The Implementation of Maximum Marginal Relevance Method on Online National and Local News Portal

Winda Yulita[1] and Feddy Setio Pribadi[2]

[1,2]Computer Science Education Department, Electrical Engineering Faculty, Universitas Negeri Semarang,
Semarang, Indonesia
winda.yulita@gmail.com[1], feddy.setio@gmail.com[2]

***Abstract*—In Indonesia, there are online national and local news portal. The structure of the sentence on the national scale news portal predominantly uses longer sentences. Meanwhile the news on local portal is more compact. Based on that problem, then the researcher developed an automatic text-summarizing system using Maximum Marginal Relevance method and word- weighting system using TF-IDF-DF algorithm in order to produce a well-summarized sentences. The sentence- summarizing process consists of text preprocessing, composed of sentence segmentation, case folding, tokenizing, filtering, dan stemming. The next step is TF-IDF-DF calculation in order to determine the weight of the word and the summarizing process using MMR method. The result showed that in comparison to online local news portal which scored precision 45,83%, recall 45,83%, and f-measure 45,83%, manual sentence – summarizing process scored precision 76,39%, recall 65,28%, and f-measure 70,4% for news document from online national news portal.**

Keywords—sentence summarizing, tf-idf-df, MMR, local and national news portal

I. Introduction

Since 2000, in Indonesia, online media has developed rapidly, it can be seen form the increasing online news portal (Nugroho, et al., 2012). There are online national and local news portal. Antaranews.com is one of the online national news portals whose news covers local, national. And even international events and its distribution area covers all regions (islands, provinces, cities). The national news portal is different to local news. The local news frequently uses compact sentences. Longer sentences have more information compared to the compact ones. Based on that problem, the researcher developed an automatic text-summarizing system using Maximum Marginal Relevance method and word-weighting system using TF-IDF-DF algorithm in order to produce a well-summarized sentences.

II. TF-IDF-DF Algorithm

Term Frequency-Inverse Document Frequency-Document Frequency (TF-IDF-DF) Algorithm is a modified version of Term Frequency-Inverse Document Frequency method (TF-IDF).TF – IDF has lack in terms of word weighting. Its lack is there is an assumption that if the dispersed words in other documents are unimportant, so that they ignore its existence. However, the frequently appearing words on other sentences can be the important ones. Consequently, the words which frequently appears in the document have high weight, but the ones which are dispersed on other documents have lower weight. Therefore, this TF – IDF method is developed further in order to gain representative weight from extracted words by considering their dispersal on other documents. Document Frequency (DF) which contains i-sequence word influences on the topic of the whole document, so that the score of word weigh is multiplied with the DF for the i word (Pramono, 2013). Algorithm equation of TF-IDF-DF can be seen as follows (1).

$$w_{i,j} = \left(tf_{i,j}\ x\ log\left(\frac{N}{df_i}\right)\right)\ x\ df_i \qquad (1)$$

In which $w_{i,j}$ is the weight of the i – sequence from the j – sequence document. $tf_{i,j}$ (term frequency) is the amount of words from the i – sequence word from the j – sequence document. $\log\left(\frac{N}{df_i}\right)$ is the equation of Invers Document Frequency (IDF), N is the amount of the whole document or sentence. df_i (document frequency) is the amount of sentences which contain i – sequence in the collection of document.

III. Maximum Marginal Relevance

According to Carbonell dan Goldstein (1998), the Maximum Marginal Relevance (MMR) summarization technique aims to capture relevant but not redundant information. Maximum Marginal Relevance (MMR) is one of the text extraction documents method which can summarize both single and multiple documents by conducting repeated ranking and comparing similarity inter documents. maximum marginal relevance method (MMR) is used to rank sentences as responses to the user – given query. MMR calculation is conducted with iterations by combining 2 cosine similarity matrix, which are relevance between query to the whole sentence and inter – sentence similarity.

