Automatic Indonesian Text Summarization Using Vector Space Model

Haris Dwi Arfianto,[a] and Sukmawati Nur Endah [a,*]

[a]*Department of Computer Science/Informatics, Faculty of Science and Mathematics, Universitas Diponegoro, Semarang, Indonesia*

**Corresponding author Email: sukma_ne@undip.ac.id*

Received on 5th September, 2016, Accepted on 13th September, 2016

Summary provides information from a text briefly to reduce the amount of effort in understanding the text. However, with the number of the existing text data, creation of a summary manually can take a long time. This process can be easily done with automatic text summarization based on Indonesian language. Automatic text summarization can be done by determining the value of similarity between sentences. This similarity value can be determined based on the vector angle sentence on the vector space model. Some sentences with the highest similarity value chosen as a representation of the text summaries. Based on testing that used data from 25 text consisted of argument, description, exposition, narration, and persuasion text that resulted in an average value of precision, recall and F-Score each 0.55, 0.49, 0.51 from expert interviewees and 0.55, 0.48, and 0.50 from common user interviewees. The summary had the same informativeness level with the expert interviewee's summaries. Precision with the highest result obtained from argument text with an average precision of 0.52 from expert interviewees and 0.46 from common user interviewees.

Keywords: Automatic text summarization, Vector Space Model

1. Introduction

A text has lots of information contained therein. This information can be obtained from a variety of sources ranging from news, scientific papers and books. The information contained in the article is not completely essential. Presentation of the text consists of a main line which is at the core of an explanatory text and a sentence is a complement of the text. Information will be more easily accepted if obtained directly from the main phrases in a summary form. The summary can be interpreted as a text resulting from one or more text contains important information from the source text with a length not exceeding half of the source text. Summary text can be presented in two forms, namely extractive and abstractive [1, 2].

Summary written properly can reduce the work in understanding the text a lot. However, with the number of the existing text data, creation of a summary manually can take a long time. So, we need a system that can create summaries automatically. The summary process is done by utilizing the information retrieval system. The text consists of several paragraphs will be processed to obtain a summary of the results automatically.

The text summarization application began to develop in 1958 [3]. The English text summarization has been developed using several methods such as scoring sentence, cluster-based, LSA, fuzzy logic, vector space model and others [4]. The Indonesian text summarization was developed by some methods such as sentence scoring, cluster-based, and LSA too. The Indonesian text summarization has not been develop by using the vector space model, whereas the English text summarization using vector space model has been shown to produce a level of accuracy of 57.86% [5]. This value does not indicate the value of high accuracy, but the Indonesian and English have a different structure. This difference is the underlying research to the device using vector space model in Indonesian text in extractive summary form.

Vector space model is a model that presents documents in a vector space. One of the advantages is adaptable to the weighting method. These advantages result in the process of looking for similarities between sentences can be more easily done with a weighting method used. The process automatically performed by comparing each sentence in the text. Some sentences that have the highest similarity value are taken to be used as a summary of the text.

2. Experimental Details

Summarization using the vector space model means modeling the sentences in a vector space. The text summarization process consists of four processes, namely the pre-processing (tokenization, stopwords removal, stemming), weighting, the similarity calculation and determine of the summary sentence. Flowchart of the text summarization is shown in Figure 1. The overall process in summarizing the points described below.

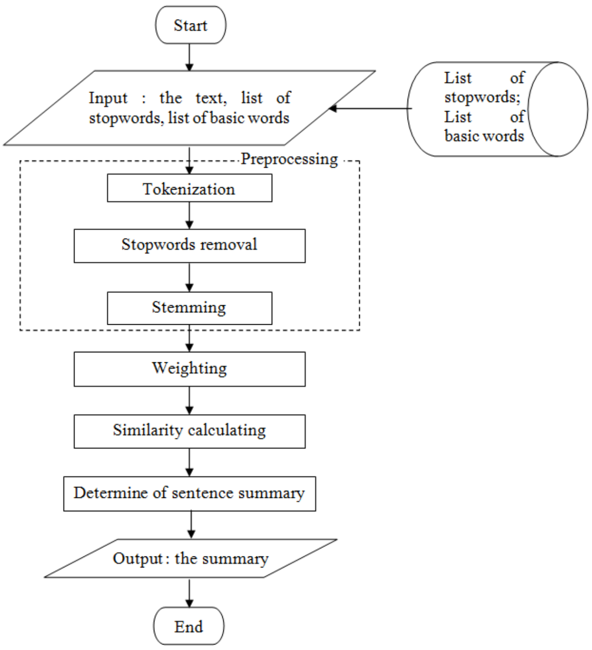

Figure 1. Flowchart of the Text Summarization

2.1. Preprocessing

Preprocessing or the indexing process is divided into several stages i.e. tokenization, stopwords removal and stemming.

1. Tokenization

   Tokenization is the process of cutting the word of a sentence in a paragraph into the form of a single word. A result of this process is only a word in lowercase with no additional punctuation, and other characters.

