Relevance Vector Machine Optimization in Automatic Text Summarization

K E Dewi, E Rainarli
Informatics Engineering Department, Universitas Komputer Indonesia, Jl. Dipatiukur 112-116 Bandung, Indonesia
Email: kania.evita.dewi@email.unikom.ac.id

Abstract. This study aims at optimizing the Relevance Vector Machine (RVM) algorithm in automatic text summarization. This research begins by studying various studies on automatic text summarization to find out what features are commonly used in the automatic text summarization process. Each feature value will be calculated as a correlation with target. The composition of features is determined by obtained correlation value, when correlation value between features and targets is greater, the feature will take precedence. The results in this study are obtained by using 4 or 6 features that obtains highest accuracy, which is 55.84%. The conclusion of this study is that the correlation coefficient can be used to determine the order of extraction features.

1. Introduction
Automatic text summarization is a process to shorten a document or more by computer program [1][2]. Documents circulating in cyberspace are very numerous, therefore it requires a lot of time to read these documents. Therefore, the need for automatic text summarization is very high. The approach to making automatic text summarization that is often used is the extractive approach [2][3]. Extractive approach is the selection of several sentences from the original document. There are numerous research in the field of automatic summarization, one of which is research conducted by Fitriaman [4]. Fitriaman’s research is aimed to make automatic text summarization using the Support Vector Machine (SVM) method. The accuracy obtained from the results of this study is 77,25%. However, in Tipping studies [5][6] it is stated that errors produced by Relevance Vector Machine (RVM) are smaller than SVM. This statement was reinforced by studies of Xu [7], Xin [8], and Rafi [9] which resulted in RVM’s accuracy being better than SVM. Then in the previous research automatic text summarization was carried out using the RVM algorithm [10]. However, the accuracy result were 63.084%.

One component that is considered influencing the results of summarization is the selection of features. The extraction feature, first offered by Lunh, was the use of word appearance [11]. Then, in Baxendale’s research states the selection of core sentences based on position and key word [12]. In the Edmundson study extraction features used cue method, location method, key method, title method [13].

Automatic text summarization researches have now combined sentence extraction features. As in Suanmal’s research, the features used are title features, sentence length, term weight, sentence positions, sentence to sentence similarity, proper noun, thematic word and numerical data [14]. In Zaman Research the features used are log TF-ISF, sentence location, sentence overlap, title overlap, sentence relative length [14]. Whereas in Deni Fitriaman’s research the extraction features used are sentence length, sentence position, numerical data, thematic words, title feature, sentences to sentences, bushy path before, and bushy path after [4]. But none of these studies have shown the use of any combination of extraction features that are most optimal in automatic text summarization using the RVM algorithm.
Therefore, this study will show a combination of extraction features which are optimal for automatic text summarization using the RVM algorithm. The correlation coefficient between extraction feature and target will determine sequence of features in the combination. The order of correlation from the largest to the smallest indicates the combination of extraction features to be used. Based on the results of testing, it is found that there is no need to use all combinations of accuracy features of text summarization in obtaining the best accuracy.

2. Method
This study aims at determining the combination of optimal extraction features in carrying out automatic text summarization using the RVM algorithm, so this study begins with a study of literature. Literature study is conducted to find out what extraction features are commonly used in automatic text summarization. After getting extraction features, then extraction feature of correlation coefficient value and target class (summary sentence or not) is calculated. The correlation coefficient will be used as a sequence of combinations of extraction features. The greater the correlation coefficient value, the more important the feature to be considered than the others. Each result of extraction features combination will be used on automatic text summarization then the accuracy result is calculated.

3. Results and Discussion
Based on the results of the literature study, extraction features commonly used in automatic text Summarization are:

a. Sentence length
   This feature is used to eliminate short sentences [4] [14] [16]. Calculating the length of the sentence is done by dividing the number of words in the sentence by the number of words in the document.

b. Sentence position
   This feature is used to indicate the position of the sentence. The initial sentence of the document tends to have a higher weight than at the end of the sentence [4] [15].

c. Numerical data
   This feature is used to show the number of numerical data contained in a sentence. Usually sentences that contain numeric data are important sentences and those sentences will be included in summaries [4].

d. Thematic word
   This feature is used to search for words related to the topic [4] [14] [16]. In this study 10 words appear in a document.

e. Title feature
   This feature is used to find the sentence that most relates to the document title [4] [13] [14] [15] [16]. The sentence that resembles the title is calculated by the ratio of the number of words in sentence that occur in the title over the number of words in title.

