Artificial Intelligence for Automatic Text Summarization

Min-Yuh Day
Department of Information Management
Tamkang University
New Taipei City, Taiwan
myday@mail.tku.edu.tw

Chao-Yu Chen
Department of Information Management
Tamkang University
New Taipei City, Taiwan
susan.cy.chen@gmail.tw

Abstract—**Automatic text summarization has played a critical role in helping people obtain key information from increasing huge data with the advantaged development of technology. In the past, few literatures are related to solve the problem of generating titles (short summaries) by using artificial intelligence (AI). The purpose of this study is that we proposed an AI approach for automatic text summarization. We developed an AI text summarization system architecture with three models, namely, statistical model, machine learning model, and deep learning model as well as evaluating the performance of three models. Essay titles and essay abstracts are used to train artificial intelligence deep learning model to generate the candidate titles and evaluated by ROUGE for performance evaluation. The contribution of this paper is that we proposed an AI automatic text summarization system by applying deep learning to generate short summaries from the titles and abstracts of the Web of Science (WOS) database.**

Keywords: Artificial Intelligence, Sequence-to-Sequence, Automatic Text Summarization, Long Short-Term Memory, Recurrent Neural Network

I. INTRODUCTION

Due to the era of change and technology development, the information people receive has been increasing enormously. Automatic text summarization has played a pivotal role in helping people obtain key information. For this issue, theories and technology are decided by text type, the approach of generating summaries, the target summary generation and the support of software and hardware. The common technologies of text summarization include statistical method, graph theory, and machine learning [5][16].

In recent years, more and more researchers have used deep learning to solve machine translation problems. However, few literatures are related to solve the problem of generating titles (short summaries) by using deep learning technology in the past. Therefore, this study focuses on evaluating the performance of applying deep learning in short summary generation.

The main objectives of this study include:

1. Use artificial intelligence technologies, which are including statistical method, machine learning ,and deep learning, to generate candidate titles, and compare the accuracy.
2. Compare the accuracy of different deep learning models.

The remainder of this paper is organized as follows: Selection 2 describes the literature on text summarization and deep learning. Section 3 shows the methodology. Section 4 shows the experimental results and discussion. Finally, Section 5 presents conclusions.

II. RELATED WORK

A. Text Summarization

The definition of a summary is a text which includes one or more words and these words represent important information in the raw text and shorter than the raw text obviously [20]. Table 1 presents the main different types of text summarization by Gambhir and Gupta [29]. The first type of summaries, amount of input document, is generated by single or multi-document. The second one has two kinds which is extraction method and abstract method. Extraction summarization is that extracts keywords and paragraphs to generate summaries. Abstraction summarization is that generates summaries by creating new texts. The third one is divided into two kinds, generic and topic-oriented summarization. Generic summarization reflects the opinion of authors, while topic-oriented summarization is related to the topics which readers are interested in.

The earliest study of text summarization started from 1958, which included the frequency of a particular word [13], the sentence position [1], and keywords [6]. Das et al. [5] and Nenkova et al. [16] reviewed the technology of different types of text summarization and main methodology. Early single-document summarization generated by extraction method which used the features to

TABLE 1. TYPE OF AUTOMATIC TEXT SUMMARIZATION

Type	Kind of Automatic Text Summarization
Amount of input documents	Single-document Summarization
	Multi-document Summarization
Generation Method of Abstract	Extraction Summarization
	Abstraction Summarization
Requirement of Users	Generic Summarization
	Topic-oriented Summarization

978-1-5386-2659-7/18/$31.00 ©2018 IEEE
DOI 10.1109/IRI.2018.00076

evaluate the extracted words and paragraphs. The common technology included Naïve-Bayes, hidden Markov models, log-linear models, machine learning. The advanced development of multi-document summarization started at 1990s and most dataset were news articles. The common technologies included graph theory, sentence clustering and domain-dependent topics. The main methodologies of text summarization include topic representation and indicator representation respectively. The former one generates topic word and TF*IDF at first, evaluates the topic of input documents, scoring sentences, and converts into summaries. The latter one uses graph theory and machine learning to evaluate each sentence, and then choose the best set of sentences in a greedy approach to generate summaries.

