International Journal of Research
Available at https://pen2print.org/index.php/ijr/
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 05 Issue 21
October 2018

Text Summary Optimization Model Based On Extraction

Jeremia Siregar[1], Tulus[2], Syahril Efendi[3]

[1]Department of Engineering Informatics, Faculty of Computer Sciences and Information Technology, University of Sumatera Utara, North Sumatera
[2]Department of Mathematics, Faculty of Math and Science University of Sumatera Utara, North Sumatera
[3]Department of Engineering Informatics, Faculty of Computer Sciences and Information Technology, University of Sumatera Utara, North Sumatera

Abstract
Text summarization can be classified into two approaches: extraction and abstraction. This paper focuses on extraction approach. The goal of text summarization based on extraction approach is sentence selection. Sentence weighting and then select the best ones. The first step in summarizing by extraction is the identification of important features. In our experiment, the source text that will be summarized is an Indonesian news article obtained from a news website or newspaper in the form of softcopy of text documents. Summarizing the text that is done is a single document summary. Summarizing results in the form of extraction from the source text and not stemming the input text. The method used is Fuzzy Based Method. Evaluation is done by comparing the content overlap between the automatic summary results and the reference summary using the ROUGE (Recall-Oriented Understanding for Gisting Evaluation) evaluation toolkit.
Keywords: Fuzzy Logic; Sentence Feature; Text Summarization

Introduction

Online information can be in the form of news articles, documents, video clips, movies, music, and so on. User needs for information in the form of news articles cause users to spend a long time to read the entire contents of the news article. Therefore, a solution is needed so that the user does not need to read the entire contents of the news article, but the user will still get brief and concise information that represents the contents of the news article. The solution is automatic text summarization, a system that is able to summarize news articles. Summarization for appropriate news articles is applied to a single document because the process is fast (Kumar and Chandra Kala, 2010)

Summary is defined as a text that is produced from one or more source texts containing significant information from the source text and no more than half the source text. Summary can be classified in two categories, extraction and abstraction. Extraction summary is a number of sentences or phrases selected from the source text with the highest value and placed together in a new, shorter text without changing the text content. While the summary of abstraction uses linguistic methods to examine and interpret the text. Most automated text summarizing systems use extraction methods to produce summaries. Summarizing automatic text works optimally on well-structured documents such as a number of articles, news, reports, and scientific papers.

Graph-based summarization algorithm or graphical text-based summarization is a language summarization method that can produce extractive summaries. Textual graph is the source text which is represented as a graph. The construction of

e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 05 Issue 21
October 2018

textual graphs uses the concept of similarity between text units. Vertex in textual graphs can be text units such as words, sentences, or paragraphs in the text. Edge in a textual graph shows the connection between vertices. The connection between vertices can be similarity between sentences or lexical or grammatical relationships between words / phrases. (Thakkar et al, 2010)

Nowdays there are many tools for text summarization, but not many have discussed the quality of the text summary. The summary quality of the text is very important in order to know the importance of a document. When you have to read in a short time, a summary of the text that is relevant to the document is needed. Fuzzy set theory can represent and overcome uncertainty which in this case leads to doubts, inaccuracies, incomplete information, and partial truths. In this study, the author uses fuzzy systems to extract important sentences in text summarization.

In this study, the author will analyze the summary quality between linear merging Graph-Based Summarization Algorithm compared to Graph-Based Summarization Algorithm with fuzzy system seen from precision, recall, and f-measure using the ROUGE Evaluation toolkit. So that from the comparative analysis of these two methods will be known better summary quality.

Statement of the problem

The main purpose of text summarization is to build a summary that maximizes text expressions, reduces excess data and maximizes the linkages between sentences. The main difficulty faced up to now is to produce a model so that an optimization form can be obtained that includes the main objectives of the summary.

