Turkish Journal of Computer and Mathematics Education *Vol.12 No.6 (2021), 4687-4694*
Research Article

A Novel Method for Multi-Variate Text Summarization

Ch Sai Prakash[a], Attuluri Rudra[a], L Rakshitha[a], Dr. J Sirisha Devi[a]
[a]Institute of Aeronautical Engineering, JNTU(H), Telangana

Article History: Received: 10 November 2020; Revised 12 January 2021 Accepted: 27 January 2021; Published online: 5 April 2021

Abstract: In this modern age, where vast quantities of data are accessible on the Internet, it is crucial to provide a better mechanism for extracting information quickly and efficiently. Manually extracting the description of a huge text document is incredibly hard and time-consuming. On the Internet, there is a wealth of text content. As a result, finding relevant documents among the large set of documents available and extracting necessary details from them is a challenge. Automatic text summarization is critical for solving the two problems listed above. The method of identifying the most important and pertinent material in a document or a group of related documents and compacting it into a condensed version while maintaining its overall significance is known as text summarization. Before precluding text summarization, it's important to know the actual import of the Summary. A summary is a text that extracts information from one or more texts and conveys it concisely. The aim of Automatic Text summarization is to covert the source material into a semantically shorter adaption. The most relevant benefit of using a summary is that it shrinks the amount of time it takes to comprehend. Extractive and Abstractive are two types of content summarization techniques. An extractive summary technique involves selecting key sentences, pieces, and other elements from the original report and connecting them into a more manageable structure. An abstractive method is an apprehension of the key ideas in a text and then expressions of those ideas in a plain regular language.

Keywords: Text summarization, TextRank, website, computer science, machine learning, CSS, HTML, extractive summary, abstractive summary, natural language processing, LSTM

1. Introduction

Today we realize that machines have got more paramount than us, and instruct us in day-to-day life, the innovations have reached a level where it does all the tasks of every individual. The field which made this happen is Machine Learning. AI instructs the machines with a specific kind of knowledge which helps them doing all cooperative work. Nowadays examines are being done in the field of investigation. As the title proposes, an Automatic Text summarizer is an online tool that helps in plotting the text without eliminating the actual essence of the original text. There is a pressing need to condense most of the text documentation into shorter chunks that simplify and collect the key points so that we can access it more efficiently and verify if the bigger documents provide the information we need. It significantly facilitates the process of skimming the most important information from the source to create an abbreviated version for a given user and mission. As a result, the aim of Automatically producing text summaries is to produce summaries that are as good as those written by humans.

In general, this tool summarizes the actual information according to the need in four different ways i.e., first, it summarizes the text according to a specific topic that is required and gives an overall synopsis. The second, it skims the text according to the text rank which usually works in the same process as a google search engine, the third slices the information according to the feature, extracts the feature of the sentence, and then evaluates according to the user need and specifications. The above methods fall under the extractive approach and the last state the abstraction-based summarization which builds up the internal explanation of the original text. Classically, most prosperous results fall under the extractive approach but the abstractive approach states the most common results to the problem. The tool mainly focuses on bridging the text with the implementation of Deep Learning ideas and techniques(algorithms). Deep Learning methods are specifically stated for the problem of Text summarization as a section-to-section learning problem. It gives better and promising results than the existing summarization tool.

2. Literature Survey

After reviewing several documents, a conclusion has reached stating that Mallick et al. [1] attempted to upgrade a graph-based approach to extractive text content description. Instead of the regular Term Frequency -Inverse Document Frequency (TF-IDF), is-weighted cosine similarity is used to calculate sentences, which produces intriguing effects as compared to newspapers. Allahyari et al. [2] This paper proposed that extraction modules generate summaries by selecting a subset of the original text of the sentences. The most important sentences from the feedback are summarized in these summaries. As input, a single document or a sequence of documents are used. They describe three distinct tasks that all summarizers must complete to get a greater understanding of how the summarization system works. 1) Create an intermediary version of the input text that conveys the most important aspects of the text.2) Using the representation as a guide to scoring the sentences.3) Build a description consisting of a few sentences. Yu et al. [3] claim that the TextRank algorithm works well for large-scale text mining, especially for automated summarization and keyword extraction, since it is an unsupervised learning method. Text Rank only addresses sentence similarity in automatic summarization systems, ignoring text form and context. To address these flaws, while calculating sentence similarities and changing the weights of nodes, the authors take mathematical and linguistic attributes such as resemblance in terms, paragraph forms, special clauses, sentence positions, and lengths.

