Automatic Text Summarization for Indonesian Language Using TextTeaser

D Gunawan[1], A Pasaribu[1], R F Rahmat[1] and R Budiarto[2]

[1] Department of Information Technology, Faculty of Computer Science and Information Technology, University of Sumatera Utara, Medan, Indonesia
[2] Department of Information System, College of Computer Science and Information Technology, Albaha University, Saudi Arabia

Email: danigunawan@usu.ac.id, anwargoog@gmail.com, romi.fadillah@usu.ac.id, rahmat@bu.edu.sa

Abstract. Text summarization is one of the solution for information overload. Reducing text without losing the meaning not only can save time to read, but also maintain the reader's understanding. One of many algorithms to summarize text is TextTeaser. Originally, this algorithm is intended to be used for text in English. However, due to TextTeaser algorithm does not consider the meaning of the text, we implement this algorithm for text in Indonesian language. This algorithm calculates four elements, such as title feature, sentence length, sentence position and keyword frequency. We utilize TextRank, an unsupervised and language independent text summarization algorithm, to evaluate the summarized text yielded by TextTeaser. The result shows that the TextTeaser algorithm needs more improvement to obtain better accuracy.

1. Introduction

Automatic Text Summarization (ATS) is the process to reduce text in order to obtain important sentences by the machine which implements particular algorithm or method. One method to produce text summary is extraction-based summarization. Extraction-based summarization method extracts important sentences in article and then unify them into one summary, therefore the sentences yielded by this algorithm are part of the original text without modification. This method is implemented by TextTeaser algorithm to summarize text. This algorithm is available freely as open source and can be downloaded from github repository. This algorithm calculates text feature, sentence length, sentence position and keyword frequency. TextTeaser is not intended to replace original text. The result will depict the whole text. TextTeaser has been tested in articles which written in English. In this research we test the TextTeaser for articles which written in Indonesian Language. We also evaluate the TextTeaser performance compared to TextRank algorithm.

Previously, a research text summarization for Indonesian news articles by using guided summarization technique has been conducted [1]. This research utilized a supervised method called SWING [2]. They define information aspects such as WHAT, WHO_AFFECTED, WHERE, WHEN, WHY, DAMAGES according to the topic of the news to obtain the text summary. On the other hand, we do not need information aspect because TextTeaser algorithm calculates title feature and keyword frequency.

Another research utilizes Latent Dirichlet Allocation and Genetic Algorithm to summarize text for Indonesian language [3]. They obtain the best F-measure value 0.556926 with precision value 0.53448 and recall value 0.58134. The summary ratio for their method is 30%. This research requires training data before actually summarizing the text. Our research does not require training data as the TextTeaser algorithm does calculation on items mentioned previously.

Massive number of information does not only occur in online news. Twitter as one of the most popular social network overwhelms the user with the information. Previously conducted research proposed a method to explain a trending topic by topic categorization, explanation generation, distance-based clustering to build subtopics, and sentence compression [4]. This research collects users' messages related to the trending topic then summarize them. There are possibilities of similar messages among different users. Our research does not consider similarity among articles.

2. Research methodology

This research follows a few assumptions in order not to widen the discussion. First of all, the texts are written in Bahasa Indonesia. Then, the sentence pattern follows official Indonesian spelling system which is called *Ejaan Yang Disempurnakan* (EYD). The system does not handle typo that was made by the authors. The last, the summarized text will not handle cohesion and coherent among sentences.

According to figure 1, original text will be the input of several processes to obtain the summarized text. The summarized text will be evaluated with TextRank algorithm [5]. TextRank succeeds to identify most importance sentences. It does not need any training data, therefore adapting TextRank to another language is expected to be easy.

