Deep Learning Method to Estimate the Focus Time of Paragraph

Zeinab Shahbazi and Yung-Cheol Byun

***Abstract*—There are a lot of reasons that helps to create a text document that one them is time. Time is an important aspect of texts which shows the value of text and it helps searching for various topics easier. We have different type of documents that some of them has time stamp or temporal documents and some of them are without time stamp which are atemporal documents. In this paper we explain problem of paragraph focus time using convolutional neural network (CNN) and natural language processing (NLP) to show the differences between publication time and focus time of document. We implement paragraph focus time which is explains the time period that document content refers and considered to document creation time. We defined a specific time period for this process that document data is related to that and it contains the publication time and also we evaluate our method on various text documents related to historical events in defined lines.**

***Index Terms*—Paragraph focus time, deep learning, convolutional neural network, temporal information retrieval, natural language processing, word2vec.**

I. Introduction

One of the important parts of any text documents or web news is time of publishing and focus time of this documents. As we know there is a big different between this two, publication time and focus time [1]. Publication time is the time that the news or story or new topic comes out and everybody can read that news or book but focus time is the time that it happened. It shows the exact event happening time [2].

The event which publish is news or Wikipedia page or as a literature happened, in real word and it has background in special time or place and for describing those usually short texts are required [3].

The main contribution of this work is finding the differences of focus time and publication time in paragraphs of historical and temporal text documents from specific websites related to history [4].

Maximum part of text document that has related parts to historical events, have different formal functions and belief. When an event happens in history and it mentioned in literature [5], [6] it is something which occur in particular time and place in real world. Considering the fact that document quality is based on time, it should be useful for categorizing text documents by temporal focus time and mapping their content into timeline [7], [8].

We propose various kind of methods for automatically calculating focus time by using set of statistics and extracting references from document.

In this approach we divided text to paragraphs and find the relationship between unique words of each paragraph in document and using cosine similarity for computing it. We also use word2vec system for getting better results in words relationship.

This process is to find the differences between publication times of the material and document we have and in other hand the exact time of the event and comparing them.

This work is useful for finding the information easily based on the creation time and date

The main part of this paper is focusing on deep learning process and effectiveness of it and also it helps to get more effective results out of this topic and convolutional neural network is used for event and topic detection and go through the document details related to events and extract information from text document. Fig. 1 shows the survey content of an example paragraph onto timeline.

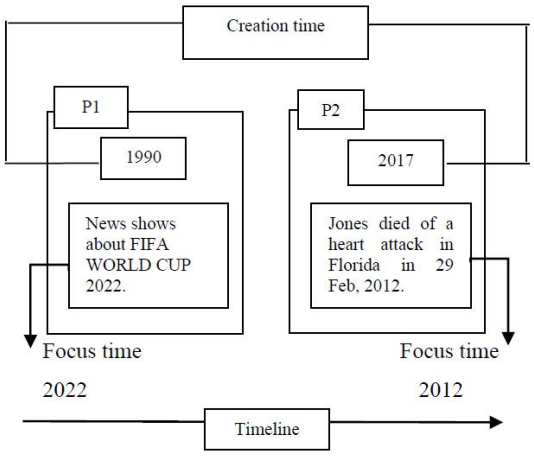


Fig. 1. Survey content of an example paragraph onto timeline.

The rest of the paper is structured as follows: in the next section we reviewed the related work. Section III shows the methodology of computing the paragraph focus time and Section IV contains the experimental results of this research and we conclude this paper in Section V.

II. Related Works

There are so many research topics related to focus time but detail of researches is very different. Some of the researchers works on document topics and some of them focus on whole document contents and other recently researchers work on

Manuscript received September 12, 2019; revised December 10, 2019.

Zeinab Shahbazi and Yung-Cheol Byun are with the Dept. of Computer Engineering, Jeju National University, Jeju, South Korea (e-mail: z.shahbazi72@gmail.com, yungcheolbyun@gmail.com).

doi: 10.18178/ijmlc.2020.10.1.901

document dating [9].

One text document has two important temporal dimensions that are creation time of text and focus time of document. Both of them shows the meaning of time temporal queries in information retrieval system [10]. By ranking the temporal expressions in text, focus time can be evaluated but it is not always a good point [11].