Lin proposed an evaluation approach called ROUGE [12]. ROUGE includes ROUGE-N, ROUGE-L, ROUGE-W, ROUGE-S, and ROUGE-SU. ROUGE-N calculates the score by an n-gram recall between a candidate summary and a set of reference summaries. ROUGE-L uses Longest Common Subsequence (LCS) to calculate the similarity score. The advantages of using LCS are requiring non-consecutive matches in sentences and including longest in-sentence common n-grams automatically. ROUGE-W is used to improve the disadvantage of ROUGE-L. ROUGE-S uses skip-bigram to generate pairs of words in their sentence order, allowing for arbitrary gaps and calculate the percentage of matched pairs between candidate summary and reference summaries as similarity. ROUGE-SU combines skip-bigram and unigram to solve the potential problem for ROUGE-S that it does not give any credit to a candidate sentence if the sentence does not have any word pair co-occurring with its references.

B. *Artificial Intelligence*

The earliest definition of Artificial Intelligence (AI) was proposed by John McCarthy in 1955. AI was the science and engineering of making intelligent machines [14]. Russell et al. [22] proposed that there were eight definitions of AI, laid out along two dimensions. The definitions on the top are concerned with thought processes and reasoning, whereas the ones on the bottom is considered of behavior. The ones on the left focuses on measure success in terms of fidelity to human performance, whereas the ones on the right measure rationality. Table 2 shows four approaches of AI. The Turing Test, proposed by Alan Turing (1950), was designed to provide a satisfactory operational definition of intelligence. A computer would need to have the following capabilities to pass the test. The capabilities are natural language processing (NLP), knowledge representation, automated reasoning, and machine learning (ML). Test summarization, the topic of this study, is a text mining problem which is a branch of NLP. We apply machine learning and deep learning to achieve our objectives. The relationship between AI, ML and deep learning (DL) is shown in Table 1. ML is a branch of the field of AI, and DL is one of methods of ML.

TABLE 2. FOUR APPROACHES OF AI

Thinking Humanly	**Thinking Rationally**
The cognitive modeling approach	The "laws of though" approach
Acting Humanly	**Acting Rationally**
Turing Test (Alan Turing, 1950)	The rational agent approach

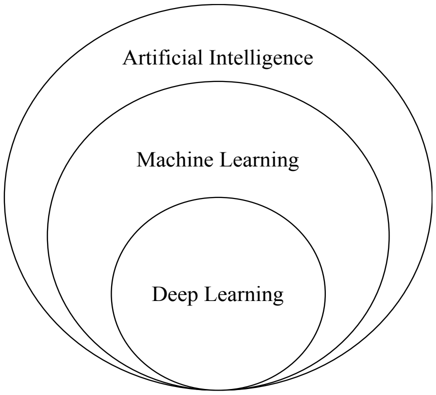


Figure 1. The relationship between AI, ML and DL

C. *Neural Network and Deep Learning*

Original concept of neural network was proposed by McCulloch and Pitts in 1948 [15], which is a parallel operation approach [23] and learned by samples [24]. Rumelhart et al. [21] proposed back-propagation that calculates a gradient and resets the weight to minorize the loss. Accuracy depends on the volume of learning data, and neural network with back-propagation is applied to variety of domains in recent years.

Traditional approach has some problems and disadvantages, such as longer training time, overfit [28], and learning capability limited by the number of hidden layer [26]. Deep learning is a branch of machine learning, and it is looked toward to do multi-linear calculation and automatically generate features with huge data and more than one hidden layer. The concept of deep learning has been existed from 1990s. The efficacy of computers and hardware was not enough good to run deep learning models, so support vector machine (SVM) has been the main technology. In recent years, deep learning has become the trend due to the advanced development of technology.

