Jurnal Teknologi Informasi dan Pendidikan
Volume 15, No. 1, Maret 2022
https://doi.org/10.24036/tip.v15i1

P.ISSN: 2086 – 4981
E.ISSN: 2620 – 6390
tip.ppj.unp.ac.id

Automatic Summarization of Indonesian Texts Using a Hybrid Approach

Kania Evita Dewi[1*]✉, Nelly Indriani Widiastuti[2]
[1,2]Information Technology Department, Universitas Komputer Indonesia, Indonesia
**Corresponding Author: kania.evita.dewi@email.unikom.ac.id*

Article Information

Article history:

No. 451
Rec. June 9, 2021
Rev. October 13, 2021
Acc. May 13, 2022
Pub. May 15, 2022
Page. 37 – 43

Keywords:

- *Automatic Summarization*
- *Indonesian Language*
- *Hybrid Approach*
- *Texts*

ABSTRACT

Automatic text summarization is a system that reduces the number of sentences without losing important information in the document. There are three approaches to making automatic text summarization: extractive, abstraction, and hybrid. The extractive approach is to choose the main sentence without changing it into a new sentence. The abstract approach is to construct new sentences that describe the document's contents. At the same time, the hybrid approach designed in this study is a combination of extractive and abstractive approaches. By designing automatic text summarization in Indonesian with a hybrid approach, the summary results obtained will be closer to those made by humans and have higher legibility. The stages of automatic text summarization are divided into two, namely preprocessing and processing. In the preprocessing step, sentence separation, tokenization, coreference resolution, stop words removal, and feature extraction are carried out. The summary designed is each document through two summary approaches. Further research can be done to input longer documents and input in the form of multiple documents.

1. INTRODUCTION

Documents with long entries are pretty tiring to read, and sometimes readers only need the information. Text summarization is a solution to finding the information contained in the text. But if it is done manually, summarizing the text will take a long time and a lot of money. Therefore, automatic text summarization is required. Automatic text summarization is the process of automated text generation by software that significantly depicts the information contained in the source text. The contents of the summarized result are no longer than the source text [1]. Automatic text summarization is not new research. Research on this has been conducted since 1958 by Luhn [2]. There have also been many studies of automatic texts in Indonesian. Mostly in making automatic text

summaries that have been created using an extractive approach. This extractive summarization has been used with various methods in machine learning, such as Support Vector Machine [3], Relevance Vector Machine [3], Naïve Bayes [4]. Some use optimization methods [5][6], using neural networks [7], Vector Space Model [8], Restricted Boltzmann Machine [9], and many more [10] [11] [12].

An abstract approach has been made in Kasyfi Ivandera's research [13]. In his research, it can be seen that the input to the system is a paragraph. Meanwhile, news documents usually do not contain just a paragraph. Therefore, in this study, a model is designed to process documents with more than one paragraph. This study will adopt the research of Wang [14]. In Wang's research, summarization uses a hybrid approach. Input in Wang's research is a Chinese news text. The first stage in Wang's research is the extractive stage. In Wang's research, this stage uses the Pagerank method. Automatic text summarization uses the Pagerank method, which is an unsupervised summary. Selection of the core sentence based on the relationship between the text title and the sentence and the relationship between the sentences. However, according to Hadyan's [15] research, which performed extractive summarization using Pagerank in Indonesian text, the F-measure obtained was not very good. So in this study, the stages of making extractive summaries will be carried out using the Support Vector Machine method, which can be seen in research [3] that can produce fairly good accuracy.

Meanwhile, the abstractive process is designed using seq2seq RNN as in Wang's research with the addition of the Pointer Mechanism method. This Pointer Mechanism is considered to be able to solve the OOV (out of vocabulary) problem, which usually becomes a problem when the features used are words. This has also been shown in Silpi's research. Although the results are still not good, with the pointer mechanism, Rouge-1, Rouge-2, and Rouge-L values increased by 18%, 34%, and 20% compared to those using the standard attention mechanism model [16].

