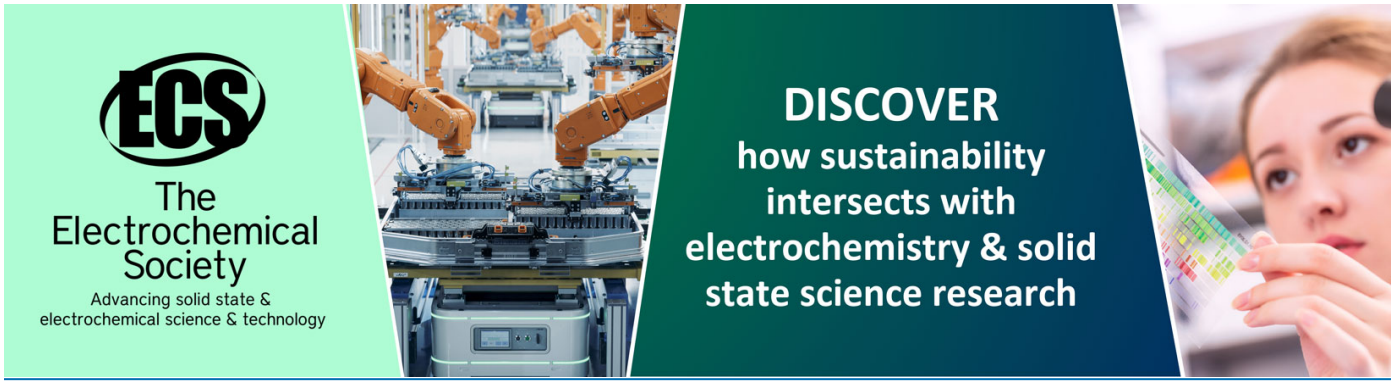
Relevance Vector Machine for Summarization



IOP Conf. Series: Materials Science and Engineering **407** (2018) 012075 doi:10.1088/1757-899X/407/1/012075

Relevance Vector Machine for Summarization

E Rainarli* and K E Dewi

Department of Informatics Engineering, Universitas Komputer Indonesia, Jl Dipatiukur 112 -116 Bandung, Indonesia

*ednawati.rainarli@email.unikom.ac.id

Abstract. This research aimed at finding relevances Vector Machine for summarization.The needed of producing an automatics text summarization create the research of text summarization continues to develop. One way to create an automatic summarization is by choosing the sentences which contain the main topics and reassembled them into a summary. The usage of Supports Vector Machine method (SVM) able to select summary sentences. The Relevance Vector Machine (RVM) appears as a further development of the SVM. This method performs a good result in a classification of Magnetic Resonance Imaging (MRI) data. Therefore, in this research, it examined the ability of RVM in the text summarization. Extracting the sentences used eight features, they are the length of the sentence, the sentence position, the containing of numerical data, the thematic words in the sentence, the similarity of the title, the sentence similarity, the sentence lexical cohesion before and after. There are 1509 training sentences and 214 testing sentences from 100 text documents. The result showed that using Radial Basis Function the accuracy of the RVM reached 63.084%. The RVM performance shows a better result than the SVM, 2% higher than the SVM result and uses fewer vector supports.

1. Introduction

Automatic text summarization is the process of making a summary with the aid of computers from a source of text's digital. The growing of online text causes the need for the automatic text summarization increase. Increasing the number of documents make more effort it takes to read and understand the information. The way to summarize text automatically is by extracting or by selecting sentences which contain the main topic and then rearranged them into a summary. Some research uses machine learning to create a summary. They identify which sentences chosen as the candidate of summary sentences. Hirao's research utilized the Support Vector Machine (SVM) to select summary sentences [1]. Hirao used the dataset from the Text Summarization Challenge (TSC) corpus. The selection of sentence summaries based on the ranking of the SVM decision value function. The test results obtained that the SVM produces a better accuracy of the summary algorithm C4.5 and C5.0 on the decision tree [1].

The SVM method has weaknesses, such as the selection of kernel functions must satisfy the Mercer's condition, the number of support vectors will increase linearly with the increasing amount of training data used [2]. Therefore, Tipping proposed the Relevance Vector Machine (RVM) method to overcome these weaknesses. The RVM algorithm works on the principle of Sparse Bayesian Learning [3]. Xiang-min [4] has compared the performance of both SVM and RVM on the heart scale data classification, breast cancer, Boston, Wdbc. The result shows that RVM requires fewer support vectors

than SVM. When the testing data added gradually, then the error of the RVM is always smaller than the SVM. It also shows that the time's requirement of training dataset using SVM is shorter than RVM. Some studies of RVM were classifying the coffee data [5], detecting arrhythmia [6], epilepsy [7], and recognizing a silent speech [8]. All these studies have shown that RVM performance is better than SVM. Matsumoto added that the results obtained RVM better than SVM when the amount of training data much [8].