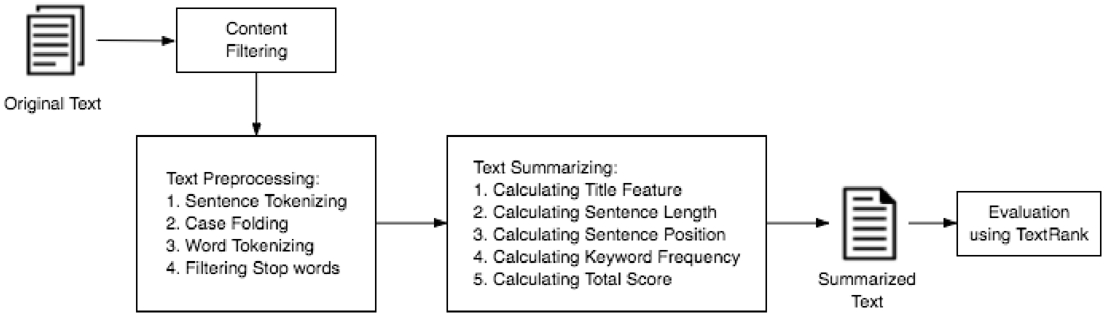
Original Text
Content Filtering
Text Preprocessing:
1. Sentence Tokenizing
2. Case Folding
3. Word Tokenizing
4. Filtering Stop words
Text Summarizing:
1. Calculating Title Feature
2. Calculating Sentence Length
3. Calculating Sentence Position
4. Calculating Keyword Frequency
5. Calculating Total Score
Summarized Text
Evaluation using TextRank

Figure 1. Text summarization process using TextTeaser and evaluated by TextRank

2.1. Preparing the Text

The required text for the research is clear from any attribution. However, the texts from the Internet are available in various data format with many different attributes for each website. Therefore, we need to prepare the text before applying TextTeaser algorithm. There are two steps of preparing the text, that are content filtering and text preprocessing.

The original texts are the main article without any extra parts from the web pages, including HTML tags, header, navigation, banners, link to other sources, and etc. However, almost all the web pages have all these extra parts to arrange the layout of the web pages. To cover this issue, there are some algorithms such as boilerplate algorithm [6] that can extract the main article from a web page. The Boilerplate algorithm does not require any inter-document knowledge, such as frequency of text block, common page layout, and etc.

Sometimes the result is still including some texts that are not directly related to the content of the article. For example, caption below the image, links to other news, author initials, or initials for news source. These extras have different structure among online newspaper publishers. Therefore, we need to adjust particular content filtering for each online newspaper publishers. This could be alleviating if the publishers follow the standard in semantic web. The publisher might embed Resource Description

Framework (RDFa) in HTML and use standardized vocabulary such as Dublin Core to make the machine easier to read. Required protocol can be developed based on available vocabulary [7].

Content filtering is a rudimentary process to obtain valid sources. The result after content filtering will be processed to the text preprocessing. The first step is case folding. This step will convert all the uppercase to lowercase in order to ease letter comparison. Next step is sentence tokenizing. This step is required as TextTeaser calculates sentence score based on its position in a paragraph. After sentence tokenizing, the next step is word tokenizing. TextTeaser will calculate word score based on its appearance in an article. Finally, we eliminates some words based on stop words list provided by Tala [8].

2.2. Text Summarization

TextTeaser uses extraction method for text summarization. It selects sentences that possess best score among others. Therefore, the calculation is done sentence by sentence. Best score sentences are obtained from the calculation regarding four elements of the article such as title feature, sentences length, sentence position and keyword frequency. Title feature is used to score the words in a sentence that have similarity to the words in the title. Sentence length is scored depends on the number of words in the sentence. Sentence position is the location of the sentence in a text. The sentence which is located in introduction and conclusion will have higher score. Keyword frequency is the frequency of the words used in the text.

The calculation of *TitleScore* is determined by calculating every word in each sentence of the article which has similarity with the words in the article title. TitleScore can be yielded by applying equation (1) to the sentence regards to the title. As shown in equation (1), K is the number of words in sentence s_i, and T is the number of words in article title t_j.

$$TitleScore(s_i) = \frac{|K \cap T|}{(|t_j| \cdot 1)} \tag{1}$$

The *SentenceLength* score is obtained by counting the number of words in a sentence. Equation (2) is used to calculate sentence length every sentence in the article. According to equation (1), $|s_i|$ is the number of words in a sentence including stop words.

$$SentenceLength(s_i) = \frac{(x - (x - |s_i|))}{x} \tag{2}$$

Table 1. Sentence position score.