The calculation principle of MMR method is by taking the sentence with the highest score for every iteration calculation. Iteration will stop, if the maximum score of MMR is ≤ 0. The parameter score λ which is used on MMR calculation is $\lambda = 0.7$ (Carbonell dan Godstein, 1998). The MMR calculation process is as follows with the note to notice that $\mathrm{Sim}_1(\mathrm{S_i},\mathrm{Q})$ is relevance query. Meanwhile, $\mathrm{Sim}_1(\mathrm{S_i},S')$ is the sentence similarity to the extracted sentences. The equation can be seen as follows (2).

$$MMR\ (S_i) = \lambda.Sim_1\ (S_i,Q) - (1-\lambda)\ .\ max\ Sim_2(S_i,S') \quad (2)$$

IV. RESEARCH DESIGN

In this research, the news text input in the automatic text – summarizing system is in the form of a single document. Broadly speaking, the process of text summarizing can be grouped into 2 stages, which are text preprocessing and text processing consisting of calculation process of Term Frequency-Inverse Document Frequency-Document Frequency (TF-IDF-DF), cosine similarity and Maximum Marginal Relevance (MMR)

A. *Text preprocessing*

The initial stage to produce automatic summary is the text preprocessing stage which is the transformation proses of raw materials in forms of paragraphs into words that are ready to be calculated for their weights. The weight of the words functions to determine the word's degree of importance in the text. The text preprocessing stage consists of few processes, which are sentence segmentation process, case folding, tokenizing, filtering, and stemming. The last process of this stage produces words in form of root words. Here are the explanation of text preprocessing stage:

1. Sentence segmentation
   sentence segmentation is an initial step of text preprocessing stage. On this process, the news text which consists of some paragraphs are divided into stand alone sentences . The segmentaton of every sentence is based on punctiation, such as full stop (.), imperative mark (!), and question mark (?).
2. Case folding
   Case folding is a transformation process of all text into characters with small letters and discard every character outside alphabet. Punctuation, numeric character, and symbols are also discarded.
3. Filtering
   Filtering is a process of stopword omission. Stopword is a group of meaningless words which often emerges in the sentences. This process is conducted by checking on the stopword dictionary. If there are words which are found in the stopword dictionary, then they will be discarded or omitted. If the stopwords are not omitted, then they will have high weight. The examples of stopword are and, which, in, if, and, etc.
4. Tokenizing
   Tokenizing is a process to transform a sentence into root words. The sentence segmentation is based on the composing delimiter, which are space (” ”). The purpose of this process will then be continued with stemming.
5. Stemming
   Stemming is a process of transforming a word to return it to its root word by applying certain rules, so that each word has the same representation. Stemming is in this research uses Nazief & Adriani's Algorithm (1996). This Algorithm deals with confix, prefix, and suffix. Infix are not included because its rare existence in the Indonesian Bahasa. Here are the steps of Nazief & Adriani's Algorithm (1996):
   a. Searching for the words which will be stemmed in the dictionary. If they are found then they are assumed that the word is the root word, so that the Algorithm will stop.
   b. Discarding inflection suffixes ("-lah", "-kah", "-ku", "-mu", atau "-nya"), If they are in the form of particles ("-lah", "-kah", "-tah" atau "-pun") then this step is repeated in order to omit possesive pronouns ("-ku", "-mu", atau "-nya").
   c. Omitting derivation suffixes ("-i", "-an" atau "-kan"). If they are found, then the Algorithm will stop
   d. Omitting derivation prefix (be-, di-, ke-, me-, pe-, se- dan te-).
      1) the Algorithm will stop if:
         a) The word in the c stem has compounded prefix and suffix which are not allowed on the table 2.
         b) Awalan yang akan dihilangkan, sama dengan awalan yang telah dihilangkan sebelumnya. The prefix which will be omitted is the same with the previous ones.
         c) The prefix has been omitted three times.
      2) Identifying the type of prefix and its disambiguity. The prefix has two types which are:
         a) Plain : di-, ke-, se- prefixes can be omitted directly.
         b) Complex : be-, te-, me-, pe- prefixes must be analyzed for their ambiguities using the table 1. me- prefix can be changed into mem- or men- depending on the first letter of the root word (see table below).
      3) Do the step a. If the word is not the root word, then repeat the d step repeatedly until finding its root word or until d1 condition.
      4) If the d1 condition has not achieved, but you haven't got the root word. Then proceed to step e.
   e. If the step d haven't yielded anything, then analyzed whether the word is on the disambiguity table or not.
   f. If all steps are carried out but they don't yield anything then the word is the root word. The process is completed.