One of the famous works in this field is by *Jatowt et al.* which calculate the focus time of document by finding the relationship between words that extracted from news articles and historical documents which associated to time.

For retrieving a temporal document creation time of text is a significant aspect of that and in search engines most of the results are based on creation time of document but there are two problems faced on document creation time [12]:

1) Text creation time is not always available.
2) Document creation time may not represent focus time of it.

Using creation time for ranking documents can reduce information retrieval system effect for example a document creation time is 2015 but it explains about an event FIFA WORLD CUP 2022 [13].

Another problem of focus time is related to document dating in non-timestamped documents by analyzing their contents. *Kanhabua et al.* explained temporal language models for comparing without time stamp documents and time stamp documents [14].

Other category explains temporal information extraction out of document collections. Here are two examples of taggers for temporal expressions and finding dates named "GuTime" and "Stanford". Lots of complex systems can built relay on temporal information extraction [15]. For example, for finding spatio-temporal information in document strotgen and Gertz represent a system for extraction, storage and etc.

Metzler *et al.*, for identifying implicit temporal information suggest mining query logs that discuss about how many times a query pre and post qualified with a given year.

Many approaches in temporal information retrieval for ranking documents is required to estimate documents temporal aspects. What they use is extracting document timestamp or explicit temporal expression from texts.

Our work is different from other researches because in this paper we used neural network method to find the relationship between words in defined categories and we used supervised learning method in our research for better implement.

III. Methodology

The overall scheme of focus time process is presenting in Fig. 2. Document pre-processing, calculate words relationship in 2 different ways and applying results in neural network system. Whereas, the dataset results are shown in Section IV.

This process is to find the differences between publication times of the material and document we have and in other hand the exact time of the event and comparing them.

A. Document Pre-processing

We can define document classification as part of categorizing collections which basis document information. In recent years because of large amount of data it become a famous topic.

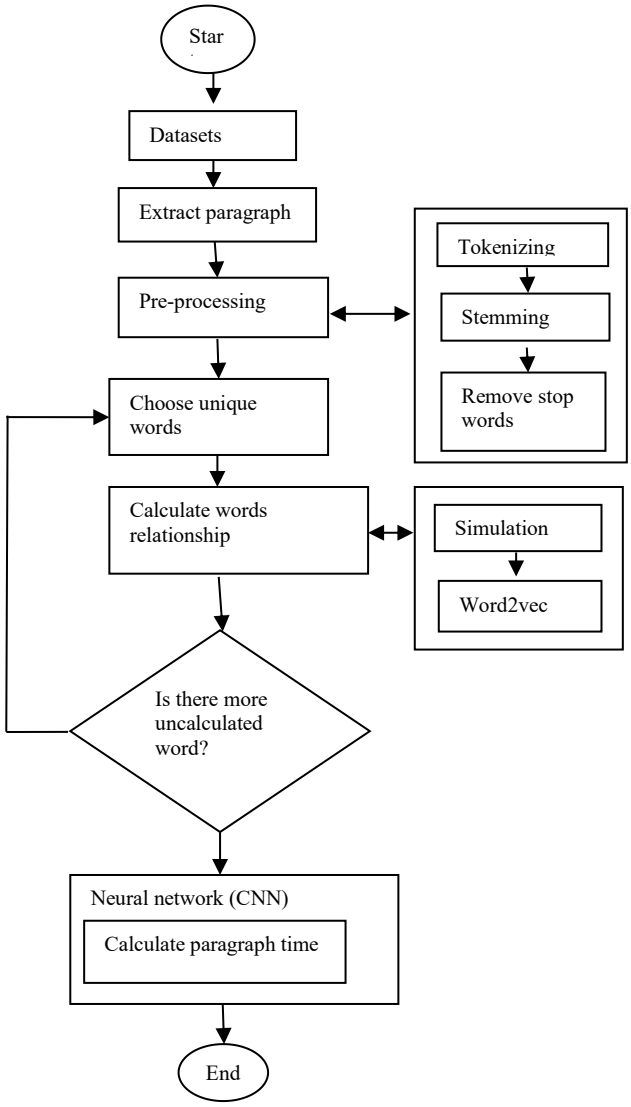


Fig. 2. Paragraph focus time algorithm.