Objective of the study

To get an optimization model from a quality web document page in accordance with the contents

Research Methodology

Ranked Positional Weight

Certain locations in the text such as headings, titles, and first paragraphs tend to contain important information. The simple method of taking the first paragraph (lead) as a summary is usually quite good, especially in news articles. Ranked Positional Weight is a method proposed by Helgeson and Birnie as an approach to solving problems in line balance and finding solutions quickly. The concept of this method is to determine the number of minimum work stations and to divide tasks into work stations by assigning position weight to each task so that all tasks have been assigned to a work station. The weight of each task, eg the i task is calculated as the time needed to do the i task plus the time to execute all the tasks that will be executed after the i-task. The sequence of steps in the Ranked Positional Weight method are as follows:

1. Perform position weight calculations for each task. The position weight of each task is calculated from the weight of a task plus the weight of the tasks after that.
2. Order tasks based on the position weight, ie from the weight of the large position to the weight of the small position.
3. Place the task with the largest volume to a work station as long as it does not violate the precedence constraint and the time of the work station does not exceed the cycle time.

e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 05 Issue 21
October 2018

4. Do step 3 until all tasks have been placed on a work station.

Cue phrase indicator criteria

In some genres of text, certain words and phrases in sentences explicitly indicate how important the sentence is. A list of cue phrases along with (positive and negative) 'goodness score' is usually built manually.

Word and phrase frequency criteria

In general, the features used to represent documents in the vector roar model are words. This is because the extraction of words from documents is relatively easy, which is only detecting a row of characters ending in a space. If it is designed that numbers are not part of the word, in Indonesian, special characters that represent words are only hypen ("-"), which indicates the repeated word, other than the alphabetical character. Research for English text involving phrases shows that involving phrases in features can improve clustering performance. Research on the detection and extraction of phrases in English has also been found quite a lot. There are various methods of seleki, starting with a statistical approach to natural language processing (NLP) approaches.

For the case of Indonesian, research in this field is still very limited. In this study, phrases are defined as two words that are close together that have certain meanings that can be different from the meaning of their single words, such as "scapegoat". The word extraction technique is taken in a simple way, which is to calculate the frequency of occurrence of the two-word pair. Furthermore, as in the words after the minimal frequency of occurrence is limited, frequency variance analysis is performed to make the selection.

Query and title overlap criteria

The simple but useful method is to score sentences according to the number of words that also appear in titles, headings, or queries.

Cohesive or lexical connectedness criteria

Lexical cohesion, which is a cohesive effect achieved through the selection of vocabulary. Second, based on the origin of the relationship, cohesion is further classified based on three things, as follows:

1. Relation to forms such as substitution, ellipsis and lexical collocation;
2. Relation to references such as lexical references and reiterations;
3. Semantic relations that are chained by conjunctions.

According to Untung Yuwono in his book entitled Enchantment of Language states that cohesion does not appear by itself, but is formally created by a language tool called cohesion marker, such as pronouns, pointing words, conjunctions, and repeated words. Cohesion markers are used appropriately to produce lexical cohesion and grammatical cohesion. Lexical cohesion is a semantic relationship between constituent elements of discourse by utilizing lexical elements or words with reiteration and collocation. Reiteration is the repetition of words in the next sentence to emphasize that the words are the focus of the conversation. Reiteration can be in the form of repetition, synonymy, hyponymy, metonymy, and antonym. While collocation is the relationship between data in the same field.

For example, [Petani] di Lampung terancam gagal memanen [padi], [sawah] yang mereka garap terendam banjir selama dua hari. While grammatical cohesion is a semantic relationship between elements that are marked by a grammatical tool, which is a language tool used in relation to grammar. Grammatical cohesion can take the form of references, substitutions, ellipsis, and conjunctions. Words can be connected in various ways, such as repetition, coreference, synonyms, and semantic associations in the thesaurus. Sentences and paragraphs can be scored based on the degree of connection of the words; the more connected is assumed to be more important.

Discourse structure criteria

Making a discourse text structure and scoring sentences based on the centrality discourse.

Graph-Based Automatic Text Summarization

Graph-based methods are relatively new in automatic text summarization. This method models the text in graphical form by making text units as vertices and adding edges to the graph based on meaningful relationships between units of text that are used as vertices, then determining the importance of each vertex based on the overall graph structure.

Textual graph is the concept of ranking web pages with pagerank that has been described to be applied to graphs on other domains. Textual graphs are constructed from text. Similar to the purpose of PageRank to rank web pages, the application of textual graph ranking is to rank the text units. From the ranking results, the most important text units that will be extractive summary compilers can be selected.