f. Sentence to sentence
   This feature is used to see similarities between sentences [4] [14] [16]. The similarity between sentences is calculated by the ratio of the number of words in sentence that occur in other sentence over the number of words in document.

g. Bushy path before
   This feature is used to view lexical bonding with the previous sentence [4]. The lexical bond with the previous sentence will be 1 if there is a lexical relationship if it is 0 if there is no lexical relationship.

h. Bushy path after
   This feature is similar to the features of a lexical bond with the previous sentence. This feature will be worth 1 if there is a lexical relationship with the following sentence and is worth 0 if there is no lexical relationship [4].

i. negative keyword
   This feature is the opposite of the thematic sentence feature. In this feature, 10 words are not related to the topic. [16] [17].
j. Connection between sentences
   This feature is used to see the number of links from a sentence that connects with other sentences. 
   This feature can be calculated by counting the number of sentences that have the same word with 
   other sentences in one document divided by a number of sentences in a document [17].

k. The weight of the connection between sentences
   This feature is used to see the relationship of a sentence with all other sentences i 
   n a document. This 
   feature can be calculated by summing the same words with other sentences in a document divided 
   by many words in the document [17].

l. Term frequent
   This feature is used to give weight to the word. In this study this feature is calculated using the TF-
   IDF formula [18].
   In this study, to determine the order of combination of features a correlation coefficient (r) between 
   the value of features and targets is used. Correlation coefficients can describe the relationship 
   between 2 variables [19]. The correlation coefficient can be calculated using formula 1:

\[ r(X, Y) = \frac{\sum(x-x)(y-y)}{\sqrt{\sum(x-x)^2}\sqrt{\sum(y-y)^2}} \]  

Where \( X \) is mean of X. The results of the correlation calculation between features and targets can be 
seen in table 1. Based on Table 1, the sequence F3, F2, F7, F5, F9, F10, F8, F4, F1, F6, F12, F11. Then 
based on this sequence, combinations of extraction features are arranged.

Table 1. The results of the calculation of the correlation coefficient between the value of the 
feature and the target

| Feature                      | Code | r     |
|------------------------------|------|-------|
| Sentence length              | F1   | -0.1352|
| Sentence position            | F2   | 0.0577 |
| Numerical data               | F3   | 0.0763 |
| Thematic word                | F4   | -0.1261|
| Title feature                | F5   | -0.0514|
| Sentence to sentence         | F6   | -0.1355|
| Bushy path before            | F7   | 0.0254 |
| Bushy path after             | F8   | -0.1064|
| Negative keyword             | F9   | -0.0838|
| Connection between sentence  | F10  | -0.0833|
| The weight of connection     | F11  | -0.1565|
| between sentence             |      |       |
| Term Frequency               | F12  | -0.1409|

The combination extraction features are then used in summarization. The accuracy of the summary 
results can be seen in Table 2. It can be seen in Table 2 that the highest accuracy is obtained by using 
combination 4, namely numerical data, sentence position, bushy path before, and title feature, with 
55.84% of accuracy. Combination 6 also produces the same accuracy of 55.84%, where the feature used 
is a combination of 4 features plus negative keyword and connection between sentence. The results of 
this study are different from the result of Zulfikar’s research [17] which says accuracy is obtained by 
using eight extraction features. Those eight extractions are sentence position, sentence to sentence, title 
feature, Named entity Recognition, numerical data, sentence length, connection between sentence, and 
the weight of the connection between sentence. Zulfikar’s research is not mentioning the algorithm used 
in making automatic text summaries. The results of this study have not been better than before, because 
it is expected that the accuracy of the summarization results in this study is better than Fitriaman research 
[4], because it has used the RVM method.
Table 2. Results of testing the combination of features with the RVM algorithm

| No | Combination | precision | recall  | f-measure | Accuracy |
|----|-------------|-----------|---------|-----------|----------|
| 1  | 2           | 0.5405    | 0.5333  | 0.5369    | 0.5306   |
| 2  | 3           | 0.5355    | 0.5287  | 0.5321    | 0.5260   |
| 3  | 4           | 0.5677    | 0.5605  | 0.5641    | 0.5584   |
| 4  | 5           | 0.5484    | 0.5414  | 0.5449    | 0.5390   |
| 5  | 6           | 0.5677    | 0.5605  | 0.5641    | 0.5584   |
| 6  | 7           | 0.5290    | 0.5223  | 0.5256    | 0.5195   |
| 7  | 8           | 0.5355    | 0.5287  | 0.5321    | 0.5260   |
| 8  | 9           | 0.5613    | 0.5541  | 0.5577    | 0.5519   |
| 9  | 10          | 0.5613    | 0.5541  | 0.5577    | 0.5519   |
| 10 | 11          | 0.5290    | 0.5223  | 0.5256    | 0.5195   |
| 11 | 12          | 0.5484    | 0.5414  | 0.5449    | 0.5390   |