D. *Recurrent Neural Network*

After recurrent neural network (RNN) was proposed [8], it has been applied to NLP, visual recognition, and voice recognition. The performance of applying strong

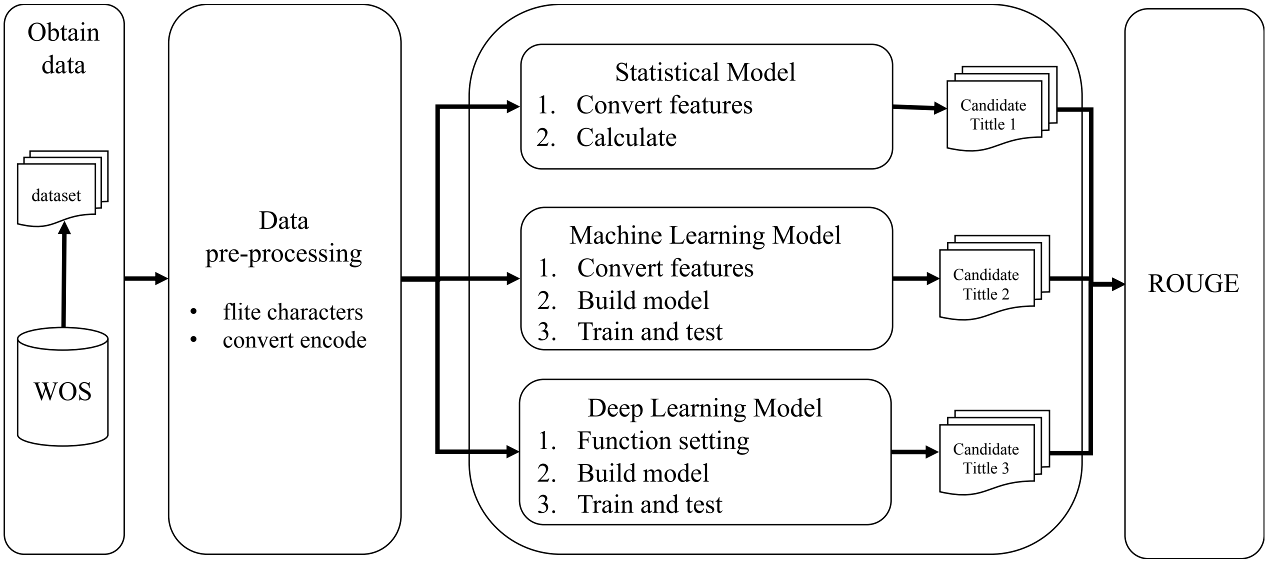


Figure 2. System architecture

dynamic RNN in the prediction of sequential data, language analysis, and word recognition is satisfied [11].

Long short-term memory (LSTM) was proposed to solve gradient vanishing or exploding occurred to long-time RNN. Evaluates data to be reset and memorized with three gates [10]. Cho et al. [2] proposed the model of Gated recurrent unit (GRU) that uses two gates instead of three gates.

Although the performance of DNN for voice recognition and machine translation is satisfied, it still has the limitation of the same length and dimensions of input and output. Chung et al. [4] and Sutskever et al. [25] proposed sequence-to-sequence (seq2seq) model. The central concept is using two LSTM as an encoder and a decoder. The former one converts the sequence into a huge vector at one time, whereas the latter one converts the prediction vector of this huge vector into the sequence as output.

III. METHODOLOGY

In this study, we use "Systems Development Research Methodology" from information system research field as our research methodology [17].

A. *System Architecture*

Figure 2 shows the system architecture of Deep Learning for Automatic Text Summarization. The system architecture of this study is shown as below:

- Obtain data: Raw data was obtained from Web of Science Core Collection database.
- Data pre-processing: This stage includes two steps that are fliting special characters and convert encode.
- Model development: We put data in three different models to generate three kinds of candidate titles.
- Evaluation: We use ROUGE as evaluation method to give the score of candidate titles.

B. *Dataset*

This study used 9 keyword indicators as research words in WOS database and obtained 50,387 essays between 1970 and 2017 as a raw dataset. These keyword indicators which are represented "Sentiment Analysis" from 1913 to 2016 [19] are as follows:

a) Sentiment Analysis
b) Sentiment Classification
c) Opinion Mining
d) Opinion Classification
e) Effective Computing
f) Sentiwordnet
g) Sentic
h) Mining sentiment
i) Mining sentiments

Essay titles and essay abstracts were extracted, flited some special characters, convert encode and converted into the format of "title-abstract" pair.

Figure 3 provides an example of generating candidate titles. For instance, we need to extract the essay title and the essay abstract from a journal paper published at MIS Quarterly in 2016. This paper is one of the raw dataset. The second step is putting these texts into three models to generate candidate titles. Calculate the similarity between the extracted essay title and candidate titles by ROUGE. Finally, select one candidate title with the highest score as the best title.

