Bangla Text Summarization using Word2Vec model for Recurrent Neural Network

[1] Md. Shafiqul Islam Mridul
[2] Mizanur Rahman
[3] Sabihatun Jannat
[4] Sajib Debnath
dept. of Computer Science and Engineering
United International University
Dhaka, Bangladesh
skmridul090@gmail.com
mrahman153102@bscse.uiu.ac.bd
sabiha.uiu.002@gmail.com
sajib.uiu.cse@gmail.com

Professor Dr. Mohammad Nurul Huda
dept. of Computer Science and Engineering
United International University
Dhaka, Bangladesh

***Abstract*—This paper describes text summarization which condenses the source text into a shorter version that preserves the content and general significance of the data. Text summarization techniques can be categorized as extractive and abstractive where prior one deals with choosing and concatenating significant phrases, paragraphs etc. from the initial document but the later one comprises the understanding and retelling the initial text in fewer words. It utilizes linguistic techniques to examine and interpret the text and then discover natural ideas and phrases to best describe it by creating an unprocessed shorter text that conveys the most significant data from the initial text document. From long and relevent information it is very difficult to find the text in a shorter way by setting some rules. Therefore an intelligent machine language based method is highly expected in this content. This study constructs Word2Vec model for Recurrent Neural Network (RNN) and compares the result with other rules based techniques based on Word Count and similarity matrix.**

***Index Terms*—condenses, preserves, concatenating, linguistic, categorized.**

I. Background

Many papers were published related to text summarization. In paper [1] Sarkar proposed an extraction based summarization method which is consist of four major steps: preprocessing, stemming, sentence ranking and summary generation.Sentences are ranked based on two important features: thematic term and position. In paper [2], Rahimi first considered the topic of text mining and its relationship with text summarization. Then a review has been done on some of the summarization approaches and their important parameters for extracting predominant sentences, identified the main stages of the summarizing process, and the most significant extraction criteria are presented along with the most fundamental proposed evaluation methods.In paper [3], Uddin has used various extraction methods for text summarization to form Bangla text summarizer. To build a bangla text summarizer, they used different method to rank sentences like: Location method, Cue method, title, Term Frequency, numerical data. In paper [4], Ferreira described and performed a quantitative and qualitative assessment of 15 algorithms for sentence scoring available in the literature which was evavaluted using three different dataset. They also suggested six common issues to improve sentence scoring. In paper [5] Gambhir presented a comprehensive survey of recent text summarization extractive approaches along with a few abstractive and multilingual text summarization approaches. They analysed extractive approaches like: Statistical based approaches, Topic based approaches, Graph based approaches, Discourse based approaches, Approaches based on machine learning etc. In paper [6], Chengzhang has represented words in an article as vectors trained by Word2vec,the sentence vector and the weight of each sentence are calculated by combining word-sentence relationship with graph-based ranking model, the weight of each word, the sentence vector and the weight of each sentence are calculated by combining word-sentence relationship with graph-based ranking model.They also compared summarization quality of the proposed algorithm with TF-IDF and TextRank.

II. Methodology

The proposed extractive summarization mainly comprises three independent tasks: creating an interim version of the input text, representing the sentences based on score and selecting number of sentences to form a summary.

In our experiment, we take a document as input, find the word frequency,determine the sentence similarity, weight the sentences, sort the sentences according to their rank and finally select the sentences with the higher rank for generating the summary. This approach is repeated for Word Count, Word2Vec and Similarity matrix method to measure which technique provides a much better summary.

A. Extractive text summarization

Extractive text summarization involves selecting expressions, sentences and wordings from the source document to generate a new summary. Methods include positioning the significance of expressions in order to select as it were those which is most noteworthy to the meaning of the source. This method works by identifying the important sections of the text and generating them precisely so that, they depend only on the extraction of sentences from the original text. One of the strategies to get reasonable sentences is to assign few numerical measures of a sentence which is called sentence scoring for generating the summary and after that select the higher score sentences to create record based on the compression rate. The compression rate determines the ratio between the length of the summary and the source text. The higher the compression rate, a larger summary is obtained along with an increased insignificant content. And if we decrease the compression rate, a shorter summary is obtained and valuable information is lost. The quality of the summary is acceptable when the compression rate is within 5-30%.