In ranking textual text, the text is represented as a graph. Vertex / node in textual graph is a text unit that will be ranked, which can be in the form of words, sentences, or paragraphs in the text. The edge / link in the graph shows significant connectivity between vertices / nodes. This connection can be in the form of similarity between sentences or lexical or grammatical relationships between words / phrases.

Selection of the type of text unit to be used as a vertex depends on the purpose of the application to be achieved. For example, for extractionyphrases usually phrases or words become indexes, while for extractive summaries usually sentences or paragraphs are selected as vertices.

Edge that connects vertices is also adjusted to the needs and text units selected. Similarity is usually used to express the relationship of a vertex with another index, or in other words, between a sentence / paragraph one with another sentence / paragraph. The types of similarity that are applied also vary and can be defined by themselves, according to the needs of the summary system to be built, including cosine similarity and simple word overlap.

Application

There are various types of summaries depending on the purpose of the summarization program to create a summary of the text, such as generic summaries or query relevant summaries. The summarization system can make both a summary of the relevant query text and generic engine generated summary depending on what the user needs. Summarization of multimedia documents,

for example pictures or movies can also be possible.

Some systems will produce summaries based on a single source document, while others can use multiple source documents (for example, a group of news on the same topic). There is a piece of text, such as a journal article, and there are results of a list of keywords or keyphrases that capture the main topics discussed in the text. In contrast, the abstractive keyphrase system will internalize content and produce keyphrases that may be more descriptive and more like what humans will produce, such as "political negligence" or "adequate protection from flooding". Note that these terms do not appear in the text and require deep understanding, which makes it difficult for computers to produce these keyphrases. Matches between proposed keyphrases and known keyphrases can be checked after originating or applying some other text normalization.

Unsupervised keyphrase extraction: TextRank

While supervised methods have several advantages, such as being able to produce rules is interpreted for what features keyphrase features, but also requires a large amount of data training. Instead of trying to learn the explicit features that characterize keyphrases, the TextRank algorithm utilizes the text structure itself to determine the keyphrases that appear "center" for text in the same way that PageRank selects important webpages. After the graph is built, it is used to form a stochastic matrix, combined with a damping factor (as in "random surfer models"), and the top rank of the vertices is obtained by finding the corresponding eigenvector to the eigenvalue 1 (i.e., the stationary distribution of the random walk on the graph).

Steemming method

Stemming technique is a search technique for the basic form of a term. What is meant by the term itself is every word that is in a text document. Stemming is done when making an index of a document. Indexing is done because a document cannot be identified directly by an information retrieval (IR) system. Therefore, the document must first be mapped into a representation using the text inside it.

Literature Review

Text Summarization

Text summarization is the process of reducing text documents with a computer program to create a summary that preserves the most important points of the original document. The Extraction method works by selecting parts of an existing word, phrase, or sentence in the original text to form a summary. Instead, the abstraction method builds an internal semantic representation and then uses natural generation language techniques to make summaries that are closer to summarizing manually. The state-of-the-art abstractive method is still quite weak, so most research has focused on extractive methods.

An article that has a long size, will cause readers will be very difficult if you have to read and absorb all the information from the article. Text Summarization will produce a text product that still has / contains important parts of the original article. The test results show that the summary process is very dependent on the type and structure of the article. The system will produce a good summary if the type of article being processed is a scientific type of argumentation. As for the structure of the article, if an article has many paragraphs and each paragraph has more than two sentences, it will get a good summary.

e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 05 Issue 21
October 2018

Whereas according to Hovy, a summary is a text produced from a text or a lot of text, which contains the information content of the original text and is not more than half the original text (Hovy, 2001). Research on automatic text summarization using various methods and approaches, beginning in 1958 by Luhn. Many techniques are used in this summarization, such as statistical approach techniques namely word frequency technique (Luhn, 1958), position in text (Baxendale, 1958), cue words and heading (Edmudson, 1969), sentence position (Lin and Hoovy, 1997) . Inverse term frequency and NLP technique is an approach technique with natural language analysis (Aone, 1990), lexical chain (Mc Keown, 1997), maximal maginal relevance (Cabonell and Goldstein, 1998).