TABLE I. THE SUFFIX STEMMING PROCESS RULES OF NAZIEF & ADRIANI'S STEMMER

<table>
<tr><th>Rules</th><th>Word Format</th><th>Segmentation</th></tr>
<tr><td>1</td><td>berV…</td><td>ber-V… | be-rV</td></tr>
<tr><td>2</td><td>berCAP…</td><td>ber-CAP… dimana C!=’r’ & P!=’er’</td></tr>
<tr><td>3</td><td>berCAerV…</td><td>ber-CaerV… dimana C!’r’</td></tr>
<tr><td>4</td><td>Belajar…</td><td>bel-ajar</td></tr>
<tr><td>5</td><td>beC l erC2…</td><td>be-C lerC2… dimana C1!={‘r’|’I’}</td></tr>
<tr><td>6</td><td>terV…</td><td>Ter-V… | te-rV…</td></tr>
<tr><td>7</td><td>terCerV…</td><td>Ter-CerV… dimana C!=’r’</td></tr>
<tr><td>8</td><td>terCP…</td><td>Ter-CP… dimana C!=“r“ dan P!=“er“</td></tr>
<tr><td>9</td><td>teC1erC2...</td><td>te-C1erC2... dimana C1!=“r“</td></tr>
<tr><td>10</td><td>me{l|r|w|y}V...</td><td>me-{l|r|w|y}V...</td></tr>
<tr><td>11</td><td>mem{b|f|v}...</td><td>mem-{b|f|v}...</td></tr>
<tr><td>12</td><td>mempe{r|l}...</td><td>mem-pe...</td></tr>
<tr><td>13</td><td>mem{rV|V}...</td><td>me-m{rV|V}... | me-p{rV|V}...</td></tr>
<tr><td>14</td><td>men{c|d|j|z}...</td><td>men-{c|d|j|z}...</td></tr>
<tr><td>15</td><td>menV...</td><td>me-nV... | me-tV</td></tr>
<tr><td>16</td><td>meng{g|h|q}...</td><td>meng-{g|h|q}...</td></tr>
<tr><td>17</td><td>mengV...</td><td>meng-V... | meng-kV...</td></tr>
<tr><td>18</td><td>menyV...</td><td>meny-sV…</td></tr>
<tr><td>19</td><td>mempV...</td><td>mem-pV... dimana V!=„e“</td></tr>
<tr><td>20</td><td>pe{w|y}V...</td><td>pe-{w|y}V...</td></tr>
<tr><td>21</td><td>perV...</td><td>per-V... | pe-rV...</td></tr>
<tr><td>23</td><td>perCAP</td><td>per-CAP... dimana C!=“r“ dan P!=“er“</td></tr>
<tr><td>24</td><td>perCAerV...</td><td>per-CAerV... dimana C!=“r“</td></tr>
<tr><td>25</td><td>pem{b|f|V}...</td><td>pem-{b|f|V}...</td></tr>
<tr><td>26</td><td>pem{rV|V}...</td><td>pe-m{rV|V}... | pe-p{rV|V}...</td></tr>
<tr><td>27</td><td>pen{c|d|j|z}...</td><td>pen-{c|d|j|z}...</td></tr>
<tr><td>28</td><td>penV...</td><td>pe-nV... | pe-tV...</td></tr>
<tr><td>29</td><td>peng{g|h|q}...</td><td>peng-{g|h|q}...</td></tr>
<tr><td>30</td><td>pengV...</td><td>peng-V... | peng-kV...</td></tr>
<tr><td>31</td><td>penyV...</td><td>peny-sV…</td></tr>
<tr><td>32</td><td>pelV...</td><td>pe-lV... kecuali “pelajar” yang menghasilkan “ajar”</td></tr>
<tr><td>33</td><td>peCerV...</td><td>per-erV... dimana C!={r|w|y|l|m|n}</td></tr>
<tr><td>34</td><td>peCP...</td><td>pe-CP... dimana C!={r|w|y|l|m|n} dan P!=“er“</td></tr>
</table>

TABLE II. THE COMBINATION OF PREFIX AND SUFFIX WHICH IS NOT ALLOWED

Prefix	Suffix
be-	-i
di-	-an
ke-	-i, -kan
me-	-an
se-	-i, -kan

B. Text Processing

Text processing TF-IDF-DF Algorithm functions as to weight the score of each word emerges from the text preprocessing stage. The score of word's weight will be higher if the word frequently emerges in a sentences or multiple sentences (Pramono, 2013). The score of word's weight will be used in the calculation of cosine similarity which the stage to bridge TF-IDF-DF algorithm and MMR method, so that the text summarization process can be carried out.