B. Similarity Matrix

Cosine similarity is a measure which determines how similar the documents are regardless of their size. It measures the cosine of the angle between two non-zero vectors of an inner product space which is projected in a multi-dimensional space. The cosine similarity is mostly used in positive space, where the outcome is neatly bounded in the range $[0, 1]$. When it is plotted on a multi-dimensional space, each dimension corresponds to a word in the document, the cosine similarity captures the angle of the documents and not the magnitude. Similarity matrix is a variation or a sub-part of cosine similarity.

Similarity matrix, also known as a distance matrix, allows us to understand how similar or far apart each pair of items is from the users' perspective. Similarity matrix is a table that shows the distance between pairs of sentences. Similarity matrices are mainly used as a data format when performing hierarchical clustering and multidimensional scaling. Data can be recorded in a similarity matrix at the time of collection.

The cosine of two non-zero vectors can be acquired by using the Euclidean dot product formula:

$$A.B = ||A||||B||Cos\theta$$

Given the two vectors of attributes, A and B, the cosine similarity, $Cos\theta$, is represented using a dot product and magnitude obtained is:

$$Cos\theta = \frac{A.B}{||A||||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

C. Proposed Method, Word2vec

Word2Vec represents words as vectors in an efficient and effective manner. Two methods can be used to obtain Word2Vec:

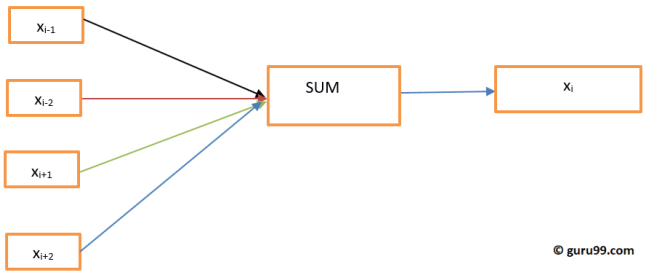


Fig. 1. Continuous Bag of Word Architecture

Skip Gram and Continuous Bag Of Words (CBOW). Skip Gram model uses the current word to forecast the adjacent window of context words. Whereas CBOW method takes the context of each word as the input and attempts to predict the word that matches the context.

Text summarization algorithm using Word2Vec includes dividing the text into sentences and dividing each sentence into words, and each word is then represented as a vector trained by Word2Vec. Weight of each word is calculated and each sentence is given a score by providing each words present in that sentence in Word2Vec model. Finally, based on the final sentence vector and the final score of the sentence, a summary is created.

Word2Vec uses CBOW method as default for building model.CBOW generally uses Bag of Word,the most common feature extraction approach for NLP. Bag of Word(BOW) looks at the histogram of word occurrences in a given corpus, without considering the order. CBOW works as follows:
[1]

Suppose V is the vocabulary size and N is the hidden layer size. Input is defined as xi-1, xi-2, xi+1, xi+2. We obtain the weight matrix by multiplying V * N.

For sentence scoring, score of each word for each sentence generated by Word2Vec model is summed up and from the summed value we generated score of each sentence using the following formula:

$$S = \alpha * STF + \beta * PV + \gamma + \lambda$$

Here, TF = summation of term frequency

D. proposed method, seq2seq Lstm model

E. proposed method, seq2seq Lstm model

F. Word Count

Word count allows us to compare documents and measure their similarities for applications such as document classification, topic modeling etc.. When a word occurs recurrent amount of times in a text the higher the score is obtained. That is, phrases which contains the most frequent words in a document are more likely to be selected for the final summary.

[1] https://www.guru99.com/images/1/111318_0826_WordEmbeddi3.png

Number of appearance of each word in the whole document is counted first using following formula:

$$number\_of\_occurrence = \sum_{i=1}^{n} word$$

Frequency of each word is then measured by dividing number of occurrence of each word by the number of occurrence of highest appeared word.

By summing up the frequency of each word present in the document, the total frequency value of a sentence is calculated.