Pelevina et al. [4] Their method employs WSD (Word Sense Disambiguation) to learn multi-prototype word embeddings. They use a sense inventory derived from crowdsourcing. Camacho-Collados et al. [5] survey proposed that correctly capturing the semantics of ambiguous terms is critical for NLP systems to understand language. They attempted to compile a list of the most important works on meaning representation learning. The primary difference between these methods is how they model meaning and where they get it. Dr. S.Vijayarani et al. [6] In their research work they tested the efficacy of the seven open-source tokenization approaches as. The token collection is used for further processing including sorting or text mining. Tokenization is used in lexical interpretation in both linguistics and computer science. This process is usually done on word-by-word analysis. However, identifying a phrase can be problematic at times. Usually, a tokenizer commits on basic heuristics. The main purpose of the tokenizer is to classify meaningful terms, which aids in the elimination of stop words. The seven open tools used Nilpotent Tokenizer, Mila Tokenizer, NLTK Word Tokenize, Text Blob Word Tokenize, MBSP Word Tokenize, Pattern Word Tokenize, and Word Tokenization with Python NLTK helped in the process of word tokenizing and let the output for the process of steaming and then the further output to the stop word removal.

Federico Barrios et al. [7] stated about the Text Rank algorithm. The ideas state about changing the method for altering the way lengths between sentences are measured to weigh the edges of the Page Rank graph. Since these similarity measurements are not related to the Text Rank model, they are easy to introduce into the algorithm. We discovered that some of these tweaks resulted in significant enhancements over the original algorithm. In this process, they have extracted the Longest Common Substring and finally, found the cosine distance using TF-IDF (Term Frequency and Inverse Document Frequency) and computed the cousins as a measure of similarity. The cosine produces the values [0,1] where 1 represents the equal vectors and 0 represents orthogonal vectors since the vectors are determined to be positive. There proposed method showed a promising accuracy of more than 2.3% than the baseline.

3. Methodology

The tool states to produce the output according to user convivence, unlike the other summarization tool which are already been existed. Figure 1 refers series of tasks performed by the text summarization tool. Data Acquisition is a method of collecting huge information (data sets) and combine to train the model with text summarization. Data pre-processing is the next step, which involves converting the whole paragraph into sentences and further processing these sentences to remove stop words after this, lemmatization is done that combines different inflected forms into a single analyzable entity. The text with later sent to the Automatic Text Summarizer which summarizes the text according to the user's convivence i.e in our research work the text can be summarized according to four main features: topic-based summary, text rank-based summary, feature-based summary, and abstract -based summary. Our works extend a step forward and make users more flexible to use this tool with a multi-lingual (English, French, Spanish) approach. Finally, the summarized text will be stated as an output with promising results.

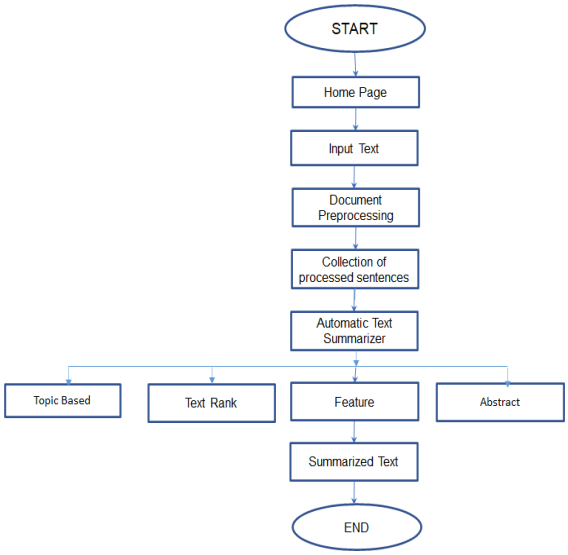

Figure 1: Automatic Text Summarization Tool

3.1. Data Acquisition

Data plays a vital role when it comes to any Natural Language Processing Techniques. During this process, data is collected from various sources(datasets) to summarize the text in abstractive summarization. The more data you get to train your model with, the more precise the results would be.