Cosine similarity is a basic calculation in order to obtain similarity score between two vectors. Cosine similarity calculation consists of two stages which are :

1. Calculating relevance between document and query or title

   Calculating Cosinus of two vectors which are W (weight) of every document or sentence with W (weight) of the query (title)

2. Calculating similarity between documents

   Calculating Cosine angle of two vectors which are W (weight) of a sentence and W (weight) of another sentence (title)

   Cosine similarity equation is as follows (2).

$$sim\ (S_1, S_2) = \frac{\sum_i t_{1i} t_{2i}}{\sqrt{\sum_i t_{1i}^2}\ x\ \sqrt{\sum_i t_{2i}^2}} \qquad (3)$$

Information :
S_1 = the vector of the candidate sentence
S_2 = the vector of the sentence beside the candidate
t_{1i} = weight of word

If the similarity between a sentence with another sentence is high, then there is redundancy. In order to overcome it, then MMR method is needed to produce a good summary by omitting the redundant sentences. The example of MMR calculation result can be seen on the Table 3

TABLE III. MMR CALCULATION RESULT

Iteration	D1	D2	D3	D4
1	**0.32269**	0.019413	0.021992	0.11908
2	-	-0.01467	-0.08974	**0.0586**
3	-	-0.01467	-0.08974	-
4	-	-	-	-

It can be seen on the table 3 that iteration number 1 the highest sentence is on the sentence number 1 (D1), so that sentence number 1 (D1) on the iteration number 2 does not have MMR score, because the sentence number 1 (D1) has ben chosen to become a summarization. There is no sentence which becomes summarization on iteration 3, becausse the MMR max score is < = 0. The result of sentence ranking can be seen on the table 4 below

TABLE IV. THE RESULT OF SENTENCE RANKING

Rank	(D) SENTENCE NUMBER	Max MMR
1	1	0.32269
2	4	0.05860

V. Results and Discussion

The summarizing process of news article which was conducted by the system on the news article on the online national and local news portal has different results. The article summarizing process from the online national and local news portal can be seen on the Table 5 and 6.

TABLE V. Accuration Score of Online National News

Article sequence	R1	R2	R3	System
1	1,3,7	1,2,3	1,2,5	1,7
2	1,4	1,2,5	1,3,4	1,4
3	1,3,4	1,4,3	1,3,4	3,1
4	1,2,4	1,4,5	1,4,5	4,1,3
5	1,2,4	1,2,3	1,2,3	1,6,3
6	1,2,9	1,2,3	1,2,3	1,2,9
7	1,3,4	1,3,4	1,5,7	1,2,3
8	1,4,6	1,4,6	1,3,4	1,6,4

TABLE VI. Accuration Score of Online Local News

Article sequence	R1	R2	R3	System
1	7,3,5	1,3,4	1,4,9	1,11,14
2	1,2,9	1,2,4	1,2,5	9,10,2
3	1,3,8	1,3,10	1,3,6	3,11,2
4	1,5,3	1,2,3	3,7,8	5,2,1
5	1,3,5	1,2,4	1,2,3	1,2,3
6	1,2,3	1,2,3	1,3,6	3,7,2
7	2,4,9	2,5,6	1,2,4	4,5,6
8	1,9,7	6,8,9	8,6,9	2,1,9

Information :

R_i : Sentence Summarizing process which is conducted by respondents (i=1,2,3)

System : Sentence Summarizing process which is conducted by the system

Table 5 contains sentences which are chosen as the results of sentence summarization. It can be seen from the Table that there are similarities between the manual summarization (R) with the system summarization. In comparison to table 6, there are many dissimilarites between the respondents summarization (R) with the system summarization. This can yield differenceon the level of system accuracy which can be seen from the score of precision, recall, and f-measure. Based on the testing results, it can be seen that the summarization of news article from the online national news portal has precision score of 76.39 %, recall score of 65.28 %, and f-measure score of 70.4 %. Meanwhile, the summarization of news article from the online local news portal has precision score of 45.83 %, recall score of 45.83 %, and f-measure score of 45.83 %.

Based on the results of precision, recall, and f-measurei from online local and national news portal, it can be seen that MMR method with word weighting using TF-IDF-DF is better to be applied in the online national news portal because it produces higher accuracy score of 70,4% compared to the online local news portal with the accuracy of 45,83%. It happens because in the online national news portal, there are words which are contrary to the title (query), but have the similar meaning, so that the system will choose other sentences which contain more words which are similar to the query. Not to mention, the news text uses short sentences, so that it affects the results of the summarization process. The score of MMR will be higher if the sentence has more similar words to the query and it consists of shorter sentences. The examples can be seen on the table 7.