Text Summarization Characteristics

There are two approaches to text summarization, shallower approaches and deeper approaches. In extraction techniques, the system copies the most important or most informative text units from the source text to a summary. The text units that are copied can be the main clause, the main sentence, or the main paragraph. While the abstraction technique involves paraphrasing the source text. Abstraction techniques take the essence of the source text, then make a summary by creating new sentences that represent the essence of the source text in a different form from the sentences in the source text. In general, abstractions can summarize the text more strongly than extraction, but the system is more difficult to develop because it applies natural language generation technology which is a topic developed separately.

Based on the number of sources, a summary can be generated from one source (single-document) or from multiple sources (multi-document). Summarizing the single-document input in the form of a text and the output in the form of a new text that is shorter. In the multi-document summarization, input is some text document that has the same theme, usually already in one cluster then output will be produced in the form of a shorter text summarizing the main information on the input cluster.

A summary can be general, that is, a summary that seeks to extract as much important information as possible that can describe the overall content of the text. In addition, information can also be obtained for summaries based on queries that are defined by the user-defined system. Queryoriented or user-oriented summarization tries to retrieve relevant information by querying users and displaying it in summary form.

Based on its function, a summary can have indicative, informative, or evaluative properties. Informative summary serves the main or most important information from the source text. The indicative summary provides suggestions for further reading on certain matters in the text content. While summary evaluators comment or evaluate the main information in the source text.

Compression rate in the summarization process will determine the summary length produced. Usually measured by a percentage of the source text, for example a summary of 10%, 25%, or 50% of the source text. Besides that it can also be measured by the number of words, for example a summary of 100 words is specified. Usually, the summary length is not more than half the source text. The picture below shows a high-level architecture of automatic text

summarization. Input in the form of text with various characteristics and outputs in the form of summary extraction and abstraction.

General Method

The method in applying Text Summarization is to use the TF-IDF (Terms Frequency - Inverse Document Frequency) and Exhaustive algorithm methods. The TF-IDF method functions to calculate the value weight of each sentence and relations between sentences. While the Exhaustive algorithm serves to generate the path of each search for points on the graph. And then the results of the path will be a summary. The Text Summarization method has three general methods, as follows:

1. Extraction-based summarization
2. Abstraction-based summarization
3. Maximum entropy-based summarization

Extraction-based Summarization

The two types of summarization that are often discussed in the literature are keyphrase extraction, which aims to select individual words or phrases to "tag" a document, and summarize the document, which aims to select entire sentences to make short paragraph summaries. In 2012, Light Filtering, one of the methods used to summarize sentences from documents assessed towards their main content, showed good results for using summarization pre-processing steps before extraction of keyphrase.

Abstraction-based Summarization

Extraction techniques only copy information that is considered to be most important by the system for summary (for example, key clauses, sentences or paragraphs), while abstraction involves paraphrasing parts of the source document. In general, abstractions can shorten text stronger than extraction, but programs that can do this are more difficult to develop because they require the use of natural language generation technology. While some processes have been carried out in abstractive summarization (creating abstract synopsis manually), most summarization systems are extractive (choosing a subset of sentences to put in summary).

Maximum Entropy-based Summarization

Although automating abstractive summarization is the goal of summarization research, the most practical system based on several forms is extractive summarization. Maximum entropy-based summarization has been applied to summarization in the news broadcast domain.

Evaluation and Results

Sentence Location

In general the initial sentence in a document is very important. So it can be made the first sentence of the document is 1.0, the second sentence is 0.8, and so on. While the fifth sentence goes to or is 0.0

Relative Length of Sentence

It is assumed that longer sentences provide more information. For s sentence in document d, the score is calculated by:

$$F2(d,s) = \frac{Sentence\ Length\ of\ s\ (s)}{Maximum\ length\ of\ h\ (si)} \quad i = 1, \dots ns$$

Where *ns* is the number of sentences.

Mean of Frequency Word Distribution

Frequency Word Distribution Size (FK) is based on assumptions:

i. The importance of the word for a document is directly proportional to its availability in the document.

International Journal of Research
Available at https://pen2print.org/index.php/ijr/
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 05 Issue 21
October 2018

ii. Document length does not affect the importance of the word. The value of Frequency Word Distribution for a t in document d is calculated by:

$$F3(d,t) = \frac{Frequency\ Word\ Distribution\ (d,t)}{Max\ Frequency\ Word\ Distribution\ (d,ti)} \quad i = 1,..,nt$$

where nt the number of words in the document.