This research used the RVM and the feature of a sentence to extract the information of each sentence. The research about text features was seen in Fattah, Anita and Begum [9, 10, 11]. We used the same features as Fitriaman's research because he has showed the optimal result for summarization in Bahasa Indonesia [12]. There are eight features used in this research. That features are the length of sentences, the position of sentences, numerical data, thematic words in sentences, title word count, sentence similarities, lexical ties of a sentence before and after. Fitriaman proofed that eight of these features influence the quality of automatic text summarization. The accuracy of each feature is greater than 58% [12]. Therefore, this study will measure the summaries resulting from the implementation of RVM and the usage of extraction features.

2. Method

This research uses a quantitative approach. The stages in the study were:

2.1. Literature study and formulation of a problem

The authors studied papers that relating to problems in automated text summarization, learned the machine learning methods used in summarization, found the extraction features used in determining sentence of the summary. At the end of this process was formulating the problems to be resolved in the study.

2.2. Collecting data set

Testing used 100 documents dataset. The dataset was the introduction of a thesis in .txt format. After collecting the data set, two of linguists chose the main sentences of each document. This summary used 50% compression. So, the expert selected half of the document's sentences as the summary sentences.

2.3. Preprocessing and feature extraction

Two things to do at this stage were performing a preprocessing and extracting feature of sentences. Preprocessing steps used were case folding, sentence separation, filtering, tokenization, and stopword removal. Feature extraction used were the length of sentences, the position of sentences, numerical data, thematic words in sentences, title word count, sentence similarities, lexical ties of a sentence before and after.

2.4. Learning and testing RVM

After the extraction process, the next step was training data set using RVM. In principle of RVM training was to find the value of the parameters used in the testing process. From the test results obtained summary sentences. The results were compared with a summary that made by experts.

2.5. Conclusion

The summary use accuracy to measure how good the result of the summary and compare it with the SVM result. This part discusses the possibility of kernel function used. The researcher compares the result with the others.

3. Result and discussion

Based on Fitriaman research [12], there are eight features to be used in the extraction process. The features are as follows:

3.1. The length of sentence

In selecting of summary sentences process, it considers the length of sentence. Candidates of the summary are the longest sentence. To calculate this feature, it's a result of dividing the number of words in a sentence against the number of words from the longest sentence. How to calculate the length of sentence given to the equation (1).

$$f_{1j} = \frac{number\ of\ word\ in\ j-th\ sentence}{number\ of\ word\ in\ document} \quad (1)$$

3.2. The position of sentence

This feature assumes the first sentence of each paragraph is the most important sentence. The equation (2) shows how to compute the position of the j th sentence.

$$f_{2j} = \frac{m-j}{m} \quad (2)$$

Where m is the number of sentences in each document, j is the index of a sentence. Index of the first sentence is 0.

3.3. Numerical data

Usually, a sentence that contains numerical data is an important sentence. Equation (3) denotes how to calculate that sentence.

$$\mathrm{f_{3j}} = \frac{\text{a lot of numeric data in the j-th sentence}}{\text{number of words in the j-th sentence}} \quad (3)$$

3.4. Thematic's words in sentences

This feature calculates the relative appearance of a keyword in a sentence. Usually, a sentence with keywords is a summary sentence. Equation (4) is calculate the value of thematic word feature in the j th sentence.

$$f_{4j} = \frac{number\ of\ thematic\ words\ appearing\ in\ the\ j-th\ sentence}{the\ total\ number\ of\ thematic\ sentences\ in\ the\ document} \quad (4)$$

3.5. Title word count

A title-like sentence is a sentence that has a vocabulary overlap between sentences with the title. Equation (5) shows how to calculate the resemblance of jth sentence.

$$\mathrm{f_{5j}} = \frac{\text{number of (word in the j-th sentence} \cap \text{word in title)}}{\text{number of (word in the j-th sentence} \cup \text{word in title)}} \quad (5)$$

3.6. Sentence similarity

The similarity of sentences counts the overlap of vocabulary between sentences with the others. To simplify it, it uses only keywords. Equation (6) shows how to calculate the resemblance of the jth sentence with another sentence:

$$f_{6j} = \frac{number\ of\ (word\ in\ \text{the}\ j-th\ sentence \cap word\ in\ other\ sentences}{number\ of\ word\ in\ document} \quad (6)$$

3.7. Lexical ties of a previous sentence

The lexical tie between the sentence and the previous sentence is the word (stem) that appears in both sentences. If the sentence has a lexical relationship then the value of this feature is 1, otherwise is 0.