For our study we have used three open-source data sets:

a) *GIGAWORD data set:* Articles and headlines are commonly found in this data. It's a statement summarization with very short input documents (31.4 tokens) and summaries (8.3 tokens). Our abstract summarization model was developed using this dataset.

b) *CNN Daily Mail set:* This data set generally consists of huge news articles. We incorporated this data set in our Pointer Generator model.

c) *Opinion's data set:* The sentences in this data set were extracted from user feedback on a specific subject. It is usually used in the ROUGE scripts. The abstractive summarization model's findings were scored using this data set.

3.2. Data pre-processing

There are indeed many methods available for extracting data from the internet. To begin the process, we have divided into three main categories:

a) *Converting sentences into words:* The whole paragraph is converted into small sentences. In this process, the most common method used is separating the paragraph in a period it appears.

b) *Removal of stop words:* Stop words are a group of common phrases that can be found in any language, stop words are useful in many applications because they enable one to concentrate on the important words while excluding the words that are often used in a language. So, the process continues to clean up the stop words.

c) *Lemmatization:* Lemmatization is a term that refers to doing it appropriately with a vocabulary and lemma is a morphological study of words to eliminate inflectional endings only and restore the root or dictionary form of a word. Finally, the words are reduced to their root forms, which aids in the speed of the process. Model selection

3.3. Model selection

3.3.1. Text summarization using LSA

Text Summarization is a technique for preventing hidden semantic structures in words and sentences that are based on algebra and statistics. It is an unsupervised process that does not require many instructions or expertise from an outside source. It extracts data from the context of the input text, such as which words are used together and which words appear in different sentences. The meaning of the sentence is determined by the words that are used in it, and the meaning of the words is determined by the sentences in which they appear. Singular Value Decomposition (SVD) is an algebraic approach, where the interrelationships between sentences and terms are discovered. SVD can model interactions between terms and sentences as well as minimize noise, which helps to increase the accuracy.

Summarization algorithms based on the LSA approach usually have three key stages:

1) *Input matrix creation:* Usually, this is represented as a matrix, with columns representing sentences and rows representing words/characters. The context of words in sentences is reflected in the cells. Filling in the values for the cells can be achieved in several ways. Since all terms are not used in all statements, the resulting matrix is usually sparse. Since it affects the SVD-calculated matrices, the way an input matrix is constructed is critical for summarization.
2) *Singular Value Decomposition:* SVD can model interactions between terms and sentences as well as minimize noise, which helps to increase the accuracy.

$$A = U \sum V^T$$

A represents an input matrix (m x n), U represents words x extracted concepts (n x n), V represents words x extracted concepts (n x n).

3) *Sentence selection:* Using the results of SVD the important sentences are selected.

3.3.2. Summarization using Text rank

The first step in this process is to locate the appropriate sentence. The Page Rank algorithm is used to determine text rank. Page Rank is a metric for accessing the value of the websites. It measures the estimated approximation of the value of the website by counting the number and quality of the links to that specific domain. The basic principle is that more relevant websites are likely to obtain more connections from other websites.

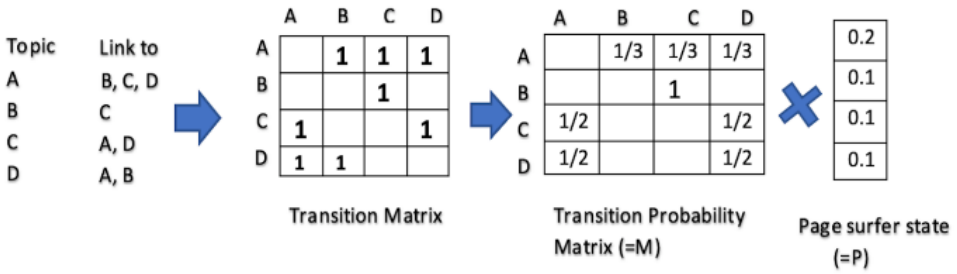

Figure 2. Matrix to a transfer frequency

The following rules state to calculate the text rank for a specific page

1) Solve Eigenvalue problem $MP = P$
2) Repeat the process until it reaches the point of convergence.

$$P_i^{'} = (1 - n) + n * M_i^T P_i$$

Matric to a transfer frequency is depicted in figure 2.

The page suffers randomly click the page with probability

$$\sum \left(P_i^{'} - P_i\right) < threshold\ of\ 1 - d$$

$$usually\ (n = 0.85)$$

On the PageRank system, Text Rank treats words or sentences as articles. The following considerations are important when using Text Rank.

1) Create the "text groups" and add them to the graph as nodes.
2) Define the "relationship" between the text units and make them the graph's edges.