2. Stopwords removal
   Stopwords defined as words that are not related to the main subject of the sentence. Stopwords can be conjunctions or other words that does not have its own meaning. The word has a frequency of more than 80% in the text does not have the benefit of information retrieval [6]. The word is a stopword that do not contain important information. So stopword need to be eliminated from the process of indexing so as not to interfere with the process of indexing.
3. Stemming
   Stemming is the process of transforming the words in a text document into form the basis word. Stemming algorithms for one language differs from stemming algorithms for other languages. For example, the English language has a different morphology with Indonesian thus stemming algorithms for both languages are also different [7].
   Indonesian has a complicated language structures with some kind of word prefixes, so selecting the right algorithm will also affect the process of indexing up to the acquisition of information. In this research, we use Indonesian stemming algorithm to be made by Bobby Nazief and Mirna Adriani often had known as Nazief & Adriani algorithm [7].

2.2. Weighting

Weighting is done by a tf.idf term weighting [8]. Every word in sentences was weighted to be stored in a sentence vector. Weight of words can be calculated using the formula (1) and (2). w_{ij} is a weight value word i in sentence j, tf_{ij} is the frequency of word i in sentence j, idf_i is the inverse document frequency of word i, N is the number of sentences in the text, and df_i is the frequency sentence containing the word i.

$$w_{ij} = tf_{ij} * idf_i \qquad (1)$$

$$w_{ij} = tf_{ij} * \log\frac{N}{df_i} \qquad (2)$$

2.3. The Similarity Calculation

Similarities are determined using the cosine similarity as in the formula (3). w_{ij} and w_{iq} is the weight of the sentence j and the sentence q.

$$\mathrm{sim}(d_j, d_q) = \frac{d_j . d_q}{|d_j| . |d_q|} = \frac{\sum_{i=1}(w_{ij} \times w_{iq})}{\sqrt{\sum_{i=1}(w_{ij})^2 \times \sum_{i=1}(w_{iq})^2}} \qquad (3)$$

2.4. Determine of The Summary Sentence

Next process calculates the total value of similarity of each sentence by summing the value of similarity one sentence with other's sentence. Then it sorts based on the total value of the greatest similarity. Sentence summary will be taken the top 41% of the number of existing text sentences. Rated 41% taken from compression ratio of the summary results of the expert.

3. Results and Discussion

3.1 Result

Results summary of the system was tested with several scenarios, testing the level of informativeness for the first scenario and testing accuracy by calculating the degree of precision, recall and F-Score for the second scenario.

1. The First Scenario

The level of informativeness calculated based on data number of the summary by system and the summary of expert person. The data are calculated using SPSS program by comparing the average similarity between data systems with any number of summary data from a summary of the expert speakers. The computing generates the group statistical data and the independent test results data as shown in Table 1 and Table 2.

Table 1. The Independent Test Result Data

Data	Equal variances	Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Expert1	A	.330	.568	-.090	48	.928	-.08000	.88566	-1.86075	1.70075
	NA			-.090	47.057	.928	-.08000	.88566	-1.86167	1.70167
Expert2	A	.845	.363	.990	48	.327	.92000	.92973	-.94935	2.78935
	NA			.990	45.764	.328	.92000	.92973	-.95171	2.79171
Expert3	A	.922	.342	1.258	48	.214	.96000	.76306	-.57424	2.49424
	NA			1.258	46.854	.215	.96000	.76306	-.57521	2.49521

Note: A = Assumed; NA = Not Assumed

Table 2. The Group Statistical Data

System	N	Mean	Std. Deviation	Std. Error Mean
Expert1	1	25	6.8000	2.90115
	0	25	6.8800	3.34564
Expert2	1	25	6.8000	2.90115
	0	25	5.8800	3.63226
Expert3	1	25	6.8000	2.90115
	0	25	5.8400	2.47790

2. The Second Scenario

This test is done by calculating the value of precision and recall and F-Score. The data used for this test was 25 texts, consisted of argument, description, exposition, narration, and persuasion text. Based on the value of all the text calculated values of precision, recall, and F-Score of each type of text that can be seen in the graph in Figure 2 to the data of experts and resource persons in Figure 3 is based on the data sources common users.

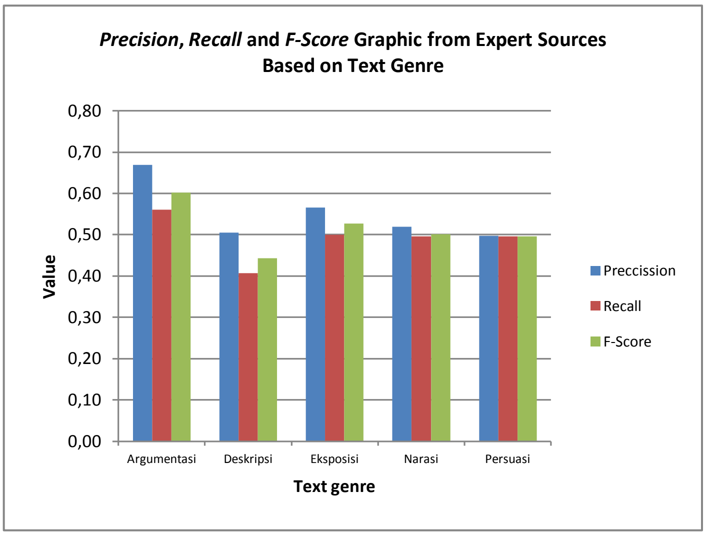

Figure 2. *Precision*, *Recall* and *F-Score* Graphic from Expert Sources Based on Text Genre