TABLE VII. The Influence of Long Sentences

Sentences	MMR
Hobi Ani **membaca**	0.252743
Terutama novel, **buku** yang Ani sukai	0.124753

Query : reading material

On the table 7, it can be seen that there 3 words on the first sentence with 1 query produces MMR score of 0.252742. The second sentence has 5 words (stopword is omitted) with 1 query produces MMR score of 0. 124753. So, the same amount of query, but with the different length of sentence produces different MMR score. The bigger MMR score will be found on shorter sentences. The Influences of the amount of query can be seen on the table 8.

TABLE VIII. The Influences of the Amount of Query

Sentences	MMR
Ani memiliki **hobi membaca**	0.339428
Terutama novel **buku** kesukaannya	0.158445

Query : hobi membaca buku

The table 8 shows that on the first sentence there are 2 queries which are ”hobi” dan ”baca” (they are root words) with the MMR score of 0.339428, but on the second sentence, there is 1 query which is ”buku” with the MMR score of 0. 158445. So, the bigger the query in a sentence, the more MMR score that the sentence will have, the more possible for the sentence to be summarized.

VI. Conclusion

Text summarization using MMR method and TF-IDF-DF on the news taken from Online National News produces a better summary compared to the news taken from Online local News. the online national news portal has precision score of 76.39 %, recall score of 65.28 %, and f-measure score of 70.4 %, Meanwhile, the summarization of news article from the online local news portal has precision score of 45,83 %, recall score of 45.83 %, and f-measure score of 45.83 %. It is caused by the words used on the query (title) are not the same or do not exist on the news text. In addition, online local news portal has more shorter sentences which influence MMR score. So the MMR score will ne bigger if there are more queries on the sentence and the sentence will be summarized.

VII. SUGGESTION

The news summarization from the online local news portal produces low f-measure score because the query which is used is not appropriate so that it creates mistakes in determining the news summary. Further development is suggested by inputting the first sentence as the consideration in determining the general summarization of the news, because the first sentence describes the content of the news as a whole.

REFERENCES

[1] Nugroho, Y., D.A. Putri, dan S. Laksmi, “Mapping The Landscape of The Media Industry in Contemporary Indonesia,” Centre for Innovation Policy and Governance, 2012.

[2] Lahari, E.P., D.V.N.S. Kumar, dan S.S. Prasad, “Automatic Text Summarization with Statistical and Linguistic Features Using Successive Thresholds,” IEEE, pp. 1519, 2014.

[3] Luhn, H.P, “The Automatic Creation of Literature Abstracts,” IBM JOURNAL, pp. 159-165, 1958

[4] Hovy, E. dan C.Y. Lin, “Automated Text Summarization in SUMMARIST,” 1997.

[5] Barzilay, R. dan M. Elhadad, “Using Lexical Chains for Text Summarization,” Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization, 1997.

[6] Aristoteles, “Penerapan Algoritma Genetika pada Peringkasan Teksa Dokumen Bahasan Indonesia,” Prosiding Semirata FMIPA Universitas Lampung, pp. 29-33, 2013.

[7] Mustaqhfiri, M., Z. Abidin, dan R. Kusumawati, “Peringkasan Teks Otomatis Berita Berbahasa Indonesia Menggunakan Metode Maximum Marginal Relevance,” MATICS, vol. 4, pp. 134-147, 2011.

[8] Xie, S. dan Y. Liu, “Using Corpus and Knowledge-Based Similarity Measure in Maximum Marginal Relevance for Meeting Summarization,” Acoustics Speech and Signal Processing, vol. 8, pp. 4985-4988, 2008.

[9] Pramono, L.H., A.S. Rohman, dan H. Hindersah, “Modified Weighting Method in TF*IDF Algorithm for Extracting User Topic Based on Email and Social Media in Integrated Digital Assistant,” Rural Information & Communication Technology and Electric-Vehicle Technology, pp. 1-6, 2013.

[10] Carbonell, J.G. dan J. Goldstein, “The Use of MMR and Diversity-Based Reranking in Document Reranking and Summarization,” IEEE, vpl.12, pp. 335-336, 1998.

[11] Nazief, B. A. A. dan M. Adriani, “Confix-Stripping : Approach to Stemming Algorithm for Bahasa Indonesia,” International Conference on Information and Knowledge Management, pp. 560-563, 1996.