# Comparison of Data Augmentation Methods in Pointer-Generator Model Using Various Sentence Ranking Methods

**Tomohito Ouchi, Masayoshi Tabuse**

*Graduate School of Life and Environmental Sciences, Kyoto Prefectural University,*
*1-5 Shimogamohangi-cho, Sakyo-ku, Kyoto 606-8522, Japan*
*E-mail: t_ouchi@mei.kpu.ac.jp, tabuse@kpu.ac.jp*

**Abstract**

In the existing research, we proposed a data augmentation method using topic model for Pointer-Generator model. In this study, we add to the sentence ranking method in the data augmentation method. Specifically, we add two ranking methods using LexRank and Luhn. LexRank is based on Google's search method and Luhn defines sentence features and ranks sentences. We compare three data augmentation methods. We considered which method is suitable for data augmentation. We confirm that most accurate model is the model using data augmentation method by topic model.

*Keywords*: automatic summarization, data augmentation, Pointer-Generator Model, Extractive summarization.

## 1. Introduction

Currently, the amount of information on the Internet is increasing exponentially, and it is said that it will reach 59ZB in the research in May 2020[1]. It is also said that the total amount of data created in the next three years exceeds the total amount of data in the past 30 years, and the total amount of data created in the next five years is more than three times the total amount of data created in the past five years. Under such circumstances, the issue of selecting information is an urgent problem. Automatic summarization struggles that issue. However, it can be said that extractive summarization that only made up with sentences is not sufficient. Since the sentence-to-sentence connection is not taken into consideration, readability is lacking. Therefore, it is needed generative summarization as a technology that looks ahead. A generative summarization basically uses the Encoder-Decoder model, which learns the relationship between input and output and generates one word at a time in the output when a new input comes in during the test. Various models have been proposed [2,3]. In this study, the Pointer-Generator model [2] uses as the baseline model. One of the issues with the generative summarization model is that data maintenance is costly. We have to attach a manual summary to each article in order to make the generative summarization model. Therefore, we focused on data augmentation as a method that can be applied to any model. This is to create extended data from existing data. As a result, it was confirmed that the accuracy of the evaluation metric ROUGE [4] of Pointer-Generator model applied by the data augmentation method is improved by about 1% compared to baseline model.

Next, we explain the method of data augmentation simply. We decide the importance of each sentence in each article. And the sentence with the lowest importance is removed to obtain extended data. In the existing research [5], the topic model was used to measure the importance of sentences. In this study, in addition to that, the method called Luhn[6] and LexRank[7] was used. These three techniques are described in Section 2. Experiments and results are described in Section 3. And discussions are given in Section 4.

## 2. Data Augmentation Method

This section describes the three models used in the data augmentation method. In each method, each sentence is scored in an article, and the sentence with the lowest score is removed to obtain extended data.

### 2.1. *Luhn*

We measure position of the top 100 in most frequent words removed stop-words. We define words with a distance of 5 or less as one cluster. The score of a cluster is the square of the number of important words in a cluster divided by the distance between the first and last words of the cluster. Finally, the maximum score of each cluster becomes the score of the sentence.

### 2.2. *LexRank*

First, we explain PageRank, which is the basis of LexRank. The basic idea of PageRank is that linked pages are good pages, and links from more linked pages are evaluated highly. This rating is equivalent to the user inflow to the page. This is because if links are provided from many pages, it is easy to flow in, and the inflow from popular pages is larger than the inflow from normal pages. Links between pages can be represented by a matrix, which is the probability that the user will transition from that page to another linked page. The matrix is made with dividing by the total number of links on each page. The purpose of PageRank is to use this matrix to determine the probability that a user will stay on each page, that is, the rating of the page. PageRank is based on the premise that the page stay probability will eventually stabilize if the page transition is repeated many times, so that the transition matrix multiplied by the stay probability vector becomes closer to the transition matrix.

In LexRank, the transition matrix is the matrix of the cosine similarity of the Tf-Idf score between sentences. The basic idea of LexRank is that sentences similar to many sentences and sentences similar to important sentences are considered to be important sentences.

