# A Literature Review of Query Relevant Summary Generation Using Machine Learning Algorithms to Generate Highlighted Version with The Most Relevant Information

Juan E. Junco
Earlham College
Richmond, Indiana
Jejunco20@earlham.edu

**ACM Reference Format:**
Juan E. Junco. 2018. A Literature Review of Query Relevant Summary Generation Using Machine Learning Algorithms to Generate Highlighted Version with The Most Relevant Information. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 3 pages. https://doi.org/XXXXXXX.XXXXXXX

## 1 INTRODUCTION

The massive number of digital texts on a single topic increases the difficulty, Generating an accurate summary of texts. Now that we live in a world where there is vast information o mostly every topic having a way to extract essential ideas from a text becomes critical. Summary generation is a field that has increasingly gained more importance due to extracting the most valuable information in the most efficient time. Although there have been many methods using vectors, machine learning, and other traditional computing algorithms, one method has not yet been explored. Designing software that highlights the most relevant sentences in a text would allow the user to identify the fundamental ideas of a text using the author's words without having to go through sentence generation or reconstruction. Currently, most methods extract the most relevant sentences or words and categorize them based on statistical analysis. It uses language processing techniques in order to show the output. This literature review will cover significant research on how text highlighting can use summary generation strategies. First, the text will explain the types of summaries extracted for an online text. Second, it will explain the different methods outstanding researchers have suggested in the field. This section will be divided into two. First, all the traditional methods that are used using known search algorithms. Second, research that uses NLP (Natural Language Processing) and statistical analysis to generate summaries. The last part of the document will conclude by suggesting possible future work in the field. Although various approaches exist for summary generation, this literature review only includes some of them and mainly focuses on Convolutional Neural Networks (CNN).[10]


*Conference'17, July 2017, Washington, DC, USA*

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/XXXXXXX.XXXXXXX

## 2 SUMMARY TYPES

Text summarization is a big field that is mainly divided into two different way to generate the desired extraction of information. A more general approach that allows a general understanding of the information displayed without taking the intentions of the user to read that specific article. Is more independent approach that merely focuses in the text content as a shorter version, in other words much like "the abstract in a paper, designed to distill salient points"[2]. A well done generic summary contains the most relevant points while avoiding redundancy [3]. The second way to summarize a text is the so called "User Focused Abstracts"[5], or query-relevant summaries. It reflects the importance of a text based on the text query made by the user.Basically it highlights the most important aspects of the text based on what the user is looking for.

## 3 DATASET

This section introduces types of text sources used by researchers. Different researchers have different assumptions and objectives. Therefore, it is important to understand why they choose one text source over another. This section of the literature review is fundamental as it shows what kinds of text have been used to summarize, and give an idea of the types of text that a research can use in order to develop a reliable analysis.

### 3.1 TIPSTER

Across several papers that deal with text summaries, the project TIPSTER came up. TIPSTER (Text Research Collection Volume or TREC) is an effort to significantly advance the state of the art in effective document detection (information retrieval) and data extraction from large, real-world data collections.

The TIPSTER Collection was used for example, in the Automatic Text Categorization research [8]. In their paper they performed two series of experiments and used I all the experiments we have used a document collection available in the TIPSTER document base[6]. They used specifically several magazines about computers. The whole TIPSTER document base contained 33,658 documents with these characteristics. A subset of these documents was randomly selected for the experiments to be reported in this section.

### 3.2 CNN

One of the most significant sources of information found when attempting a new sumarization method were news channels and articles. Using news articles are one of the strongest sources of data when it comes to creating summaries, most of the time, the rigor behind a news article is far less strict than a scientific paper. In

other words if the proposed summarization method is capable of generating a a well build summary with a subjective semantic field, then it will work when using it in reviewed documents like papers or books.

In the text Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis, They used CNN Worldview news programs a their data source to conduct different techniques of text summary. "Their evaluation database consists of closed captions of 549 news stories whose lengths are in the range of 3 to 105 sentences. As summarizing short articles does not make much sense in real applications, for our evaluations we eliminated all the short stories with less than ten sentences, resulting in 243 documents.[3]"

## 4 METHODS

This section describes the types of methods used by major researchers in the Field. This section analyses two types of research: studies that developed and utilized classical methods to create text summaries, and studies that used NLP to create query based summaries.

### 4.1 NLP

Baldwin and Morton [1] implemented NLP in order to perform the analysis to generate a query relevant summarized texts. They implemented the CAMP NLP framework, that allowed them to utilize a wide variety of features of linguistic information. Some of their main components:

- Entity recognition
- Tokenization
- Sentence detection
- Part of speech tagging
- Morphological Analysis
- Parsing
- Argument Detection

In their analysis they broke down three possible relationships between words and sentences in the document. The relationships between nouns were made on the basis of string matches, acronym matching and dictionary look up. Whenever the first words of a proper noun match with an acronym the word was assigned as explanation for that specific acronym. Using the a reverse dictionary lookup Baldwin and Morton were able to associate cities with the corresponding countries they belonged. Using this and some other techniques Baldwin and Morton were able to assign a score to the sentences in the documents by implementing seven different ranking techniques. Later, scores 1-6 were ranked in descending order while score 7 is ranked in ascending order. The final location of the sentences in ascending order after performing all the different categorization ended up with the sentences that were most relevant for the summary production. One of the strongest aspects of this research is the fact that they utilized query based look ups and then proceeded to test the output summary for specific tokenization of the words related with the query on the text.