Sentence Position	Distributed Probability
0.0 < x ≤ 0.1	0.17
0.1 < x ≤ 0.2	0.23
0.2 < x ≤ 0.3	0.14
0.3 < x ≤ 0.4	0.08
0.4 < x ≤ 0.5	0.05
0.5 < x ≤ 0.6	0.04
0.6 < x ≤ 0.7	0.06
0.7 < x ≤ 0.8	0.04
0.8 < x ≤ 0.9	0.04
0.9 < x ≤ 1.0	0.15

Another calculation regarding the sentence is the calculation of its position. The sentences which have higher scores are found at the beginning or the conclusion of the news articles rather than the other positions. Sentence position score, which is represented with $PositionScore(o, S)$, can be calculated by using equation (3). Sentence position in an article is represented with o and the number of all sentences is represented with S. $PositionScore(o, S)$ is determined based on table 1.

$$PositionScore(o, S) = \frac{o}{(S \cdot 1)} \tag{3}$$

Keyword frequency is the number of word occurrence in the whole article. Keyword frequency is calculated by using two methods, that are Density-Based Selection (DBS) and Summation-Based Selection (SBS) [9]. Before commencing the calculation of keyword frequency using DBS and SBS, article keywords and the score is determined by finding unique words. Unique words are the words that has no duplication and not included in stop words list. Next, keywords and their frequency are ordered from the highest weight to the lowest weight. We take ten top keywords to be calculated by using equation (4) to obtain keyword score (represented with $KeywordScore(k_i)$). Where k_i is the keyword that will be calculated, $|k_i|$ is the frequency of keywords occurrence in the text, and $|W|$ is the number of unique keywords in the text. DBS and SBS can be calculated after each keyword has its score.

$$KeywordScore(k_i) = (1 \cdot |k_i| \div |W|) \cdot 1.5 \tag{4}$$

Density-Based Selection is used to determine keyword rank in the text based on several parameters. Those parameters are scored keyword set, words in the sentence that will be calculated, and the words in keyword set [9]. Then the parameters will be calculated by using equation (5), where $dbs(s_i)$ is density-based selection value, K is the number of words in sentence s_i which also in top keywords list, $dbs(w_j)$ and $dbs(w_{j+1})$ are the weight of keyword w_j and w_{j+1} respectively, and $distance(w_j, w_{j+1})$ is the value of two adjacent keyword (w_j and w_{j+1}) and also not a keyword or stop word in s_i.

$$dbs(s_i) = \frac{1}{K \cdot (K+1)} \cdot \sum_{j=1}^{K-1} \frac{dbs(w_j) \cdot dbs(w_{j+1})}{distance(w_j, w_{j+1})^2} \tag{5}$$

Summation-Based Selection is used to determine the sentence weight regarding the occurrence of top keywords in a sentence. The value will be higher if there are more top keywords in a sentence. The weight is calculated by using equation (6), where $sbs(s_i)$ is summation-based selection value, s_i is a sentence in a text, $|s_i|$ is the number of words in s_i, then parameter to take the word weight (w_k) which represents the content of text with $\tau(\tau > 0)$.

$$sbs(s_i) = \frac{1}{|s_i|} \cdot \Big(\sum_{w_k \in s_i} sbs(w_k)^{\tau} \Big)^{\frac{1}{\tau}} \tag{6}$$

Keyword frequency, represented by $KeeywordFreq(s_i)$, is obtained by summing the values of $dbs(s_i)$ and $sbs(s_i)$, then divided by 2.0 * 10.0 as shown in equation (7).

$$KeywordFreq(s_i) = \frac{(dbs(s_i) + sbs(s_i))}{2.0 \cdot 10.0} \tag{7}$$

Title feature, sentence length, sentence position and keyword frequency calculation are used to obtain sentence score. The sentence score calculation is stated in equation (8), where $TotalScore(s_i)$

represents sentence score, TS is title score, SL is sentence length, SP is sentence position and KF is keyword frequency.

$$TotalScore(s_i) = \frac{(TS \cdot 0.5 + SL \cdot 0.5 + SP \cdot 1.0 + KF \cdot 2.0)}{4.0} \tag{8}$$

2.3. Evaluation

This research uses intrinsic evaluation method with method recall (R), precision (P), and F-Score (F). Recall in the context of automatic text summarization is number of sections of text that is relevant with the original text based on the number of all sentences in the text. Precision is the number of relevant text that are obtained from all the sentences from original text. F-Score is average weight of recall and precision value. The value of F-Score is the range between 0 and 1. The summarized text from both algorithms will be considered similar if the value is close to 1.