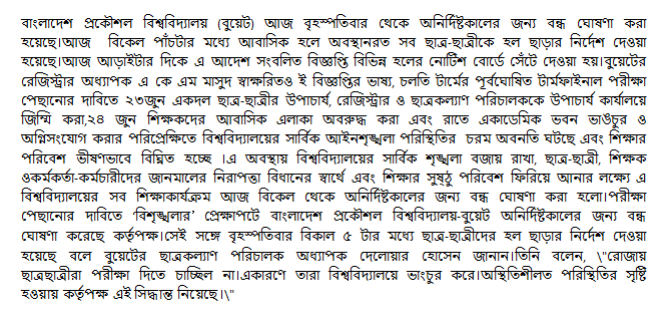
G. Summary generation

Number of sentence that will be present in summary is determined using the following formula :

$$SummaryFrequency = [\sqrt{TopRankedSentences}]$$

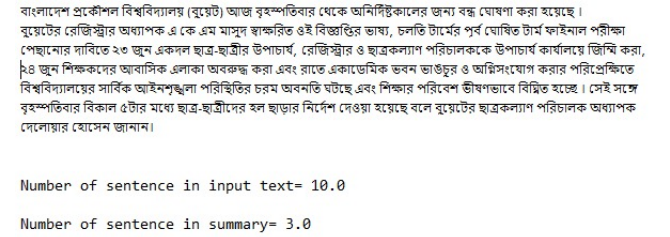
III. Result

We have built a Bangla text summarizar using three different methods to generate summary from a single document. To demonstrate these summarization method we take a sample input text as follows from a Bangla newspaper and generated three different summary using Word Count, Word2Vec and Similarity matrix. Finally we compare the result using statistical approach which consider compression rate of each method.



বাংলাদেশ প্রকৌশল বিশ্ববিদ্যালয় (বুয়েট) আজ বৃহস্পতিবার থেকে অনির্দিষ্টকালের জন্য বন্ধ ঘোষণা করা হয়েছে।আজ বিকেল পাঁচটার মধ্যে আবাসিক হলে অবস্থানরত সব ছাত্র-ছাত্রীকে হল ছাড়ার নির্দেশ দেওয়া হয়েছে।আজ আড়াইটার দিকে এ আদেশ সংবলিত বিজ্ঞপ্তি বিভিন্ন হলের নোটিশ বোর্ডে সেঁটে দেওয়া হয়।বুয়েটের রেজিস্ট্রার অধ্যাপক এ কে এম মাসুদ স্বাক্ষরিতও ই বিজ্ঞপ্তির ভাষ্য, চলতি টার্মের পূর্বঘোষিত টার্মফাইনাল পরীক্ষা পেছানোর দাবিতে ২৩জুন একদল ছাত্র-ছাত্রীর উপাচার্য, রেজিস্ট্রার ও ছাত্রকল্যাণ পরিচালককে উপাচার্য কার্যালয়ে জিম্মি করা,২৪ জুন শিক্ষকদের আবাসিক এলাকা অবরুদ্ধ করা এবং রাতে একাডেমিক ভবন ভাঙচুর ও অগ্নিসংযোগ করার পরিপ্রেক্ষিতে বিশ্ববিদ্যালয়ের সার্বিক আইনশৃঙ্খলা পরিস্থিতির চরম অবনতি ঘটছে ।এ অবস্থায় বিশ্ববিদ্যালয়ের সার্বিক শৃঙ্খলা বজায় রাখা, ছাত্র-ছাত্রী, শিক্ষক ওকর্মকর্তা-কর্মচারীদের জানমালের নিরাপত্তা বিধানের স্বার্থে এবং শিক্ষার সুষ্ঠু পরিবেশ ফিরিয়ে আনার লক্ষ্যে এ বিশ্ববিদ্যালয়ের সব শিক্ষাকার্যক্রম আজ বিকেল থেকে অনির্দিষ্টকালের জন্য বন্ধ ঘোষণা করা হলো।পরীক্ষা পেছানোর দাবিতে 'বিশৃঙ্খলার' প্রেক্ষাপটে বাংলাদেশ প্রকৌশল বিশ্ববিদ্যালয়-বুয়েট অনির্দিষ্টকালের জন্য বন্ধ ঘোষণা করেছে কর্তৃপক্ষ।সেই সঙ্গে বৃহস্পতিবার বিকাল ৫ টার মধ্যে ছাত্র-ছাত্রীদের হল ছাড়ার নির্দেশ দেওয়া হয়েছে বলে বুয়েটের ছাত্রকল্যাণ পরিচালক অধ্যাপক দেলোয়ার হোসেন জানান।তিনি বলেন, \"রোজায় ছাত্রছাত্রীরা পরীক্ষা দিতে চাচ্ছিল না।একারণে তারা বিশ্ববিদ্যালয়ে ভাংচুর করে।অস্থিতিশীলত পরিস্থিতির সৃষ্টি হওয়ায় কর্তৃপক্ষ এই সিদ্ধান্ত নিয়েছে।\"