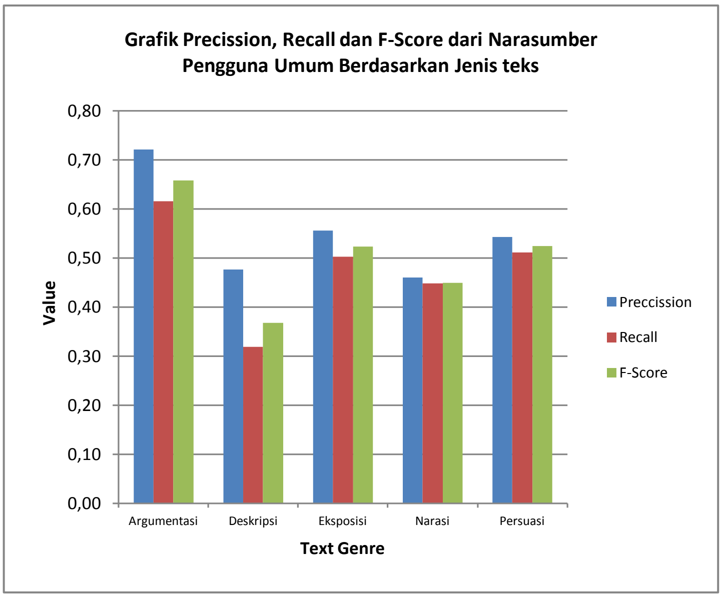

Figure 3. *Precision*, *Recall* and *F-Score* Graphic from Common User Sources Based on Text Genre

3.2 Discussion

H_0 value for the first expert is acceptable because the t value is greater than t table value, i.e. -0.090> -1.677. H_0 value for the second expert is acceptable because the value of t is greater than t table value, i.e. 0.990> -1.677. H_0 value for the third expert is acceptable because the value of t is greater than t table value, i.e. 1.258> -1.677. The test results on the first of scenario for each expert is H_0 value received, so the difference between the summary of the expert and by system does not have a significant difference. Informativeness level summary of the application have the same level of informativeness with ideal summary.

The second scenario is based on the results of expert sources of data obtained an average value of precision, recall and F-Score each of 0.55, 0.49, and 0.51. While based on the data sources common users gained an average value of precision, recall and F-Score each of 0.55, 0.48, and 0.50. The resulting value is quite low because the text used all kinds and of each speaker gives a different summary. Each sample text is used to produce different accuracy results even though some of the text is the same type of text.

Type text with the highest precision value is the text argument with an average value of expert user precision of 0.67 and 0.72 of the general users. Otherwise the value of the lowest accuracy is a narrative text with an average value of precision of 0.52 and 0.46 of expert speakers from general users.

4. Conclusion

Based on data from Indonesian expert, the summary by system has the informativeness value equal to the ideal summary. Results of application development the automatic Indonesian text summarization generate an average value of precision, recall and F-Score each of 0:55, 0:49, and 0:51. Based on the data sources common users gained an average value of precision, recall and F-Score each of 0:55, 0:48, and 0:50.

The results tend to be low, but there are some values that indicate high value on some text. These results are obtained due to each individual has a different way of determining a summary of the text, so that the summary results can vary. Type the text that has the highest accuracy values is the text of the argument that this application is more suitable to do a summary on the argument text.

References

1. Hovy, E., *Automated Text Summarization. In R. tkov Handbook of computation linguistics*, Oxford University Press, England (2001).
2. Pimpalshende A, N., Overview of Text Summarization Extractive Techniques. *International Journal Of Engineering And Computer Science*, (2013) 1205-14.
3. Luhn, H.P., The Automatic Creation of Literature Abstracts. IBM Journal of Research Development 2, (1958) 159-65.
4. Gupta, V. and Lehal, G.S., A Survey of Text Summarization Extractive. *Journal of Emerging Technologies in Web Intelligence*, (2010) 258-68.
5. Kageback, M., Mogren, O., Tahmasebi, N. and Dubhashi, D., Extractive Summarization using Continuous Vector Space Models. *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality*, (2014) 31-39.
6. Yates, R.B. and Neto, B.R., *Modern Information Retrieval, Addison Wesley-Pearson international edition*. Boston, USA, (1999).

7. Agusta, L., Perbandingan Algoritma Stemming Porter Dengan Algoritma Nazief & Adriani Untuk Stemming Dokumen Teks Bahasa Indonesia. *Konferensi Nasional Sistem dan Informatika 2009*, (2009)196-201.
8. Lee D, L., Chuang, H. and Seamons, K., Document Ranking and the Vector-Space Model. *Journal IEEE Software,*14 *(*1997) *2*, 67-75.