-measure the new method showed better results by a few decimal points.

### Conclusions

## While the results showed that the proposed method was better there was no review of what the contents of the documents to be evaluated were and how the size of the document, its sentences might affect the summarization techniques or how the nine features contribute to the model. The number of documents used were also particularly small.

## The RBM used as a control was too similar to the proposed method a more mainstream version of automated summarization should have been used as a control.

## Paper 2

### Models

Models

### Experiments

Experiments

### Evaluation

Evaluation

### Conclusions

conclusion

## Paper 3

### Models

Models

### Experiments

Experiments

### Evaluation

Evaluation

### Conclusions

conclusion

## Paper 4

### Models

Models

### Experiments

Experiments

### Evaluation

Evaluation

### Conclusions

conclusion