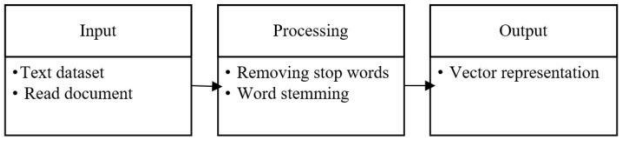


Fig. 3. Document pre-processing procedure.

Text pre-processing is famous for text tokenization. The main goal of text pre-processing is to find key words of document and detect relationship between text and words.

Document pre-processing include set of processes for some special formats in text data likewise format of numbers. Most current words include data formats that uncertainly help to text mining.

In this section after loading dataset, text pre-processing divide document into words and in this process stop words, special characters and etc. removed. For the next step text stemming decrease the stemming process for standardizing document by doing some process of roots.

B. Distributed Representation of Words

This part of process helps us to get better performance in words similarity by helping learning algorithms and it effects

to get better results in natural language processing task. The earliest use of words representation is for 1986 related to Rumelhart, Hinton and Williams [14].

Many machine learning algorithms also require the word representation and one of the most common one is bag-of-words. This feature has two weakness:

1) They lose the ordering of the words.
2) They also ignore the semantics of the words.

In this part we are going to find words relationship by using Dice's coefficient and word2vec method.

1) Evaluating word time relationship

In this method, we have two type of text categories from Wikipedia collection and web dataset from two famous websites, "History world[1]" and "Infoplease[2]" which is related to history. These documents are from specific time period "1900-2013".

$$S(X,V) = \frac{2Tz}{TX + TV} \quad (1)$$

T_Z is collection of sentences that shows the relationship between word X and word vector V.

For estimating paragraph focus time we need to find the exact time of the event and for calculating, first we need to extract unique words in each sentence by word2vec method that it helps us to know each unique words is related to which historical event.

$$S(x,t) = \frac{1}{|W|}\sum_{Y=1}^{|W|} S(y,x)^2\, S(y,t) \quad (2)$$

$S(y, x)2$ reduces faint associated of word x. We have also experimented with the method that uses non-squared $S(y, x)$ but we found it results in inferior performance. t is a focus time of words which appear together in the same sentence [15].

2) word2vec

We used this method for finding relationship between words because word2vec consist of large corps words and its process is similar to neural network but it is not a part of neural network method.

The whole process behind word2vec represent words base on its context that it means if the words become evident together they can embedded similarly too. This includes synonyms, opposites, and semantically equivalent concepts.

C. Text Creation Time

As we know not all of the documents have focus time. For example, homework report or test scores of students. The meaning of document creation time explains document publishing date which shows in that date, everybody read about that news or they upload in social media programs.

Generally, one of the challenges of this process is to achieve to exact and trustworthy timestamp of paragraphs in document. One of the dominant parts is temporal information retrieval that it is in most type of documents and actually it is a special topic in documents. Finding advantages of this propose in text processing shows that extracting information and estimating measures can clear further detection tasks.

[1] http://www.historyworld.net
[2] http://www.infoplease.com

We need text creation time for evaluating temporal expressions and this is the problem of how to find the document timestamp.

Typically, text document focus time can be useful and it has the possibility for improving information retrieval process.

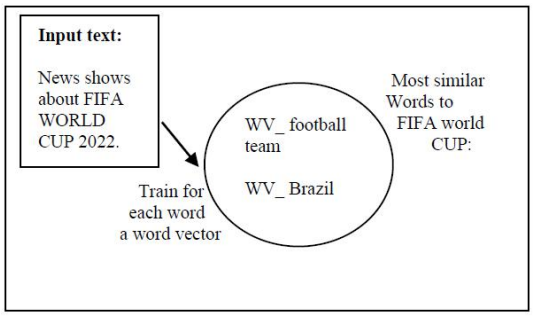


Fig. 4. Word2vec process.

IV. Experiments

For testing our algorithm we have two type of datasets: "Wikipedia dataset" and "Web dataset". First category contains 250 temporal documents which has creation time in defined time period that we mentioned before and second category contains 819 documents from web historical news of two internet websites mentioned in pervious sections.