Figure 2: Sample input text



বাংলাদেশ প্রকৌশল বিশ্ববিদ্যালয় (বুয়েট) আজ বৃহস্পতিবার থেকে অনির্দিষ্টকালের জন্য বন্ধ ঘোষণা করা হয়েছে ।
বুয়েটের রেজিস্ট্রার অধ্যাপক এ কে এম মাসুদ স্বাক্ষরিত ওই বিজ্ঞপ্তির ভাষ্য, চলতি টার্মের পূর্ব ঘোষিত টার্ম ফাইনাল পরীক্ষা পেছানোর দাবিতে ২৩ জুন একদল ছাত্র-ছাত্রীর উপাচার্য, রেজিস্ট্রার ও ছাত্রকল্যাণ পরিচালককে উপাচার্য কার্যালয়ে জিম্মি করা, ২৪ জুন শিক্ষকদের আবাসিক এলাকা অবরুদ্ধ করা এবং রাতে একাডেমিক ভবন ভাঙচুর ও অগ্নিসংযোগ করার পরিপ্রেক্ষিতে বিশ্ববিদ্যালয়ের সার্বিক আইনশৃঙ্খলা পরিস্থিতির চরম অবনতি ঘটছে এবং শিক্ষার পরিবেশ ভীষণভাবে বিঘ্নিত হচ্ছে । সেই সঙ্গে বৃহস্পতিবার বিকাল ৫টার মধ্যে ছাত্র-ছাত্রীদের হল ছাড়ার নির্দেশ দেওয়া হয়েছে বলে বুয়েটের ছাত্রকল্যাণ পরিচালক অধ্যাপক দেলোয়ার হোসেন জানান।

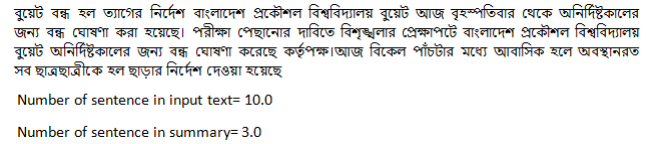
Number of sentence in input text= 10.0

Number of sentence in summary= 3.0

Figure 3: Human generated summary

IV. EVALUATION

Comparison considering compression rate of different document was done based on summary using **Word Count** as shown in the table below:

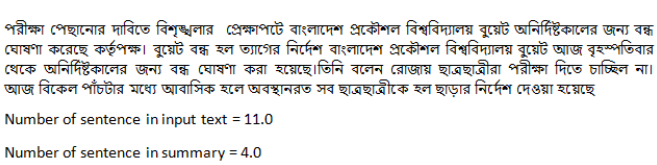


বুয়েট বন্ধ হল ত্যাগের নির্দেশ বাংলাদেশ প্রকৌশল বিশ্ববিদ্যালয় বুয়েট আজ বৃহস্পতিবার থেকে অনির্দিষ্টকালের জন্য বন্ধ ঘোষণা করা হয়েছে। পরীক্ষা পেছানোর দাবিতে বিশৃঙ্খলার প্রেক্ষাপটে বাংলাদেশ প্রকৌশল বিশ্ববিদ্যালয় বুয়েট অনির্দিষ্টকালের জন্য বন্ধ ঘোষণা করেছে কর্তৃপক্ষ।আজ বিকেল পাঁচটার মধ্যে আবাসিক হলে অবস্থানরত সব ছাত্রছাত্রীকে হল ছাড়ার নির্দেশ দেওয়া হয়েছে