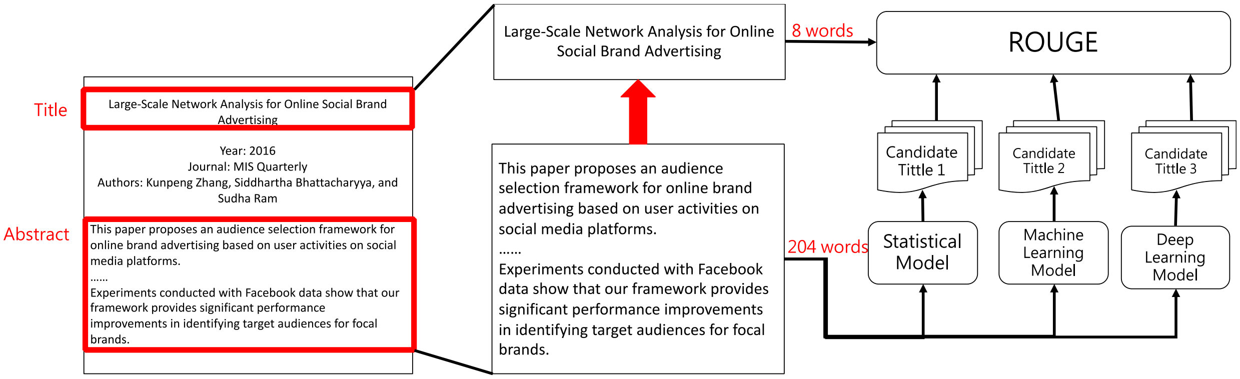
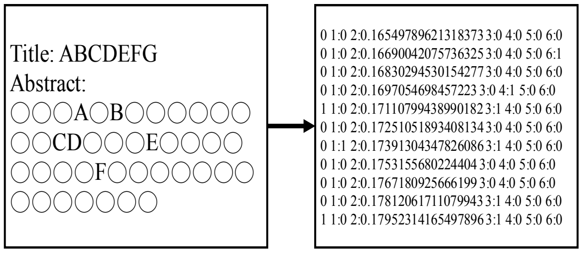

Title
Large-Scale Network Analysis for Online Social Brand Advertising
Year: 2016
Journal: MIS Quarterly
Authors: Kunpeng Zhang, Siddhartha Bhattacharyya, and Sudha Ram
Abstract
This paper proposes an audience selection framework for online brand advertising based on user activities on social media platforms.
......
Experiments conducted with Facebook data show that our framework provides significant performance improvements in identifying target audiences for focal brands.
Large-Scale Network Analysis for Online Social Brand Advertising
8 words
ROUGE
Candidate Tittle 1
Candidate Tittle 2
Candidate Tittle 3
204 words
Statistical Model
Machine Learning Model
Deep Learning Model

Figure 3. Summary example

C. *Feature Processing*

We extracted six features from abstracts as the input of machine learning model. Table 3 presents six features. F01 and F02 are Named Entity Recognition (NER), position. Other features represent that the token is the special part of speech or not.

In this study, we use LibSVM [30] to train and predict that the token is same as the part of candidate title or not. Figure 4 shows the processing of tokens converting into the special format for LibSVM. In figure 4, there are some matched tokens located in different part of the abstract. If the token is same as one token of the title, the class label is set as 1. Other tokens are not same, so the class labels are set as 0.



```
0 1:0 2:0.1654978962131837 3:0 4:0 5:0 6:0
0 1:0 2:0.1669004207573632 5 3:0 4:0 5:0 6:1
0 1:0 2:0.1683029453015427 7 3:0 4:0 5:0 6:0
0 1:0 2:0.169705469845722 3 3:0 4:1 5:0 6:0
1 1:0 2:0.1711079943899018 2 3:1 4:0 5:0 6:0
0 1:0 2:0.1725105189340813 4 3:0 4:0 5:0 6:0
0 1:1 2:0.1739130434782608 6 3:1 4:0 5:0 6:0
0 1:0 2:0.175315568022440 4 3:0 4:0 5:0 6:0
0 1:0 2:0.176718092566619 9 3:0 4:0 5:0 6:0
0 1:0 2:0.1781206171107994 3 3:1 4:0 5:0 6:0
1 1:0 2:0.1795231416549789 6 3:1 4:0 5:0 6:0
```

Figure 4. The processing of tokens converting into the special format

TABLE 3. TABLE OF FEATURE

Feature ID	Feature Name
F01	NER
F02	Position
F03	cntPOSNoun
F04	cntPOSVerb
F05	cntPOSAdj
F06	cntPOSAdv

D. *Parameter Setting*

In order to increase accuracy, we set some parameters.