A. Wikipedia Dataset Group

For creating set of this dataset we collected 250 articles from the English Wikipedia that are related to historical events. This events time period is from 1900 to 2013. The events are from different topics related to wars, battles, treaties, strikes, elections, and any other key events that give information about past.

B. Web Dataset Group

This dataset created with using popular websites which focused on history. We assemble 819 temporal documents form named websites: "History World[1]" and "Info please[2]". We evaluate each paragraph as a separate document and define it to take dates from the paragraph's title or if the paragraph has not any date in title we add it manually.

C. Document Focus Time

The concept of focus time is showing document content onto timeline and find the other events related to that topic and it also shows the last update of that news. In this part we use the last report start date of the event rather than its end date as the paragraph may have been updated for the last time after the start date of the event.

The time which mentioned in content of document can be shown as focus time of that text. For better explanation, time can be seen as an essential construct in human life since our thinking is often in the form of chronologically arranged events stretching from past to the present, and to the future. Each instance of time is a point-in-time value.

For starting our process using CNN in the first step we need to train our data and after test our method. The results of focus time in each table shows focus time of documents after finding relationship between words in defined dataset. Table I shows the detail result of each decade in training process by

deep learning.

TABLE I: DOCUMENT WIKI DATASET TRAINING RESULTS

Time period	Wiki data	Focus time
1900-1910	6	1901
1911-1920	9	1915
1921-1930	9	1926
1931-1940	12	1933
1941-1950	37	1949
1951-1960	96	1958
1961-1970	15	1970
1971-1980	13	1974
1981-1990	12	1986
1991-1999	28	1998
2000-2013	10	2011

TABLE II: DOCUMENT WEB DATASET TRAINING RESULTS

Time period	Wiki data	Focus time
1900-1910	96	1901
1911-1920	144	1919
1921-1930	64	1922
1931-1940	61	1935
1941-1950	106	1950
1951-1960	255	1957
1961-1970	63	1963
1971-1980	63	1980
1981-1990	50	1989
1991-1999	103	1999
2000-2013	58	2001

For starting the training process firstly, we divide the dataset into 2 parts. We used 80% of the data for the training process and 20% for the testing process. As is shown in Table I and Table II our training process in deep learning completed and we can see the focus time of whole the process after test set in Table III.

TABLE III: DOCUMENT FOCUS TIME

datasets	Total documents	Focus time	Total result
Wiki	250	1958	91%
Web	819	1957	90%

- Result Comparing with Adam Jatowt Method

After all the process we compare our results to other previous works which was about calculating focus time of document. As you can see below Table IV and V shows the result of estimating focus time of document and generic method for detecting focus time of document which written by Adam Jatowt. In these 2 papers they used temporal documents of WIKI and Web datasets for their first research they describe the concept of document focus time and they provide a range of methods for its estimation. Their approach harnesses corpus statistics and the important characteristic of their approach is that their method also works for documents that don't contain any temporal expressions or contain a few of them. The time period which they used for their process is from 1990 – 2010. Table IV shows their datasets statistics and results [15].

TABLE IV: ADAM JATOWT DATASETS STATISTICS RESULT

datasets	Total documents	Avr. sent	Focus time
Wiki	250	179	1958
Web	819	18.3	1957

In their second research which is about generic method for detecting focus time of documents they propose estimating the focus time of documents which is defined as the time period to which documents content refers and which is considered complementary dimension to the document's creation time. They test different combinations which use document terms. These approaches do not use the dates that appear in texts. Table V shows the classification result of their research process.

TABLE V: ADAM JATOWT DATASETS WIKI CLASSIFICATION RESULTS

Wiki dataset classifier	Accuracy
LDAC	0.88
Perceptron	0.86
LLSVRC	0.88

TABLE VI: ADAM JATOWT DATASETS WEB CLASSIFICATION RESULTS

Web dataset classifier	Accuracy
LDAC	0.91
Perceptron	0.89
LLSVRC	0.90

By comparing the results of 2 different researches with deep learning method we can see that deep learning increase results of document focus time up to 5% of our method.