Mean of Frequency Word Distribution - 1KD

Frequency Distribution – Inverse document distribution (FK-1KD) requires one additional assumption, as follows:

iii. A word in a document that rarely appears in a document is more important than a word that often appears in a document. For a word t in document d given a c collection, the FK-1KD value is calculated as follows:

$$F4(c,d,t) = F3(d,t)Log\frac{nd}{df(c,t)}$$

Where nd is the total number of documents in the text and the frequency of the df document says the number of documents where the word is distributed

Title Similarity

This feature views vocabulary between sentence and document title, which is calculated by:

$$F5(d,s) = \frac{|word-s\ \cap word-t\,|}{|word-s\ \cup word-t|}$$

Where the words s and t are words that are distributed consecutively in sentence s and title t.

Keywords

This feature and the next two features assume that sentences distributed in certain types of items contain important information about documents. This feature calculates the number of keywords in a sentence.

$$F6(d,s) = \frac{Keywords\ (s)}{Length\ of\ sentence\ (s1)}$$

Name Entity

The name entity is calculated by:

$$F7(d,s) = \frac{Name\ entity\ (s)}{Length\ of\ sentence\ (s)}$$

Numeric data

This feature calculates the number of numeric words in a sentence

$$F8(d,s) = \frac{Numeric\ data\ (s)}{Length\ of\ senctence\ (s)}$$

Centrality Sentence

This feature measures vocabulary between a sentence and other sentences in a document. This indicates the importance of a document calculated by:

$$F9(d,s) = \frac{word-c}{nt}$$

The c is the number of common words that occur in sentences and sentences d.

Conclusion

In this paper, we have presented a fuzzy logic aided sentence extractive summarizer that can be as informative as the full text of a document with better information coverage. A prototype has also been constructed to evaluate this automatic text summarization scheme using as input some news articles collection provided by DUC2002. We extracted the important features for each sentence of the document represented as the vector of features consisting of the following elements: title feature, sentence length, term weight, sentence position, sentence to sentence similarity, proper noun, thematic word and numerical data We have done experiments with 125 data set, comparing our

e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 05 Issue 21
October 2018

summarizer with Microsoft Word 2007 and baseline using precision, recall and f-measure built by ROUGE. The results show that the best average precision, recall and f-measure to summaries produced by the fuzzy method. In conclusion, we will extend the proposed method using combination of fuzzy logic and other learning methods and extract the other features could provide the sentences more important.

Reference

[1] Aristoteles, Hardiyeni, Ridha, and Adisantoso. 2012. Text Feature Weighting for Summarization of Documents in Bahasa Indonesia Using Genetic Algorithm. International Journal of Computer Science Issues 9(1):1-6.

[2] Bawden and Robinson. 2009. The Dark Side of Information: Overload, Anxiety and Other Pathologies. Journal of Information Science 35(2):180-191.

[3] Berker and Gungor. 2012. Using Genetic Algorithms with Lexical Chains for Automatic Text Summarization. ICAART 1:595-600.

[4] Budhi, Intan, Silvia, and Stevanus. 2007. Indonesian Automated Text Summarization. Proceeding ICSIIT 2007.

[5] Gholamrezazadeh, Salehi, and Gholamzadeh. 2009. A Comprehensive Survey on Text Summarization System. Proceedings of CSA 9:1-6.

[6] Gong and Liu. 2001. Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis. Proceedings of The 24th International ACM SIGIR Conference on Research and Development in Information Retrieval hlm. 19-25.

[7] Gupta and Lehal. 2010. A Survey of Text Summarization Extractive Techniques. Journal of Emerging Technologies in Web Intelligence 2(3):258-268.

[8] Jezek and Steinberger. 2008. Automatic Text Summarization (The State of The Art 2007 and New Challenges). Znalosti 2008 hlm. 1-12.

[9] Jurafsky and Martin. 2006. Speech and Language Processing: An Introduction To Natural Language Processing, Computational Linguistics, And Speech Recognition 2nd Edition. New Jersey: Pearson Prentice Hall.

[10] Kumar and Salim. 2012. Automatic Multi Document Summarization Approaches. Journal of Computer Science 8(1):133-140.