On the other hand, a comparative study done by Munot and Govilkar [7] also showed how utilizing Natural processing techniques is a viable option when it comes to summary generation. In they comparative study they show how "Abstractive text summarization method generates a sentence from a semantic representation and then use natural language generation techniques to create a summary that is closer to what a human might generate. Such a summary might contain words not explicitly present in the original." Their analysis essentially proceeds to understand the original text and then using less words to re created the meaning. It consists of understanding the original text and re-telling it in fewer words.

### 4.2 Neural Networks

Kaiaikhah [4] implemented a succesful Neural Network (NN) model that generated very accurate summaries of online texts. In his approach Each document is converted into a list of sentences, meaning that sentences can now be represented as a vector, composed of seven features.

**Table 1: Seven Features of a Document**

| | |
|---|---|
| $f_1$ | Paragraph follows title |
| $f_2$ | Paragraph location in document |
| $f_3$ | Sentence location in paragraph |
| $f_4$ | First sentence in paragraph |
| $f_5$ | Sentence length |
| $f_6$ | Number of thematic words in the sentence |
| $f_7$ | Number of title words in the sentence |

**Figure 1: Framework of the Project**

Kaiaikhah broke down the research into three different steps. First training a neural network to recognize the type of sentences that should be included in the summary. Second, prunes the neural network and collapses the hidden layer unit activation into discrete values with frequencies. And, finally the neural network to filter the text and to select only the highly ranked sentences. Sinha, Yadav,

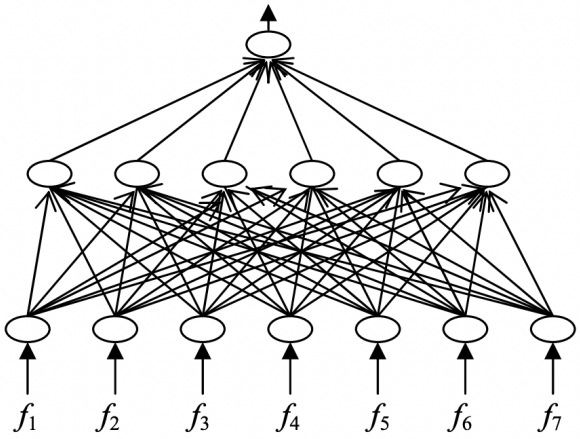


**Figure 1: The Neural Network after Training**

**Figure 2: Framework of the Project**

and Gahlot [9] also implemented a NN to perform an extractive summarization. "Extractive summarizers take sentences as input and produce a probability vector as output." those vectors will later show the probability of a sentence being included in the summary. Their main constrain the length of the document they were summarizing in many cases, approaches like Recurrent Neural Networks and End to End learning have been proposed to deal with such constrain, but in this case they used a recursive solution by breaking amount of words into pages: "Let the number of sentences in the document be '$doc_len$'.Now we divide the document into segments, each having a fixed number of sentences. Each such segment is called a 'page' and let this fixed number be '$page_len$'. In this way we obtain '$num_pg$' pages, where '$num_pg$' equals to ceil($doc_len/page_len$)" Their model consisted of a simple NN with one input layer, one hidden layer and one output layer, with a A softmax activation function in the latter layer. Each entry of the obtained vector denotes the weight associated with the corresponding sentence which represents the measure of belief of the sentence being included in the summary.

## 5 CONCLUSION

This literature review discussed how NN and NLP are used for summary generation. First, I explained what types of online-generated summaries had been used by researchers and the underlying assumptions behind their choices. The abstractive or query-specific methodology completely changes the way a text is summarized based either on the contents of the text itself or on the query the user used to highlight query-relevant results. Secondly, I explained the different sources of data researchers used when generating summaries—starting with the most common among several pieces of research. This TIPSTER project contains a set of digitilized documents categorized by date and topic. The second possible approach is retrieving CNNs news and using them as long text inputs. This would make the approach more resilient as it would treat less strict discourse. Finally presented the different approaches when generating summaries highlighting the methods based on NLP and NN.

## REFERENCES


[1] Breck Baldwin and Thomas S Morton. 1998. Dynamic coreference-based summarization. In *Proceedings of the third conference on empirical methods for natural language processing*. 1–6.

[2] Adam Berger and Vibhu O Mittal. 2000. Query-relevant summarization using FAQs. In *Proceedings of the 38th annual meeting of the association for computational linguistics*. 294–301.

[3] Yihong Gong and Xin Liu. 2001. Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. 19–25.

[4] Khosrow Kaikhah. 2004. Text summarization using neural networks. (2004).

[5] Inderjeet Mani. 2001. *Automatic summarization*. Vol. 3. John Benjamins Publishing.

[6] R Merchant. 1994. The proceedings of the TIPSTER text program: Phase I.

[7] Nikita Munot and Sharvari S Govilkar. 2014. Comparative study of text summarization methods. *International Journal of Computer Applications* 102, 12 (2014).

[8] Joel Larocca Neto, Alex A Freitas, and Celso AA Kaestner. 2002. Automatic text summarization using a machine learning approach. In *Brazilian symposium on artificial intelligence*. Springer, 205–215.

[9] Aakash Sinha, Abhishek Yadav, and Akshay Gahlot. 2018. Extractive text summarization using neural networks. *arXiv preprint arXiv:1802.10137* (2018).

[10] Oguzhan Tas and Farzad Kiyani. 2007. A survey automatic text summarization. *PressAcademia Procedia* 5, 1 (2007), 205–213.