D. Paragraph Focus Time

In this process, we calculated the document focus time and for completing our process we need to calculate the focus time of paragraph. The process of paragraph focus time is as below:

1) Extracting *paragraphs from our train data*.
2) Finding the relationship between words.
3) Apply our documents in deep learning process.
4) Test all the paragraphs that we have as our data.

TABLE VII: PARAGRAPH WIKI DATASET TEST RESULTS

Time period	Wiki data	Focus time
1900-1910	9	1909
1911-1920	18	1919
1921-1930	20	1930
1931-1940	37	1933
1941-1950	116	1947
1951-1960	296	1960
1961-1970	41	1969
1971-1980	40	1973
1981-1990	34	1987
1991-1999	85	1998
2000-2013	25	2011

TABLE VIII: PARAGRAPH WEB DATASET TEST RESULTS

Time period	Web data	Focus time
1900-1910	111	1907
1911-1920	294	1920
1921-1930	181	1925
1931-1940	182	1938
1941-1950	260	1942
1951-1960	350	1959
1961-1970	185	1966
1971-1980	216	1978
1981-1990	272	1990
1991-1999	192	1995
2000-2013	118	2012

TABLE IX: PARAGRAPH FOCUS TIME

datasets	Total documents	Focus time	Total result
Wiki	250	1960	96%
Web	819	1959	94%

The reason that we start to find focus time of paragraph is, paragraphs shows the detail of each document. For example, if we have a document which, related to timeline 1910, after dividing text to paragraph and estimate each paragraph and find time of them we can see how many percent of the text related to that time period and it shows that our text focus time is true or false.

E. *Human Being*

In this section we care going to do our process by human being. After doing all the focus time method by deep learning and finding the relationship between words using word2vec and comparing our results with other methods and other authors result we are going to check our paragraph data's by 3 different people and 3 different ideas to find out if the process works well and benefits of the deep learning method. In Table X we show the result of human being with 3 different people.

TABLE X: HUMAN BEING RESULTS

categories	True			False		
1900-1910	95%	98%	98%	5%	2%	2%
1911-1920	96%	97%	95%	4%	3%	5%
1921-1930	97%	98%	97%	3%	2%	3%
1931-1940	95%	98%	98%	5%	2%	2%
1941-1950	97%	99%	99%	3%	1%	1%
1951-1960	97%	98%	98%	3%	2%	2%
1961-1970	95%	98%	96%	5%	2%	4%
1971-1980	97%	97%	97%	3%	3%	3%
1981-1990	96%	97%	97%	4%	3%	3%
1991-1999	98%	99%	96%	2%	1%	4%
2000-2013	97%	98%	97%	3%	2%	3%

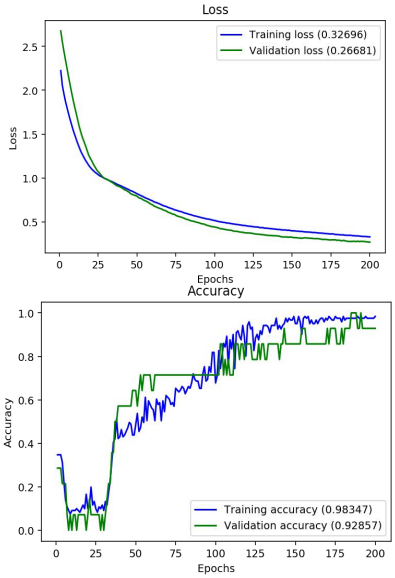


Fig. 5. Document focus time Wiki dataset loss and accuracy.

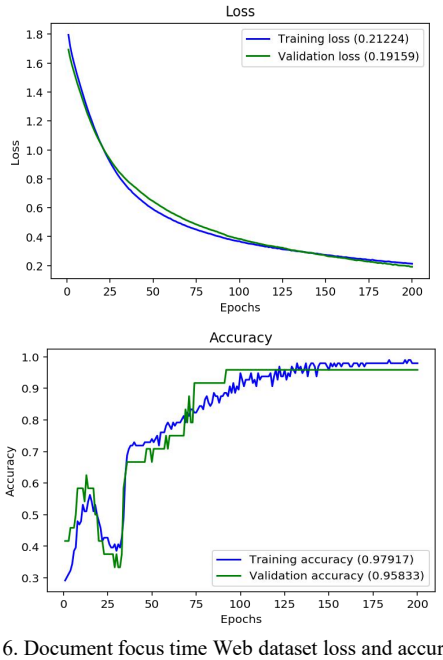


Fig. 6. Document focus time Web dataset loss and accuracy.