TABLE 4. PARAMETERS OF DEEP LEARNING MODEL

Parameter	Value
Dropout	0.5
Loss Function	categorical_crossentropy
Optimizer	rmsprop

Table 4 shows how we set these parameters. Dropout is the percentage of dropping out neurons in each hidden layer for providing overfit. Hinton et al. [8] proposed that 0.5 is better. Loss function calculates the bias between current answers and candidate answers. The bias is smaller, and the accuracy is higher. According to different kinds of data and format of predicted results, we need to choose different loss function. In this study, we have to convert essay titles and essay abstracts into the one-hot format, so we choose categorical_crossentropy. Optimizer is using different algorithms to raise accuracy. We refer the sample code from Keras and use rmsprop as the optimizer.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

A. *Experiment Design*

According to the system architecture, we use ROUGE as evaluation method and put pre-processed data in three models. Select fixed number of candidate titles from three kinds of candidate titles, and find the matched standard essay titles from dataset as reference titles. Use ROUGE-1, ROUGE-2, ROGUE-L, and ROUGE-SU to calculate the avenged similarity score as the performance of each model.

TABLE 5. RESULTS OF STATISTCAL MODEL

KS Results (Keep Stop words)									
	ROUGE-1			ROUGE-2			ROUGE-L		
	F	P	R	F	P	R	F	P	R
count	50174	50174	50174	50174	50174	50174	50174	50174	50174
mean	**0.25**	**0.24**	**0.27**	0.01	0.01	0.01	0.16	0.16	0.17
std	0.12	0.11	0.13	0.02	0.02	0.03	0.08	0.07	0.08
min	0	0	0	0	0	0	0	0	0
25%	0.17	0.17	0.18	0	0	0	0.12	0.11	0.13
50%	0.26	0.24	0.27	0	0	0	0.16	0.15	0.17
75%	0.33	0.31	0.36	0	0	0	0.21	0.20	0.22
max	0.80	0.77	0.83	0.44	0.40	0.50	0.61	0.57	0.67
RS Results (Remove Stop words)									
	ROUGE-1			ROUGE-2			ROUGE-L		
	F	P	R	F	P	R	F	P	R
count	50174	50174	50174	50174	50174	50174	50174	50174	50174
mean	0.17	0.16	0.18	0.01	0.01	0.01	0.12	0.11	0.12
std	0.11	0.10	0.12	0.03	0.02	0.03	0.07	0.07	0.08
min	0	0	0	0	0	0	0	0	0
25%	0.09	0.08	0.09	0	0	0	0.07	0.07	0.07
50%	0.16	0.15	0.17	0	0	0	0.11	0.11	0.11
75%	0.24	0.22	0.25	0	0	0	0.16	0.15	0.17
max	**0.86**	**0.89**	**1.00**	**0.80**	**0.67**	**1.00**	**0.82**	**0.75**	**1.00**
F: F-score, P: Precision, R: Recall									

B. *Results of Statistcal Model*

TF*IDF is used as the main method for statistical model. Figure 5 shows the processing of this model. After tokenized the abstracts, raw dataset is divided into two sub datasets, keeping stop words (KS) and removing stop words (RS). According to standard titles, extract the fixed number of tokens which the TF*IDF score are top in the whole abstract. This model generates two results. Table 5 presents the details of these two results evaluated by ROUGE. In Table 5, it is clear that the performance of KS results evaluated by ROUGE-1 is the best. However, the performance of RS results shows that this method can generate the higher similarity for one case.