Number of sentence in input text= 10.0

Number of sentence in summary= 3.0

Figure 4: Summary using Word Count

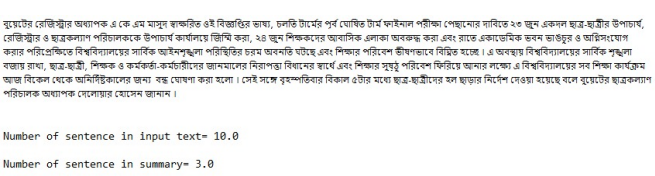


পরীক্ষা পেছানোর দাবিতে বিশৃঙ্খলার প্রেক্ষাপটে বাংলাদেশ প্রকৌশল বিশ্ববিদ্যালয় বুয়েট অনির্দিষ্টকালের জন্য বন্ধ ঘোষণা করেছে কর্তৃপক্ষ। বুয়েট বন্ধ হল ত্যাগের নির্দেশ বাংলাদেশ প্রকৌশল বিশ্ববিদ্যালয় বুয়েট আজ বৃহস্পতিবার থেকে অনির্দিষ্টকালের জন্য বন্ধ ঘোষণা করা হয়েছে।তিনি বলেন রোজায় ছাত্রছাত্রীরা পরীক্ষা দিতে চাচ্ছিল না। আজ বিকেল পাঁচটার মধ্যে আবাসিক হলে অবস্থানরত সব ছাত্রছাত্রীকে হল ছাড়ার নির্দেশ দেওয়া হয়েছে

Number of sentence in input text = 11.0

Number of sentence in summary = 4.0

Figure 5: Summary using Similarity Matrix



বুয়েটের রেজিস্ট্রার অধ্যাপক এ কে এম মাসুদ স্বাক্ষরিত ওই বিজ্ঞপ্তির ভাষ্য, চলতি টার্মের পূর্ব ঘোষিত টার্ম ফাইনাল পরীক্ষা পেছানোর দাবিতে ২৩ জুন একদল ছাত্র-ছাত্রীর উপাচার্য, রেজিস্ট্রার ও ছাত্রকল্যাণ পরিচালককে উপাচার্য কার্যালয়ে জিম্মি করা, ২৪ জুন শিক্ষকদের আবাসিক এলাকা অবরুদ্ধ করা এবং রাতে একাডেমিক ভবন ভাঙচুর ও অগ্নিসংযোগ করার পরিপ্রেক্ষিতে বিশ্ববিদ্যালয়ের সার্বিক আইনশৃঙ্খলা পরিস্থিতির চরম অবনতি ঘটছে এবং শিক্ষার পরিবেশ ভীষণভাবে বিঘ্নিত হচ্ছে । এ অবস্থায় বিশ্ববিদ্যালয়ের সার্বিক শৃঙ্খলা বজায় রাখা, ছাত্র-ছাত্রী, শিক্ষক ও কর্মকর্তা-কর্মচারীদের জানমালের নিরাপত্তা বিধানের স্বার্থে এবং শিক্ষার সুষ্ঠু পরিবেশ ফিরিয়ে আনার লক্ষ্যে এ বিশ্ববিদ্যালয়ের সব শিক্ষা কার্যক্রম আজ বিকেল থেকে অনির্দিষ্টকালের জন্য বন্ধ ঘোষণা করা হলো । সেই সঙ্গে বৃহস্পতিবার বিকাল ৫টার মধ্যে ছাত্র-ছাত্রীদের হল ছাড়ার নির্দেশ দেওয়া হয়েছে বলে বুয়েটের ছাত্রকল্যাণ পরিচালক অধ্যাপক দেলোয়ার হোসেন জানান ।

Number of sentence in input text= 10.0

Number of sentence in summary= 3.0

Figure 6: Summary using Word2Vec

TABLE I: Statistical result of summary using Word Count

No.	No. of sentences in			Compression rate by	
	original document	summary by human	summary by program	RNN	human
01	21	06	04	19.0%	28.5%
02	10	03	03	30.0%	30.0%
03	09	02	02	22.2%	22.2%
04	47	12	07	14.0%	25.5%
05	13	05	03	23.0%	38.4%

Table 1 shows the statistical result of the summary generated using Word Count by testing on 5 documents. This experiment was conducted by calculating the compression rate of the number of sentences present in the original document and the number of sentences present in the summary using Word Count. By considering the compression rate of the five documents, we can say that on average **22.0%** compression is done using Word Count.

The table shows the compression rate by considering the number of sentences present in the original document and the number of sentences present in the human generated summary which obtained an average of **28.9%**. So based on the compression rate the summary using Word Count gave a better performance.