3.8. Lexical ties of a next sentence

The lexical tied between the sentence and the next sentence is the word (stem) that appears in both sentences. The same previous feature, the value will be 1 if it has a lexical relationship and 0 if it does not have.

After all the documents extracted, the RVM algorithm uses the result of feature extraction to generate summaries. Tipping shows that the detail of Bayesian Sparse Learning and RVM relationship in classification context [13]. It used the library provided by Mike Tipping to apply the RVM algorithm [14]. Dataset used are 100 of the introduction's part of the thesis. From 100 documents, it

obtains 1723 sentences. The two of linguists asked to summarize the text manually. There are 838 sentences chosen as the sentences of summary and 885 as the sentences that were not summary from 1723 sentences. Table 1 shows the training and the testing data used. There were 1509 sentences used as trainer data. That sentences consist of 733 summary sentences and the other of 776 sentences not selected as summary sentences. The test used 214 sentences. There are 105 sentences of the summary sentences and the other of 109 sentences not selected as summary sentences.

The composition of the data in Table 1 will be used to test the performance of the RVM method in the summary. As a comparison, there is also a summary process using the SVM. The LibSVM package used to implement the SVM method [15]. There are several kernel functions tested in this research, namely: linear, polynomial, Radial Basis Function (RBF), and sigmoid. Performed several tests to obtain the optimal parameter value of each kernel function usage. The optimal value is determined based on the highest accuracy value obtained from the test data. Table 2 shows the details of the optimal parameter values used in each kernel function. The optimal parameter values in the SVM will be used to test the performance of the RVM. Table 2 shows the SVM and the RVM accuracy (See Table 1).

Table 1. Details of training data and tested data.

	Summary Sentences	**Non-Summary Sentences**	**Total Sentences**
Training Sentences	733	776	1509
Tested Sentences	105	109	214

Table 2 shows that both the SVM and the RVM get the highest accuracy when it uses Radial Basis Function Kernel. The RVM accuracy is better than the SVM, except for the polynomial kernel function. The number of support vectors required in the RVM is less than the SVM. For the RBF, the number of support vectors in the SVM is 1191 vectors, whereas for the relevance vector on RVM there are only 21 vectors. This result is consistent with what Tipping has said and Xiang-ming's result [2,4]. The SVM method produces a large support vector when the data is sparse, but not with the RVM method [2,3]. It also indirectly describes the distribution of sentence summaries selected by the expert. The results of this RVM test are same as those done by Lima et. al [7]. They have shown that the use of RBF is suitable for scattered and sparse data.

After comparison of the performance of RVM with SVM for each kernel function is selected, then the next election will be the best parameter values for the use of the RBF. Table 3 showed the rated accuracy of the gamma value changes on the RBF. The highest accuracy reached when gamma was 9. The accuracy of RVM was 63.084% and 27 relevance vectors used. However, the best accuracy shown in table 3 was better than Putra's research [16]. Putra used RVM for the document summary. This study only used TF-IDF in its feature extraction and produce accuracy as much as 53%. While in Arifin's research, multi-documentary summarization using RVM obtained 67% accuracy [17]. Our result is better than Putra's research but not more than Arifin's. Arifin used the feature of entity word of the sentences whereas we used the thematic's word as a feature. This result shows that it needs to evaluate the optimal features of the summary and see the conformity with the RVM method. Although it has yet to show results, this study has shown that the use of RVM allows selecting the sentence of summarization and the performance of RVM consistently better than SVM, especially using RBF Kernel (See Table 2).

Table 2. The performancy result of RVM and SVM.