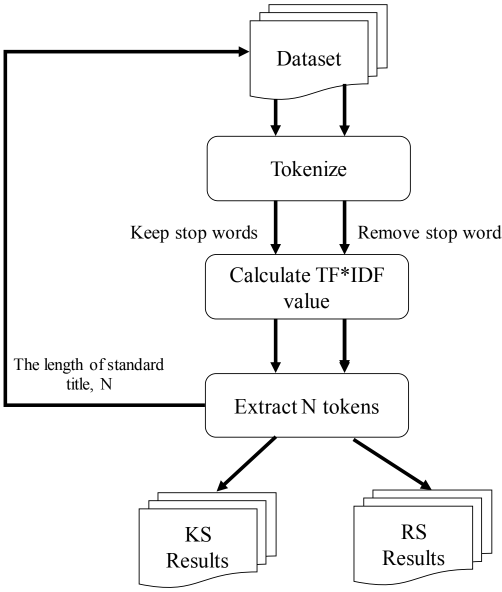


Figure 5. The processing of statistical model

C. *Results of Machine Learning Model*

We divide 1000 randomly abstracts into the training part (80%) and testing part (20%). These two parts convert into 185132 and 46322 vectors respectively. After these training vectors trained 34179 iterations, the trained model has 66574 support vectors. Put testing vectors into this trained model, and then get the accuracy which is 82.47%. The table 6 shows the predictive result of this model. This model predicts 38204 correct vectors and 8118 non-correct vectors.

TABLE 6. PREDICTIVE RESULT OF MACHINE LEARNING MODEL

Accuracy	**82.47%**
# of Correct	38204
# of Non-correct	8118
Total	46322

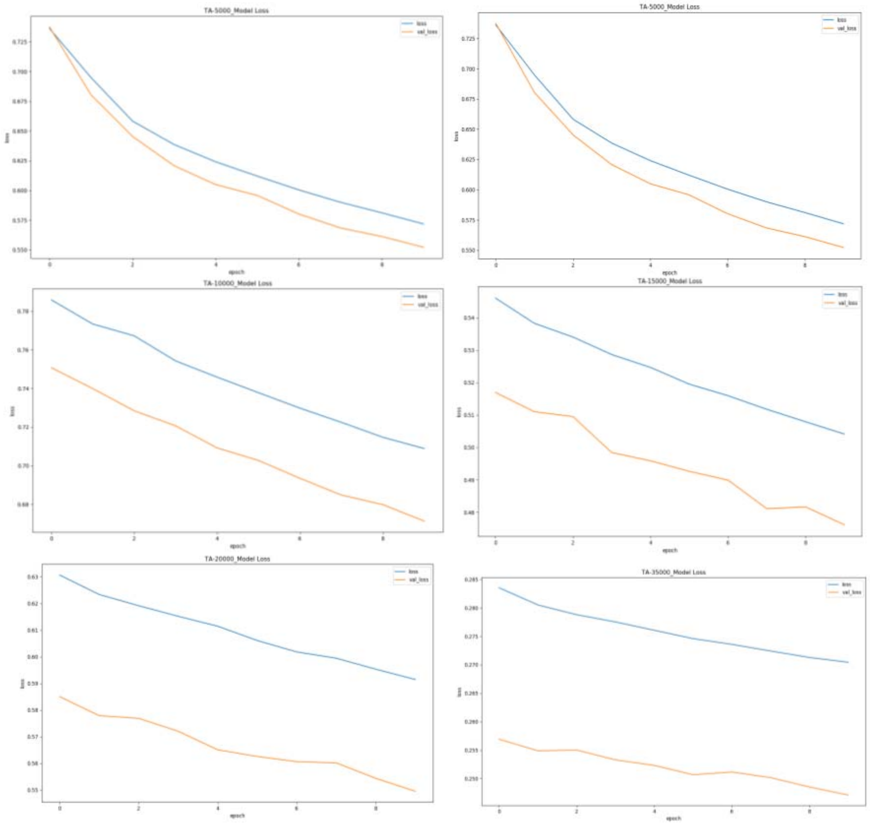


Figure 6. Loss and val-loss of 30,000 essay and 10 epochs

D. *Results of Deep Learning Model*

We divided the dataset into 10 sub datasets, and set input layer as character level. Used seq2seq model built by Keras to train, batch size was 64, and epoch was 10. Figure 6 shows the value of loss and val-loss of six sub datasets. Observe the model that is overfit or not through the change of the value of loss and val-loss. At first, both values decreased from 0.72. Finally, both values remained at 0.2. This seq2seq model is not convergence.