### 2.3. *Topic Model*

This method was used in the existing method. For how to determine the importance of sentences using the topic model, we referred to existing research [8]. The topic model is one of the language models that assumes that one document consists of multiple topics. In addition, each topic has an appearance word distribution. The method of determining the importance of a sentence is as follows.

1. Calculate the frequency of occurrence in a topic with words that make up a sentence
2. Sum of all the words that make up the sentence
3. Divide by the square root of the sentence length
4. Sum on all topics

## 3. Experiment and Results

In this section, the experimental conditions and the results of additional experiments using the Luhn method and the LexRank method are included.

### 3.1. *Parameter Setting*

The CNN / DailyMail dataset is used as the dataset for training, evaluating, and testing. The training data, evaluating data, and test data are 287,226 articles, 13,768 articles, and 11,490 articles, respectively. The model used for the experiment is the Pointer-Generator model, which is divided into a copy mechanism and a coverage mechanism when learning. The Copy mechanism calculates the error of the evaluating data each time the epoch ends and we uses the model of the epoch with the lowest error in Early Stopping. Early Stopping what we mean here, uses a model that waits twice as many epochs as the error seems to be the minimum, unless the minimum value is updated. Next, in the coverage mechanism, the same processing is performed in the coverage loss. We use ROUGE as using for evaluation on existing research.

The program used in this research uses PyTorch. It has been confirmed that this program can achieve the same result as [2]. The hidden layer vector size was set to 256 and the embedded vector size was set to 128. The batch size was set to 8. In the original paper, the batch size is 16, so double learning is required to learn the same number of articles. The beam size was set to 4. The beam search will be described later. The number of vocabulary was set to 50,000. The learning rate was set to 0.15.

In this program, the number of words used to encode an input article is limited to 400. This setting has no effect on learning an extended data. Specifically, an extended data is the same as an original data. This is because the extracted sentence may not be within 400 words from the beginning. We must confirm that the extracted sentence is present in the input article. Therefore, I found the article with the most number of words among the articles used in the training data. The number of words with the most words was 2,380. And the upper limit of the number of words used in encoding the input article was set to 2,380. Table 1 shows the values of ROUGE when the maximum number of words is 400 and 2,380. In the Table 1, f, r, and p represent the F value, recall, and precision, respectively.

Table1 the values of ROUGE when the maximum number of words is 400 and 2,380

| | ROUGE-1-f | ROUGE-1-r | ROUGE-1-p | ROUGE-2-f | ROUGE-2-r | ROUGE-2-p |
|---|---|---|---|---|---|---|
| 400 | 0.3935 | 0.4372 | 0.3800 | 0.1709 | 0.1891 | 0.1662 |
| 2380 | 0.3958 | 0.4181 | 0.3994 | 0.1741 | 0.1832 | 0.1770 |

| | ROUGE-L-f | ROUGE-L-r | ROUGE-L-p |
|---|---|---|---|
| 400 | 0.3616 | 0.4014 | 0.3493 |
| 2380 | 0.3644 | 0.3846 | 0.3679 |

Table 1 shows when the upper limit of the number of words is increased from 400 to 2,380, the value of ROUGE increases slightly. In the following, the experiment is performed with the upper limit of the number of words set to 2,380.

### 3.2. *Beam search*

Greedy method contrasts with beam search. This is because, when generating a word, one word with the highest generation probability is selected, while in beam search, processing is performed while holding the top K words. Then, we make the final summarizations by multiplying the probabilities of each word generation, and make the highest one the final summarization. In this experiment, this K value is set to 4.

### 3.3. *Results*

The results are as below.

Table 2 Results of learning 287226 articles using 6 methods

| | normal | extended | ri | sr |
|---|---|---|---|---|
| rouge-1 | 0.3820 | 0.3948 | 0.3856 | 0.3869 |
| rouge-2 | 0.1640 | 0.1724 | 0.1649 | 0.1683 |
| rouge-L | 0.3514 | 0.3624 | 0.3527 | 0.3551 |

| | rs | rd | Luhn | LexRank |
|---|---|---|---|---|
| rouge-1 | 0.3911 | 0.3817 | 0.3739 | 0.3777 |
| rouge-2 | 0.1696 | 0.1657 | 0.1640 | 0.1616 |
| rouge-L | 0.3596 | 0.3499 | 0.3449 | 0.3489 |

In order of good results, the existing method[5], baseline, ri, sr, rd, LexRank, rs, and Luhn.