Kernel Function	Parameter	SVM		RVM	
		Accuracy (%)	Number of Support Vectors	Accuracy (%)	Number of Relevance Vector
$\vec{u}' \cdot \vec{v}$	-	52.333	1284	58.411	7
$(\gamma(\vec{u}' \cdot \vec{v}) + C)^{\text{degree}}$	$\gamma = 8,\ C = 10,\ \text{degree} = 2$	60.280	1151	57.009	8
$\exp(-\gamma \cdot \|\vec{u} - \vec{v}\|^2)$	$\gamma = 8$	61.682	1191	62.150	21
$\tanh(\gamma \cdot \vec{u}' \cdot \vec{v} + C)$	$\gamma = 1/8,\ C = 0$	52.804	1314	58.411	7

Table 3. The performancy of RVM using radial basis function kernel

	$\gamma = 7$	$\gamma = 8$	$\gamma = 9$	$\gamma = 10$	$\gamma = 11$
Accuracy (%)	61.682	62.150	63.084	63.084	63.084
Number of Vector	19	21	27	26	26

4. Conclusion

This research has shown that RVM allows for use in automated text summarization. The results showed the comparison of accuracy between the RVM and the SVM. Based on the testing, the RVM works better than the SVM. It is seen from the accuracy obtained and from the number of support vectors needed to determine the candidate of sentence summarization. The Radial Basis Function remains the first choice for use in the classification process by the RVM method.

Acknowledgements

We thank Heryandi A, M.T. for helping to preprocess dataset and West Java Language Center for validating the test results. This research was part of Internal Research 2017 that supported by Research Institute and Community Service of Universitas Komputer Indonesia.

References

[1] Hirao T, Isozaki H, Maeda E and Matsumoto Y 2002 *Proc. of the 19th Int. Conf. on Computational Linguistics (Taiwan)* **1** (Pennsylvania: ACL) p 1184
[2] Tipping M E 1999 *Conf. Neural Information Processing Systems (Canada)* **12** (Canada: NIPS Fondation) p 652
[3] Tipping M E 2001 Sparse Bayesian Learning and the Relevance Vector *J. of Machine Learning Research* **1** 211
[4] Xiang-min X, Yun-feng M and Jia-ni X 2007 *Int. Workshop on Anti-Counterfeiting, Security and Identification (Xiamen)* (*Xiamen*: IEEE) p 208-211
[5] Wang X, Ye M and Duanmu C J 2009 Classification of Data from Electronic Nose using Relevance Vector Machines *J. Sensors and Actuators B: Chemical* p 140-143
[6] Gayathri S, Suchetha M and Latha V 2012 ECG Arrhythmia Detection and Classification Using Relevance Vector Machine *J. Procedia Engineering* **38** 1333
[7] Lima C A, Coelho A L and Chagas S 2009 Automatic EEG Signal Classification for Epilepsy Diagnosis with Relevance Vector Machines *J. Expert Systems with Applications* **36** 10054
[8] Matsumoto M and Hori J 2014 Classification of Silent Speech using Support Vector Machine and Relevance Vector Machine *J. Applied Soft Computing* **20** 95

[9] Fattah M A and Ren F 2008 Automatic Text Summarization *J. World Academy of Science, Engineering and Technology* **37** 192
[10] Kulkarni A R and Apte M S 2009 An Automatic Text Summarization using Feature terms for Relevance Measure *IOSR J. of Computer Engineering* **9** 62
[11] Begum N, Fattah M A and Ren F 2009 Automatic Text Summarization using Support Vector Machine *Int. J. of Innovative Computing Information and Control* **5** 1987
[12] Fitriaman D, Khodra M L and Trilaksono B R 2011 *Peringkasan Teks Otomatis Berita Berbahasa Indonesia Pada Multi-Document Menggunakan Metode Support Vector Machines* [Theses] (Bandung: Institut Teknologi Bandung)
[13] Tipping M E 2004 Bayesian : An Introduction to Principles and Practice in Machine Learning *Advanced Lectures on Machine Learning* (*Lecture Notes in Computer Science* vol 3176) ed O Bousquet, U V Luxburg and et al (Berlin: Springer) p 41-62
[14] Tipping M E 2006 *Sparse Bayes Software* [Internet] [cited 30 Agustus 2017] Available from: http://www.miketipping.com/downloads.htm
[15] Chang C C and Lin C J. 2011 LIBSVM: a library for support vector machines *J. ACM transactions on intelligent system and technology* TIST **2** 27
[16] Putra A B E 2017 *Implementasi Metode Relevance Vector Machine Dalam Peringkasan Teks Otomatis* [Undergraduate Theses] (Bandung: Digital Library of Universitas Komputer Indonesia)
[17] Arifin T 2017 *Peringkasan Otomatis Pada Multi Dokumen Menggunakan RVM* [Undergraduate Theses] (Bandung: Digital Library of Universitas Komputer Indonesia)