3.3.3. Luhn's Heuristic method for text summarization

It is a function-based model which collects the sentence attributes before assessing its relevance. Here is representative research Luhn's algorithm is an approach based on TF-IDF. Certain features are used – The position of the sentence in the source text, identifying the verdin the sentence, calculating the sentence length, the frequency of the term, tag on a certain individual Named Entity (NE), font design. It chooses the best words based on their frequency. The terms that come at the start of the text are given greater weight. Instead of dividing the number of relevant terms by the total number of words, the value is calculated by dividing them by the word duration.

3.3.4. Text summarization using Abstraction Techniques

Abstractive text summarization focuses to trim the long text records into a human-readable format that includes the most important information from the actual document. However, current methods do have a poor degree of real abstraction, as calculated by novel phrases that do not appear in the source text. Although, Abstractive-based Text Summarization can be stated in two perspectives: Structure-based approach and Semantic-based approach. Structure-based strategies rely on prior information and psychological feature schemas, such as models and extraction rules, as well as flexible alternate structures such as graphs which help to encode the vital data. The semantic-based approach feeds a natural language processing engine with linguistic representations of documents with the primary goal of distinguishing noun and verb phrases. In this method we used the Sequence-to-sequence(seq2seq) method of modeling is used to transform one sequence to another sequence. Long Short-Term Memory (LSTM) encoder and decoder are used to perform the seq2seq modeling. Deep Learning techniques are used for the implementation of seq2seq modeling. This method would be used to translate the English sentences into French sentences. The LSTM encoder and decoder will be illustrated in the below figure 3.which is used for the language conversion.

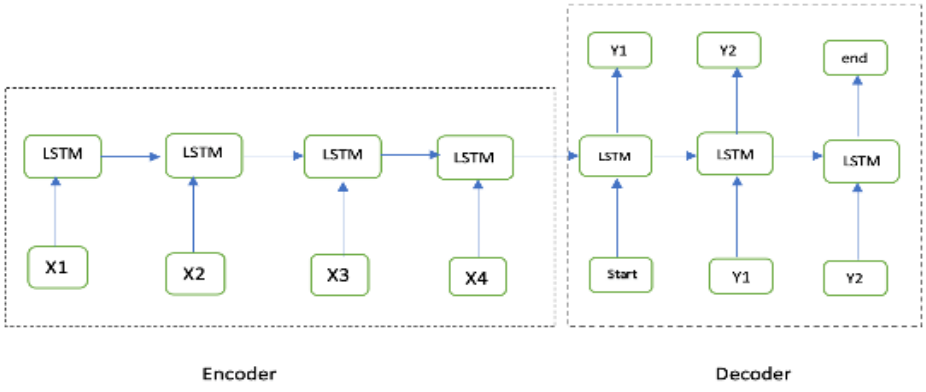

Figure 3. LSTM Encoder and Decoder

3.3.5. Model Evaluation

Rouge - It consists primarily of a collection of metrics for assessing automated text summarization and machine translation. It works by comparing a description or translation generated automatically with a collection of reference summaries easily. In our, paper ROUGE is used as a metric, were used to calculate the accuracy rate of performance measures such as precision, recall, and FI-score.

4. Results and Discussions

This project is being carried out in Google collab. and complete implementation of the text summarization is done using flask framework, HTML. The comparative study between different models is done using an evaluation metric called ROUGE where the performance measures such as precision, accuracy, and f1-score are produced from ROUGE.

4.1. Analysis

In this analysis, the text summarization algorithms Text Rank, LSA, LUHN, and Abstractive are applied to a dataset to check how the speed of the summarizer's scale with the size of the dataset. The tests were performed on a GPU NVIDIA K80. The tests were run on the book "The adventures of Sherlock Holmes" by Arthur Conan Doyle.

The plots below in figure 4, together with data size, visualize the running times. Prefixes of text are selected from the book, in other words, we took the first n characters from the book, to create datasets of different sizes. The algorithm seems polynomial on time, so you must be careful before you plug into the summarizer a large dataset. The algorithms were run to produce the results in such a way that it produces $1/10^{th}$ of the content is generated from the input.