This study aims to recommend the design of automatic text summarization in Indonesian. So that the results of the summarized text can be read correctly, in this study, an automated text summarization system is designed using a hybrid approach. The design of this system is based on the results of a literature study. The results of this study are in the form of recommendations for automatic text summarization designs.

2. RESEARCH METHOD

This study will discuss design recommendations for forming an automatic text summarization system for long texts. The design of this system was based on the results of previous research. This research consists of several stages, namely:

P.ISSN: 2086 – 4981
E.ISSN: 2620 – 6390
tip.ppj.unp.ac.id

Formulation of the problem

It is done by reviewing the literature on summarization that previous researchers have done at this stage. The literature sought was journals or papers about automatic text summarization in Indonesian. This stage is helpful for building problems in this study.

Literature Search

At this stage, it is done by reviewing the literature on automatic text summarization. The literature searched is not limited to automatic text summarization in Indonesian. The literature searched for books, journals, or papers that have the latest automatic text summarization topics and produce the best accuracy in automatic text summarization.

The design of an automatic text summarization system uses a hybrid approach

In this research, the automatic text summarization system designed is automatic text summarization with a hybrid approach, which combines extractive and abstractive approaches. The design of this system is based on the literature obtained in the second stage of this study. The system design consists of two steps, namely preprocessing and processing.

3. RESULTS AND DISCUSSION

Automatic text summarization is the process of shortening a document without losing the information contained therein. There have been many cases of summarization of Indonesian texts, but mostly using an extractive approach. With this approach, besides being less legible, the resulting accuracy is also not good. The development of automatic text summarization with an abstract approach has been carried out [13]. But from the paper that is read, it can be seen that the input is in the form of 1 paragraph consisting of only a few sentences, and there are still problems with words that are not in the training data or what is often called Out of Vocabulary (OOV). So, this study designed an automatic text summarization with a hybrid approach. With this hybrid design, it is hoped that the summary results will be more legible. This automatic text summarization design will be divided into two stages: the preprocess and the summarization process.

The first stage that must be done is the preprocess. This process is done to clean the document from noise. Apart from being clean from noise from the preprocessing results, the sentence weight will be used as input for extractive summarization. At this stage are the processes usually carried out in the preprocess for extractive summarization [3] [1].

3.1. Split Sentence

This stage is commonly carried out in summarizing text with an extractive approach because the output of extractive summarization is a sentence. This process is almost present in every extractive summarization study. This stage breaks the paragraph into sentences, where the delimiter is the point.

3.2. Tokenizing

This stage is commonly carried out in summarizing text with an extractive approach because the output of extractive summarization is a sentence. This process is almost present in every extractive summarization study. This stage is done by breaking the paragraph into sentences, where the delimiter is the point

3.3. Coreference Resolution

This stage helps change pronouns for people and things. Extractive summarization selects sentences considered introductory sentences in a document as summary sentences without changing the sentence. If coreference resolution is not made first, then when a word is chosen to have a person or thing pronoun, it will confuse the reader

3.4. Filtering

This stage helps remove punctuation marks. The results of this stage are words and numbers only

3.5. Case Folding

This stage is done to homogenize the letters. Either uniform it to capital letters and all lowercase.

3.6. Stop Words Removal

Eliminate words that are considered unrelated to the topic. In this study, Tala's stoplist was used [18].

3.7. Feature Extraction

This stage helps give weight to the sentence of the candidate for the summary. Feature extraction that is often used is TF-IDF [17] [18]. At the same time, other studies also use other features such as sentence length, sentence position, feature title, sentence to sentence, negative keywords, and the connection between sentences [3] [19] [20].