In total we use 2 datasets (Wiki and Web). All the Wikipedia articles were processed by using CLIPS library and subject to processing in order to extract core content and segment it into paragraphs. The Figures below shows document focus time loss and accuracy results.

In paragraph focus time total datasets that we have is 2 datasets (Wiki and Web). Figures below show the result of loss and accuracy in paragraph focus time.

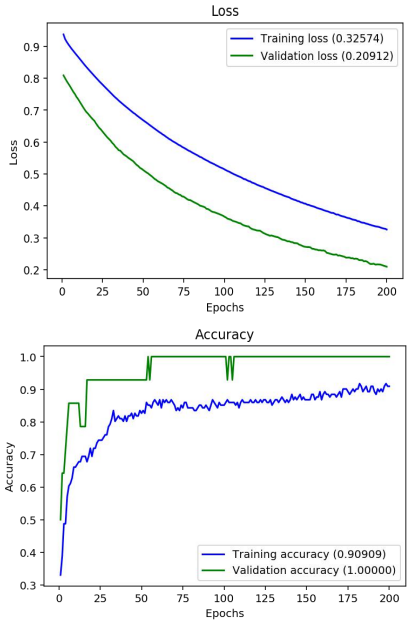


Fig. 7. Paragraph focus time Wiki dataset loss and accuracy.

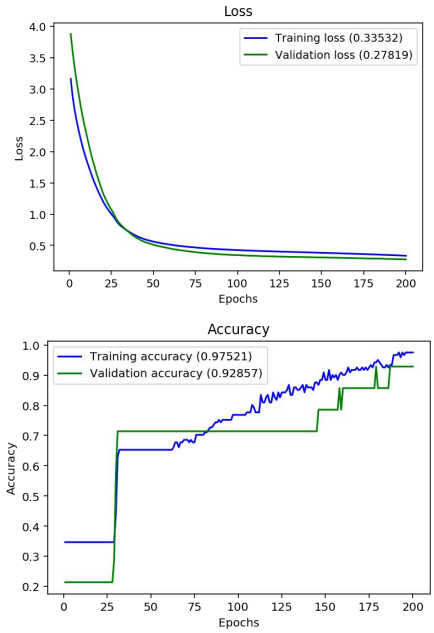


Fig. 8. Paragraph focus time Web dataset loss and accuracy.

F. *Precision and Recall*

In this section we are presenting the output of final result in two different dataset which shows the comparison between document and paragraph focus time. Table XI and XII shows the final result of proposed system.

TABLE XI: DOCUMENT FOCUS TIME F-MEASURE

Document					
Wiki			Web		
Precision	Recall	F1	Precision	Recall	F1
0.94	0.92	0.92	0.96	0.95	0.95

TABLE XII: PARAGRAPH FOCUS TIME F-MEASURE

Paragraph					
Wiki			Web		
Precision	Recall	F1	Precision	Recall	F1
0.86	0.86	0.86	094	0.92	0.92

V. CONCLUSION

Time is a significant aspect of text documents that help to value of text and real historical events. Time matters greatly in our lives. It is also an important aspect of texts. We think that properly estimating the content time of temporal documents should improve temporal information retrieval and strengthen our means of analyzing and understanding documents and temporal references in texts. The main contribution of this work is calculating focus time of paragraph by applying deep learning method and using CNN for this process. In this paper, we describe the concept of focus time and our approach to deep learning especially uses absolute references to past years in a news article. The intriguing characteristic of our proposal is that it also works for documents which do not contain any temporal expressions. Besides estimating document focus time we also demonstrate the classification approach for detecting temporal documents. The intriguing characteristic of our proposal is that it also works for documents which do not contain any temporal expressions.

ACKNOWLEDGMENT

This work was supported by the 2019 education, research and student guidance grant funded by Jeju National University.