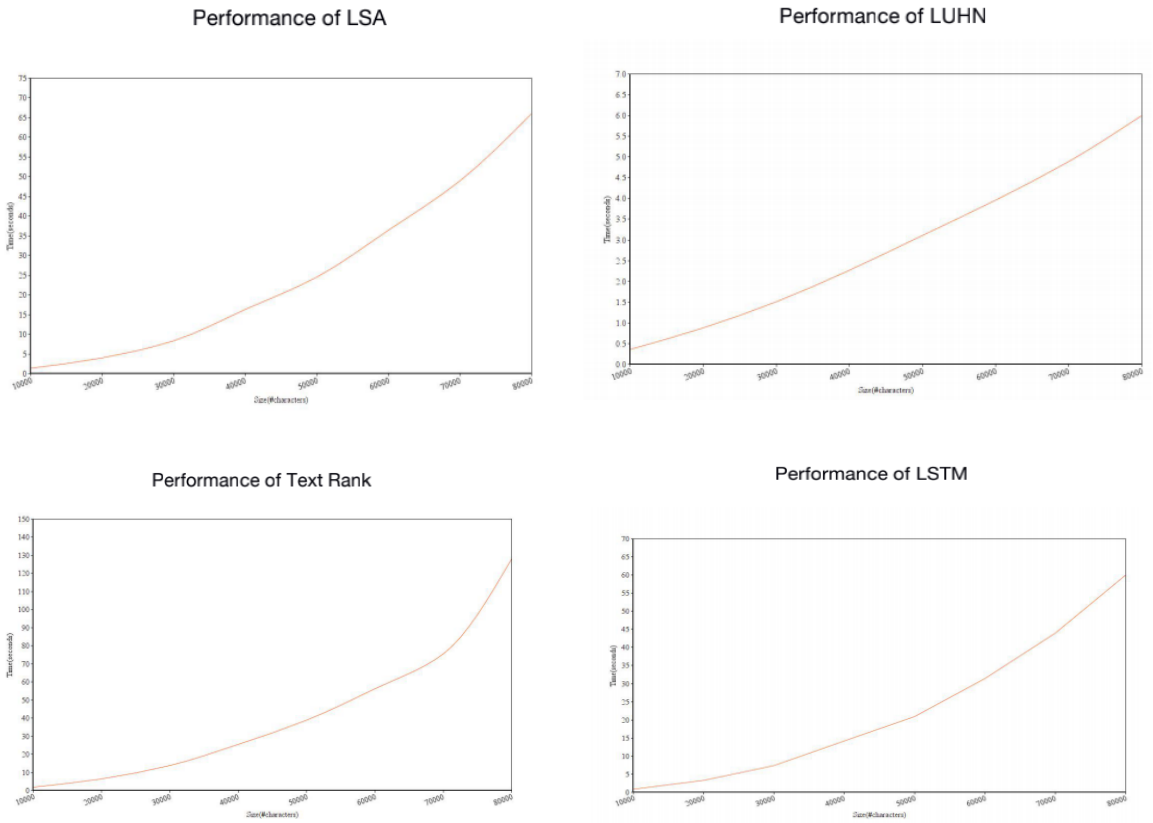

Figure 4. *Line charts of the number of characters and time taken (sec)*

4.2. Accuracy

4.2.1. Precision: Precision refers to what you are fundamentally measuring or How much of the system summary is relevant or necessary.

$$Precision = \frac{numberofoverlappingwords}{totalwordsinsystemsummary}$$

It is shown in table [1] for each classifier the precision is represented.

4.2.2. Recall: Recall refers to how much the system summary is recovered or captured by the reference summary. If we just look at the words, they can be calculated as

$$Recall = \frac{numberofoverlappingwords}{totalwordsinreferencesummary}$$

In table [1] for each model, the recall is represented.

4.2.3. F1-score: It simply uses harmonic mean to fuse recalls and accuracy.

Rouge report for different text summarization models is represented in table 1.

Table 1. Experimental Analysis on ROUGE of Text Summarization

Model	Rouge Report		
	Precision	Recall	F1-score
LSA	0.23	0.62	0.34
Text Rank	0.24	0.62	0.35
LUHN	0.22	0.63	0.36
LSTM	0.54	0.60	0.57

4.3. Tool built to utilize all the models

The Framework has been built to Utilize all four models as required by the user. It is a combination of Flask, Tensor-flow. The tool is language autonomous and an option is provided to the user to select the language. Also depending on how many sentences does the user wants it to be summarized the option to select the number of sentences is also provided. The following figure 5 illustrates the tool for the summarization of various models we have worked on.