The table below shows the comparison considering compression rate of different document based on the summary using **Similarity Matrix**:

TABLE II: Statistical result of summary using Similarity Matrix

No.	No. of sentences in			Compression rate by	
	original document	summary by human	summary by program	RNN	human
01	21	06	05	23.8%	28.5%
02	10	03	04	40.0%	30.0%
03	09	02	04	44.4%	22.2%
04	47	12	07	14.8%	25.5%
05	13	05	04	30.7%	38.4%

Table 2 shows the statistical result of the summary obtained using Similarity Matrix by experimenting on 5 documents. This experiment was conducted by calculating the compression rate of the number of sentences present in the original document and the number of sentences present in the summary using Similarity Matrix. By considering the compression rate of the five documents, we can say that on average **30.7%** compression was done using Similarity Matrix. The table shows the compression rate by considering the number of sentences present in the original document and the number of sentences present in the human generated summary which acquired an average of **28.9%**. So based on the compression rate, the summary using human generated summary gave a better performance.

The table below shows the comparison considering the compression rate of different document based on the summary using **Word2Vec**:

TABLE III: Statistical result of summary using Word2Vec

No.	No. of sentences in			Compression rate by	
	original document	summary by human	summary by program	RNN	human
01	21	06	05	23.8%	28.5%
02	10	03	03	30.0%	30.0%
03	09	02	03	33.3%	22.2%
04	47	12	07	14.8%	25.5%
05	13	05	04	30.7%	38.4%

Table 3 shows the statistical result of the summary obtained using Word2Vec by experimenting on 5 documents. This experiment was conducted by calculating the compression rate of the number of sentences present in the original document and the number of sentences present in the summary using Word2Vec. By considering the compression rate of the five documents, we can say that on average **26.5%** compression was done using Word2Vec. The table shows the compression rate by considering the number of sentences present in the original document and the number of sentences present in human the generated summary which acquired an average of **28.9%**. So based on the compression rate, the summary using human generated summary gave a better performance.

Considering above experiments, we can conclude that in case of summary generation based on compression rate, Word Count performs better than other two method.

V. CONCLUSION

We have used textual data to summarize text and considered three method Word Count, similarity matrix and proposed Word2Vec. The paper concludes the following:

- Word Count produces more compressed summary rather than other two method.
- Summary generated using Word2Vec highlights main content without skipping any information.
- We focused on producing summary more effectively using rule based models.

The authors would like to implement text summarization using CNN and seq2seq LSTM model.

VI. LIMITATION AND FUTURE WORK

We have done extractive text summarization for Bangla language which works for single document. Extractive summarization only reduces number of sentence in a text based on their rank whereas abstractive summarization uses own phases to represent the main content of any given text.We will try to improve our summarizer by transforming it to abstractive summarizaion.Also we will make a GUI summarizer system which has not built yet.Finally we will try to resolve following issues to improve sentence scoring : Morphological transformation, Similar semantics, Co-reference, Ambiguity, and Redundancy.

ACKNOWLEDGMENT

Our work is dedicated for United International University.

REFERENCES

[1] Kamal Sarkar, BENGALI TEXT SUMMARIZATION BY SENTENCE EXTRACTION, Proceedings of International Conference on Business and Information Management (ICBIM-2012), NIT Durgapur, PP 233-245, 11 Jan 2012.

[2] Shohreh Rad Rahimi, Ali Toofanzadeh Mozhdehi, Mohamad Abdolah, An Overview on Extractive Text Summarization, 2017 IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI), 26 March 2018.

[3] Md. Nizam Uddin ; Shakil Akter Khan "A study on text summarization techniques and implement few of them for Bangla language," in 10th international conference on computer and information technology 2007,27-29 Dec. 2007.

[4] R. Ferreira,L. Souza Cabral,R. Dueire Lins, G. Pereira Silva,F. Freitas, George D.C.Cavalcanti,Rinaldo Lima, Steven J.Simske, LucianoFavaro, "Assessing sentence scoring techniques for extractive text summarization," ,Expert Systems with Applications,Volume 40, Issue , Pages 5755-576414, 15 October 2013.

[5] Mahak Gambhir · Vishal Gupta, "Recent automatic text summarization techniques:a survey",Artificial Intelligence Review ,Volume 47, Issue 1, pp 1–66, January 2017.

[6] Xu Chengzhang ,Liu Dan, "Chinese Text Summarization Algorithm Based on Word2vec," Journal of Physics: Conference Series, Volume 976, conference 1.