A summary example generated in each model is shown.

**reference**
marseille prosecutor says `` so far no videos were used in the crash investigation " despite media reports .
journalists at bild and paris match are `` very confident " the video clip is real , an editor says .
andreas lubitz had informed his lufthansa training school of an episode of severe depression , airline says .

**normal**
new : `` a person who has such a video needs to immediately give it to the investigators " robin 's comments follow claims by two magazines , german daily bild and french paris match , of a cell phone video showing the harrowing final seconds from on board germanwings flight 9525 .
paris match and bild reported that the video was recovered from a phone at the wreckage site .

**extended**
marseille prosecutor brice robin says he was not aware of any video footage from the plane .
robin 's comments follow claims by two magazines , german daily bild and french paris match .
`` one can hear cries of ` my god ' in several languages , " paris match reported .

**ri**
marseille prosecutor brice robin told cnn that `` so far no videos were used in the crash investigation " robin 's comments follow claims by two magazines , german daily bild and french paris match .
all 150 on board were killed .

**sr**
new : `` it is a very disturbing scene , " official says .
new : `` one can hear cries of ` my god ' in several languages , " prosecutor says .
`` one can hear cries of ` my god ' in several languages , " official says .

**rs**
marseille prosecutor brice robin says he was not aware of any video footage from on board .
robin 's comments follow claims by two magazines , german bild and french paris match .
the video was found by a source close to the investigation .

**rd**
french prosecutor leading an investigation into the crash of germanwings flight 9525 .
robin 's comments follow claims by two magazines , german daily bild and french paris match , of a cell phone video showing the harrowing final seconds from on board the plane .
german airline lufthansa confirmed tuesday that co-pilot andreas lubitz had a `` previous episode of severe depression "

**Luhn**
new : `` so far no videos were used in the crash investigation , " official says .

new : `` a person who has such a video needs to immediately give it to the investigators , '' says a forensic psychologist .

**LexRank**
marseille prosecutor brice robin told cnn that `` so far no videos were used in the crash investigation '' french president francois hollande says it should be possible to identify all the victims using dna analysis by the end of the week .

## 4. Conclusion

In this study, in addition to the existing studies, we experimented with a data augmentation method using the Luhn and LexRank methods. The results confirmed that the best data augmentation method is to use the topic model of the existing research. In the future, we would like to confirm the effectiveness of data augmentation for state-of-the-art models. When extracting a sentence from an article, I would like to try a method for extracting multiple sentences instead of one sentence. We also expect that the number of sentences will change depending on the length of the article.

## References

1. International Data Corporation (IDC) https://www.idc.com/
2. Abigail. See, Peter J. Liu, et al. “Get To The Point: Summarization with Pointer-Generator Networks” arXiv:1704.04368(2017)
3. Yang Liu, Mirella Lapata “Text Summarization with Pretrained Encoders” arXiv:1908.08345v2(2019)
4. Chin-Yew Lin, “ROUGE: A Package for Automatic Evaluation of Summaries” Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004, Barcelona, Spain (2004)
5. T. Ouchi, M. Tabuse, “Effectiveness of Data Augmentation in Pointer-Generator Model” ICAROB(2020)
6. Luhn, H.P., “The automatic creation of literature of abstracts” IBM journal of research and development, vol. 2, No. 2, pp.159-165 (1958)
7. Gunes Erkan, Dragomir R Radev, “LexRank : graph-based lexical centrality as salience as text summarization” journal of Artificial Intelligence Research, vol. 22, pp. 457-479 (2004)
8. H.Sigematsu, I.Kobayashi “Generation of abstracts considering importance of potential topics” The Association for Natural Language Processing (2012) (In Japanese).