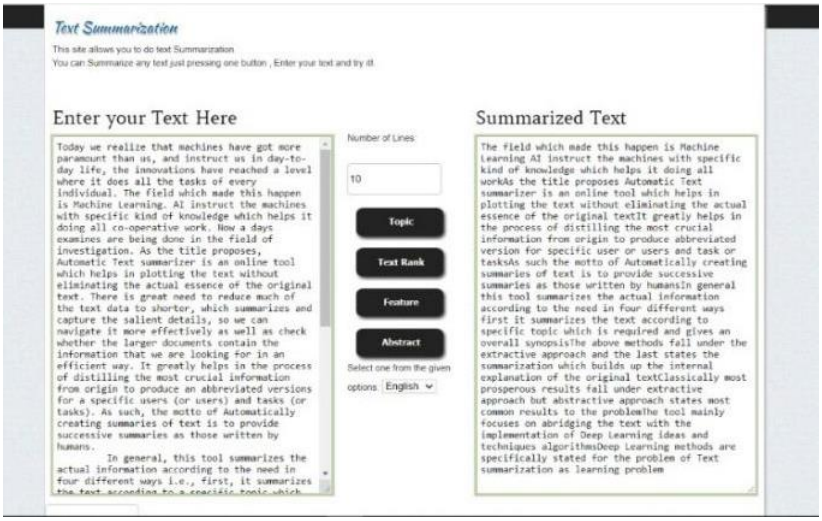

Figure 5. Summarization tool

5. Conclusion

Identifying insight into the needs of the user between vast amounts of information is an immediate problem with the growth of textual resources. Text summary strategies are designed and assessed to solve this problem. The research on synthesis began with the extraction of simple data and continued using various methods including lexicon, linguistic, statistical, graphical, and algebraic approaches.

All approaches were evaluated on an English book called "Adventures of Sherlock Holmes" and the results were taken comparing with other text-summary approaches. For the assessment of the results, we used ROUGE scores. The results show that among various approaches LSTM - seq2seq works better than all others. In the future, the tool "automatic summarizer model" will be added where it runs the text across multiple models and picks the best one for the user.

References

[1]. Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assef," Text summarization techniques: a brief survey", preprint arXiv:1707.02268. 2017.

[2]. Chianti Mallick,Ajit Kumar Das,MAdhurima Dutta,Asit Kumar Das," Graph-based text summarization using modified TextRank. In Soft Computing in Data Analytics ", Springer; 2019. p. 137-146.

[3]. Yu,S., Su, J.,LI, P., Towards high-performance text mining," a TextRank-based method for automatic text summarization", International Journal of Grid and High-Performance Computing (IJGHPC). 2016; 8(2): p. 58-75.

[4]. Maria Peliva, Nikoley Arefyev,Chris Biemann,Alexander Penchanko," Making sense of word embeddings.",arXiv preprint arXiv:1708.03390. 2017.

[5]. Jose Cmacho-Callodos ,Mohmamad Taher Pilevar," From word to sense embeddings: A survey on vector representations of meaning", Journal of Artificial Intelligence Research. 2018; 63: p. 743-788.

[6]. Dr. S.Vijayarani,Ms.R.Janani," Text mining: open source tokenization tools-an analysis. Advanced Computational Intelligence", An International Journal (ACII). 2016; 3(1): p. 37-47

[7]. Yllias Chali and Shafiq R Joty," Improving the performance of the random walk model for answering complex questions. In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies" Short Papers. Association for Computational Linguistics, 2008 9–12.

[8]. Yihong Gong and Xin Liu," 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. 2001 ACM, 19–25.

[9]. Yogesh Sankarasubramaniam, Krishnan Ramanathan, and Subhankar Ghosh" Text summarization using Wikipedia", Information Processing & Management 50, 3 2014, 443–461.

[10]. Ted Pedersen, Satanjeev Banerjee, and Siddharth Patwardhan," Maximizing semantic relatedness to perform word sense disambiguation", University of Minnesota supercomputing institute research report UMSI, 25,2005.

[11]. Fei Tian, Hanjun Dai, Jiang Bian, Bin Gao, Rui Zhang, Enhong Chen, and Tie-Yan Liu,"A probabilistic model for learning multiprototype word embeddings" In COLING, pages 151–160, 2014,Dublin, Ireland.

[12]. Dominic Widdows and Beate Dorow.,"A graph model for unsupervised lexical acquisition", In Proceedings of the 19th international conference on Computational linguistics, pages 1–7, Taipei, Taiwan, 2002.