The next stage is the stage of making summaries. The draft made by the document that has been cleaned goes into the extractive approach text summarization stage. This stage aims to shorten the time for making abstractive summaries. The results of this stage are sentences that are considered important in a document. Selection of sentences using the supervised method, the Support Vector Machine method. The way these methods work is

P.ISSN: 2086 – 4981
E.ISSN: 2620 – 6390
tip.ppj.unp.ac.id

with the training dataset. This method creates a hyperplane that limits the summarized sentences and those that aren't. If the entered sentence is a summary sentence, the system will produce a positive value. In contrast, if the entered sentence is not a summary sentence, the method will produce a negative value. This method is deemed sufficient to produce a similar summary to other methods. Enter data which was a document consisting of sentences, D={S_1,S_2,S_3,…,S_n }, with n many sentences, the output of this method is the selected sentence, for example D_1={S_1,S_2,..,S_k }, where k<n.

The next stage is to form summary sentences. This stage was designed using the seq2seq RNN method as in Wang's study [15]. In Wang's research, it can be seen that with this method, the summary results are better than the usual RNN. At this stage, the resulting sentences are broken down into words. Then the words are made into vectors using word embedding. Then, with the RNN encoding process, a summary sentence is formed. So the input of the extractive summarization results is D_1={S_1, S_2,.., S_k }, using the RNN method will produce D_2={w_1,w_2,..,w_m }, where m is the number of words that is smaller than the number of words which is on D_1. This process can be seen in Figure 1. The difference between Kasyifi and Wang's RNN is that in Kasyifi's research, he used the development of the RNN, namely LSTM, whereas Wang used the seq2seq RNN method. In the RNN method, a point mechanism is added, which is helpful to reduce interference from the word Out Of Vocabulary (OOV), which is a problem in Kasyifi's research.

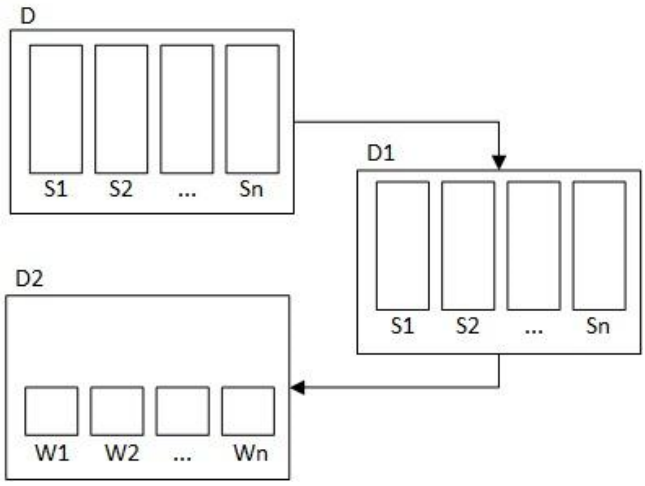

Figure 1. Suggested Auto Text summarization design

4. CONCLUSION

Based on the results and discussion, it is found that this design can handle long documents so that it can be continued to make abstracts. In this research, the new input

data is in the form of a single document so that in the future, we can try to design a system that can handle multi-document input.