4. Conclusion
Based on the Results and Discussion, the accuracy has not been improved by using features, but the calculation of the correlation of features to targets can be used to sequence which features to be used. By using 4 features, namely numerical data, sentence position, bushy path before, and title feature obtained the highest accuracy of 0.5584. Accuracy is also obtained when the feature is added 2, namely negative keyword and connection between sentence.

References
[1] Manuel J, Moreno T. 2014. *Automatic Text Summarization* (Great Britain: ISTE Ltd and John Wiley & Sons, Inc.)
[2] Antunes J, Dueire R L, Lima R, Oliveira H, Riss M, Simske S J. 2018. *Computer Speech & Language* 55 (1), pp. 141-64
[3] Fang C, Mu D, Deng Z, Wu Z. 2016. *Expert Systems with Applications* 72 (1), pp. 189-95
[4] Fitriaman D, Khodra M L, Trilaksono, B R Peringkasan Teks Otomatis Berita Berbahasa Indonesia Pada Multi-Document Menggunakan Metode Support Vector Machines (SVM) [Online] [Accessed 9 March 2019] Available: https://docplayer.info/36453812-Peringkasan-teks-otomatis-berita-berbahasa-indonesia-pada-multi-document-menggunakan-metode-support-vector-machines-svm.html
[5] Tipping M E. 1999. *Advance in Neural Information Processing System* 12 (1), pp. 652-8
[6] Tipping M E. 2001. Sparse Bayesian Learning and the Relevance Vector Machine *J. of Machine Learning Research* 1(2), pp. 211-44
[7] Xiang-min X, Yun-feng M, Jia-ni X. 2007. *Int. Workshop on Anti-Counterfeiting, Security and Identification (Xiamen)* (Xiamen: IEEE) 2 (1), pp 208-11
[8] Rekha, S. N., Jeyanthy, P. A., & Devaraj, D. 2019. Relevance vector machine based fault classification in wind energy conversion system. International Journal of Electrical and Computer Engineering, 9(3), pp. 1506.
[9] Kumar, N. P., & Chandra, A. P. 2019. Development of Remote Instrumentation and Control for Laboratory Experiments using Handheld Devices. International Journal of Online Engineering, 15(9), pp. 1-12
[10] Rainarli E, Dewi K E. 2018. *Proc of INCITEST (Bandung)* 407 (Bandung: IOP Conference Series) p.012075
[11] Yasunaga, M., Kasai, J., Zhang, R., Fabbri, A. R., Li, I., Friedman, D., & Radev, D. R. 2019. Scisummnet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks. *In Proceedings of the AAAI Conference on Artificial Intelligence*. 33, pp. 7386-7393.
[12] Baxendale P B 1958 Machine-Made Index For Technical Literature-an Experiment *J. IBM Journal of Research and Development* 2(3), pp.354.
[13] Edmundson H P 1969 New Method in Automatic Extracting *J. of the ACM* 16 (3), pp.264
[14] Suanmali L, Salim N, Binwahlan M S 2008 Automatic Text Summarization Using Feature-Based Fuzzy Extraction *J. of Teknologi Maklumat* 20(2), pp.105
[15] Zaman B, Winarko E.2011. Analisis Fitur Kalimat untuk Peringkas Teks Otomatis pada Bahasa Indonesia *J.IJCCS* 5(1), pp.60
[16] Fattah M A, Ren F. 2008. Automatic Text Summarization *J. World Academy of Science, Engineering and Technology* 37(2), pp.192.
[17] Zulkifli, Wibowo A T, Septiana G. 2015. *E-Proceeding of Engineering (Bandung)* 2 (Bandung: Universitas Telkom)
[18] Dewi K E, Sagala R E 2018 *International Conference on Informatics, Engineering, Science and Technology (Bandung)* 407 (Bandung: IOP Conference Series) 012157
[19] Boediono, Koster W. 2008. *Teori dan Aplikasi Statistika dan Probabilitas* (Bandung: PT Remaja Rosdakarya)