CONFLICT OF INTEREST

The authors whose names are listed immediately below certify that they have NO affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

AUTHOR CONTRIBUTION

Zeinab Shahbazi and Yungcheol Byun designed the model and the computational framework and analyzed the data. They carried out the implementation and performed the calculations. They wrote the manuscript with input from all authors and conceived the study and were in charge of overall direction and planning.

REFERENCES

[1] A. Kaushik and S. Naithani, "A comprehensive study of text mining approach," *IJCSNS International Journal of Computer Science and Network Security,* vol. 16, no. 2, February 2016.

[2] R. J. Mooney and U. Y. Nahm, "Text mining with information extraction," in *Proc. the 4th International MIDP Colloquium on Multilingualism and Electronic Language Management*, September 2003, Bloemfontein, South Africa, pp. 141-160.

[3] S. Vijayarani, J. Ilamathi, and Nithya, "Preprocessing techniques for text mining — An overview," *International Journal of Computer Science & Communication Networks*, vol. 5, no. 1, pp. 7-16.

[4] A. Kloptchenko, T. Eklund, J. Karlsson, B. Back, H. Vanharanta, and A. visa, "Combining data and text mining techniques for analyzing financial reports" *Intelligent System in Accounting, Finance and Management, Intell. Sys. Acc. Fin. Mgmt,* vol. 12, pp. 29-41, 2004.

[5] R. Talib, M. K. Hanif, S. Ayesha, and F. Fatima, "Text mining: Techniques, applications and issues," *International Journal of Advanced computer Science and Applications,* vol. 7, no. 11, 2016.

[6] K. Khelif, R. Dieng-Kuntz, and P. Barbry, "An ontology-based approach to support text mining and information retrieval in biological domain," *Journal of Universal Computer Science*, vol. 13, no. 12, pp. 1881-1907.2007.

[7] M. Krallinger and A. Valencia, "Text-mining and information-retrieval services for molecular biology – Review," *Genome Biology*, vol. 6, no. 224, 2005.

[8] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning review," *Nature*, vol. 521, 2015.

[9] B. Mitra and N. Crawell, Neural models for information retrieval," arXiv preprint arXiv:1705.01509， 2017.

[10] I. Syu, S. D. Lang, and N. Deo, "A neural network model for information retrieval using latent semantic indexing," *0-7803-3210-5/96 $4.00©1996 IEEE.*

[11] H. Li and Z. Lu, "Deep learning for information retrieval," *SIGIR' 16*, Pisa, Italy, July 17-21, 2016.

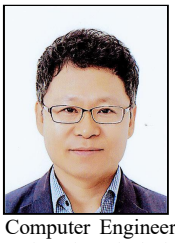
[12] T. Schnabel, I. Labutov, and D. Mimno, *Evaluation Methods for Unsupervised Word Embeddings*, Lisbon, Portugal, September 17-21, 2015.

[13] I. Mokris and L. Skovajsova, *Development of Neural Network Information Retrieval System from Text Documents.*

[14] T. Mikolov, I. Sutskever, and K. Chen, *Distributed Representations of Words an Phrases and Their Compositionality.*

[15] A. Jatowt, C. A. Yeung, and K. Tanaka, *Estimating Document Focus Time.*

Zeinab Shahbazi received her B.S. in software engineering from Pooyesh University, IRAN. In March 2017, she moved to Republic of Korea for M.S studies and started working in internet laboratory, Chonbuk National University (CBNU). After completing her master in 2018, she moved to Jeju-do in March 2019 and started working as a Ph.D. research fellow in Machine Learning Laboratory (MLL), Jeju National University. Her research interests include artificial intelligence and machine learning, natural language processing, deep learning and data mining.



Yung Cheol Byun received his B.S. from Jeju National University, Korea in 1993, M.S and Ph.D degrees from Yonsei University in 1995 and 2001. He worked as a special lecturer in Samsung Electronics in 2000 and 2001. From 2001 to 2003, he was a senior researcher of Electronics and Telecommunications Research Institute and he promoted to join Jeju National University as an assistant professor in 2003, where he is currently a professor in the Department of Computer Engineering. From 2012 to 2014, he had research activities at University of Florida as a visiting professor. His research interests include the areas of pattern recognition & image processing, artificial intelligence & machine learning, security based on pattern recognition, home network and ubiquitous computing, u-healthcare and RFID & IoT middleware system.