REFERENCES

[1] J.-M. Torres-Moreno, *Automatic text summarization*. John Wiley \& Sons, 2014.
[2] H. P. Luhn, "The automatic creation of literature abstracts," *IBM J. Res. Dev.*, vol. 2, no. 2, pp. 159–165, 1958.
[3] E. Rainarli and K. E. Dewi, "Relevance Vector Machine for Summarization," in *IOP Conference Series: Materials Science and Engineering*, 2018, vol. 407, no. 1, p. 12075.
[4] S. Raharjo and E. Winarko, "Klasterisasi, klasifikasi dan peringkasan teks berbahasa indonesia," *Pros. KOMMIT*, 2014.
[5] A. N. Ammar and S. Suyanto, "Peringkasan Teks Ekstraktif Menggunakan Binary Firefly Algorithm," *Indones. J. Comput.*, vol. 5, no. 2, pp. 31–42, 2020.
[6] Z. Zulkifli, A. T. Wibowo, and G. Septiana, "Pembobotan Fitur Ekstraksi Pada Peringkasan Teks Bahasa Indonesia Menggunakan Algoritma Genetika," *eProceedings Eng.*, vol. 2, no. 2, 2015.
[7] M. N. Rachmatullah and A. Primanita, "IMPLEMENTASI JARINGAN SYARAF TIRUAN PADA SISTEM PERINGKASAN TEKS OTOMATIS MENGGUNAKAN EKSTRAKSI CIRI."
[8] C. Slamet, A. R. Atmadja, D. S. Maylawati, R. S. Lestari, W. Darmalaksana, and M. A. Ramdhani, "Automated text summarization for Indonesian article using vector space model," in *IOP Conference Series: Materials Science and Engineering*, 2018, vol. 288, no. 1, p. 12037.
[9] R. Widiastutik, J. Santoso, and others, "Peringkasan Teks Ekstraktif pada Dokumen Tunggal Menggunakan Metode Restricted Boltzmann Machine," *J. Intell. Syst. Comput.*, vol. 1, no. 2, pp. 58–64, 2019.
[10] R. Indrianto, M. A. Fauzi, and L. Muflikhah, "Peringkasan Teks Otomatis Pada Artikel Berita Kesehatan Menggunakan K-Nearest Neighbor Berbasis Fitur Statistik," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 1, no. 11, pp. 1198–1203, 2017.
[11] M. Mustaqhfiri, Z. Abidin, and R. Kusumawati, "Peringkasan teks otomatis berita berbahasa Indonesia menggunakan metode Maximum Marginal Relevance," *Matics*, 2011.
[12] C. Fang, D. Mu, Z. Deng, and Z. Wu, "Word-sentence co-ranking for automatic extractive text summarization," *Expert Syst. Appl.*, vol. 72, pp. 189–195, 2017.
[13] K. Ivanedra and M. Mustikasari, "Implementasi Metode Recurrent Neural Network Pada Text Summarization Dengan Teknik Abstraktif," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 6, no. 4, pp. 377–382, 2019.
[14] S. Wang, X. Zhao, B. Li, B. Ge, and D. Tang, "Integrating extractive and abstractive

P.ISSN: 2086 – 4981
E.ISSN: 2620 – 6390
tip.ppj.unp.ac.id

models for long text summarization," in *2017 IEEE International Congress on Big Data (BigData Congress)*, 2017, pp. 305–312.

[15] F. Hadyan, M. A. Bijaksana, and others, "Comparison of Document Index Graph Using TextRank and HITS Weighting Method in Automatic Text Summarization," in *Journal of Physics: Conference Series*, 2017, vol. 801, no. 1, p. 12076.

[16] A. S. Alpiani and S. Suyanto, "Pointer Generator dan Coverage Weighting untuk Memperbaiki Peringkasan Abstraktif," *Indones. J. Comput.*, vol. 4, no. 2, pp. 169–176, 2019.

[17] N. G. dan Indriati Indriati dan Ratih Dewi, "Peringkasan Teks Otomatis Secara Ekstraktif Pada Artikel Berita Kesehatan Berbahasa Indonesia Dengan Menggunakan Metode Latent Semantic Analysis," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 9, pp. 2821–2828, 2018.

[18] A. Najibullah and W. Mingyan, "Otomatisasi peringkasan dokumen sebagai pendukung sistem manajemen surat," *… Ilm. Teknol. Sist. Inf.*, 2015.

[19] B. Zaman and E. Winarko, "Analisis Fitur Kalimat untuk Peringkas Teks Otomatis pada Bahasa Indonesia," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 5, no. 2, 2011.

[20] M. A. Fattah and F. Ren, "GA, MR, FFNN, PNN and GMM based models for automatic text summarization," *Comput. Speech \& Lang.*, vol. 23, no. 1, pp. 126–144, 2009.