V. Conclusions

The best performance of statistical model is the ROUGE-1 result of KS results, which is 0.25 meanly. However, the three results of RS results have the higher similarity in one case. Machine learning model uses 1000 randomly abstracts to train and test. The trained model gets the accuracy which is 82.47% to predict the token is the part of the candidate title or not. The result of deep learning model shows that both values decreased from 0.72 to 0.2 approximately. The seq2seq model was not convergence, we could try to change input layer as word level and set more parameters.

The contribution of this paper is applying deep learning to generate short summaries, comparing with different methods, and the training and testing of automated generation of English essay titles and abstracts from 1970 to 2017.

For the research in the future, we propose the following points:

- Extend the domains of essay and choose international journals that have higher rankings.
- Set more parameters and algorithms, such as attention mechanism.

Limitation of the Study is decreasing the fluency of candidate titles and then evaluate with correct titles.

VI. Acknowledgement

This research was supported in part by Tamkang University (TKU) research grant.

VII. References

[1] Baxendale, P. B. (1958). Machine-made index for technical literature—an experiment. IBM Journal of Research and Development, 2(4), 354-361.

[2] Cho, K., Van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:1409.1259.

[3] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.

[4] Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555.

[5] Das, D., & Martins, A. F. (2007). A survey on automatic text summarization. Literature Survey for the Language and Statistics II course at CMU, 4, 192-195.

[6] Edmundson, H. P. (1969). New methods in automatic extracting. Journal of the ACM (JACM), 16(2), 264-285.

[7] Graves, A., Mohamed, A.-r., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. Paper presented at the Acoustics, speech and signal processing (icassp), 2013 ieee international conference on.

[8] Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. science, 313(5786), 504-507.

[9] Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580.

[10] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

[11] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444. doi:10.1038/nature14539

[12] Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. Paper presented at the Text summarization branches out: Proceedings of the ACL-04 workshop.

[13] Luhn, H. P. (1958). The automatic creation of literature abstracts. IBM Journal of research and development, 2(2), 159-165.

[14] McCarthy, J. (1998). What is artificial intelligence?

[15] McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics, 5(4), 115-133.

[16] Nenkova, A., & McKeown, K. (2012). A survey of text summarization techniques. Mining text data, 43-76.

[17] Nunamaker Jr, J. F., Chen, M., & Purdin, T. D. (1990). Systems development in information systems research. Journal of management information systems, 7(3), 89-106.

[18] Olah, C. (2015). Understanding lstm networks. GITHUB blog, posted on August, 27, 2015.

[19] Piryani, R., Madhavi, D., & Singh, V. K. (2017). Analytical mapping of opinion mining and sentiment analysis research during 2000–2015. Information Processing & Management, 53(1), 122-150.

[20] Radev, D. R., Hovy, E., & McKeown, K. (2002). Introduction to the special issue on summarization. Computational linguistics, 28(4), 399-408.

[21] Rumelhart, D., Hinton, G., & Williams, R. (1986). Learning internal representation by back propagation. Parallel distributed processing: exploration in the microstructure of cognition, 1.

[22] Russell, S. J., & Norvig, P. (2010). Artificial intelligence: a modern approach (3 ed.): Prentice hall Upper Saddle River.

[23] Sondak, N. E., & Sondak, V. K. (1989). Neural networks and artificial intelligence. Paper presented at the ACM SIGCSE Bulletin.

[24] Specht, D. F. (1991). A general regression neural network. IEEE transactions on neural networks, 2(6), 568-576.

[25] Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. Paper presented at the Advances in neural information processing systems.

[26] Utgoff, P. E., & Stracuzzi, D. J. (2002). Many-layered learning. Neural Computation, 14(10), 2497-2529.

[27] Zaccone, G. (2016). Getting Started with TensorFlow: Packt Publishing Ltd.

[28] Zheng, X., Chen, H., & Xu, T. (2013). Deep Learning for Chinese Word Segmentation and POS Tagging. Paper presented at the EMNLP.

[29] Gambhir, M., & Gupta, V. (2017). Recent automatic text summarization techniques: a survey. Artificial Intelligence Review, 47(1), 1-66.

[30] Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: a library for support vector machines. ACM transactions on intelligent systems and technology (TIST), 2(3), 27.