TextTeaser algorithm will be compared with another automatic text summarization algorithm called TextRank. TextRank algorithm is a graph-based ranking model for text processing. This algorithm deciding the importance of a sentence by the vertex which represents that sentence. The higher value of vertex means the sentence is more important than other sentences. Besides, the link between one vertex to another one also has role to determine the importance of the sentence. The advantages of this algorithm is unsupervised and language independent, which means TextRank is expected to work with any language.

The calculation for Recall, Position and F-Score are shown in equation (9), where S is the result of text summarization using TextTeaser algorithm, and T is the result of text summarization using TextRank.

$$F = \frac{2PR}{(P + R)}; P = \frac{|S \cap T|}{|S|}; R = \frac{|S \cap T|}{|T|} \tag{9}$$

3. Result and discussion

This research uses 3075 samples of news articles from various online newspaper publishers in Indonesia. After content filtering and text preprocessing, the articles are summarized by using TextTeaser algorithm. We evaluate the summarized text by calculating recall (R), precision (P), and F-score (F) value between TextTeaser algorithm and TextRank algorithm.

Table 2. F-Score value after evaluation.

F-Score Range	Number of Articles	Percentage (%)
1	233	7.6
0.9 – 0.99	36	1.17
0.8 – 0.89	615	20
0.7 – 0.79	778	25.30
0.6 – 0.69	725	23.58
0.5 – 0.59	150	4.87

Table 2 shows the number of occurrences every F-Score value of 3075 summarized texts. There are 28,75% of summarized texts have F-Score ≥ 0.8, where 233 summarized texts have F-Score = 1. Meanwhile, 31.36% of summarized texts have F-Score between the range 0.5-0.79. Moreover, 39.89% of summarized texts have F-Score below 0.5. According to the results, TextTeaser algorithm performance for summarizing text in Indonesian language is still average. There are a lot of rooms for improvement to obtain better text summarization result for Indonesian language.

4. Conclusion

TextTeaser is one of the text summarization algorithms that use extraction method. It selects the best sentences among text by calculating title feature, sentence length, sentence position and keyword frequency. This research evaluates TextTeaser algorithm to summarize text in Indonesian language by using TextRank algorithm. We find that utilizing TextTeaser algorithm to summarize text in Indonesian language still need more improvements. Our future work will modify the text preprocessing step and compare it with the original TextTeaser.

5. References

[1] Massandy D T and Khodra M L 2014 Guided summarization for Indonesian news articles *Advanced Informatics: Concept, Theory and Application (ICAICTA), 2014 Int. Conf. of* 140-5 (IEEE)

[2] Ng J-P, Bsyani P, Lin Z, Kan M-Y and Tan C-L 2011 SWING: Exploiting Category-Specific Information for Guided Summarization *Proc. of the Fourth Text Analysis Conf.* (Gaithersburg)

[3] Silvia, Rukmana P, Aprilia V R, Suhartono D, Wongso R and Meiliana 2014 Summarizing Text for Indonesian Language by Using Latent Dirichlet Allocation and Genetic Algorithm *Int. Conf. on Electrical Engineering, Computer Science and Informatics (EECSI 2014)* 148-153 (Yogyakarta)

[4] Winatmoko Y A and Khodra M L 2013 Automatic Summarization of Tweets in Providing Indonesian Trending Topic Explanation *Procedia Technology* **11** 1027-33 (Elsevier)

[5] Mihalcea R and Tarau P 2004 TextRank: Bringing order into texts *Proc. of EMNLP 2004* 404-11 (Association for Computational Linguistics)

[6] Kohlschütter C, Fankhauser P and Nejdl W 2010 Boilerplate Detection using Shallow Text Features *Proc. of the Third ACM Int. Conf. on Web Search and Data Mining* 441-50 (New York, ACM)

[7] Gunawan D 2015 Protocol for E-Commerce data harvesting *Technology, Informatics, Management, Engineering & Environment (TIME-E), 2015 Int. Conf. on* 1-5 (Samosir, IEEE)

[8] Tala F Z 2003 A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia *Master Thesis* (Netherlands, Universiteit van Amsterdam)

[9] Hu M, Sun A and Lim E-p 2007 Comments-oriented Blog Summarization by Sentence Extraction *Proc. of the Sixteenth ACM Conf. on Conf. on Information and Knowledge Management* 901-4 (Lisbon, ACM)