Inclusion-Based and Exclusion-Based Approaches in Graph-Based Multiple News Summarization

Nongnuch Ketui and Thanaruk Theeramunkong

School of Information, Computer and Communication Technology,
Sirindhorn International Institute of Technology, Thammasat University
nongnuch.ketui@studentmail.siit.tu.ac.th,
thanaruk@siit.tu.ac.th
http://www.siit.tu.ac.th

Abstract. As combination of information extraction and relation analysis, constructing a comprehensive summary from multiple documents is a challenging task. Towards summarization of multiple news articles related to a specific event, an ideal summary should include only important common descriptions of these articles, together with some dominant differences among them. This paper presents a graph-based summarization method which is composed of text preprocessing, text-portion segmentation, weight assignment of text portions, and relation analysis among text portions, text-portion graph construction, and significant portion selection. In the process of portion selection, this paper proposes two alternative methods; inclusion-based and exclusion-based approach. To evaluate these approaches, a set of experiments are conducted on fifteen sets of Thai political news articles. Measured with ROUGE-N, the result shows that the inclusion-based approach outperforms the exclusion-based one with approximately 2% performance gap (80.59 to 78.21%).

Keywords: Thai Text Summarization, Multiple News Summarization, Graph-based Summarization.

1 Introduction

Nowadays a gigantic number of news articles are produced in any language all over the world. Not an exception, we can also find a large pile of Thai news articles online. The situation of having too much information make us difficult to find needed information and requests us a long time to read many related news issues before catching the occurring events. Especially it is usual to find news articles relating to an event, from several sources with similar and different facts. To solve this problem, it is necessary to study an efficient and effective method to make a summary from multiple written Thai news articles. The expected summarization gives a short comprehensive description of the similarity and the difference among related news.

In this paper, we present a graph-based summarization method which is composed of text preprocessing, text-portion segmentation, weight assignment of

T. Theeramunkong et al. (Eds.): KICSS 2010, LNAI 6746, pp. 91–102, 2011.

text portions, relation analysis among text portions, text-portion graph construction, and significant portion selection. In the process of portion selection, this paper proposes two alternative methods; inclusion-based and exclusion-based approach. Both approaches are evaluated with a set of Thai political news articles and measured with ROUGE-N. Section 2 describes related works on multi-document summarization and two alternative methods are presented in Section 3. In Section 4, we show experimental settings and results. Finally, conclusion and future work are given in Section 5.

2 Related Works

In past decades, summarization has been recognized as an interesting application in the field of natural language processing (NLP). In an earlier period of summarization research, several works placed interest on scientific documents and proposed various paradigms to extract salient sentences from texts by making use of features like word and phrase frequency [1], position in the text [2] and key phrases [3]. In general, two main summarization approaches are extraction-based and abstraction-based methods. The extraction-based methods [4], [5] and [6] usually involve a process to assign weights for each processing unit, mostly sentence unit or phrase unit. In fact, many approaches differ on the manner of their problem formulation. Various researchers published their work concentrated on newswire data. Many approaches are developed under different problem setting for summarization. One of the important issues in summarization is how to evaluate the obtained summary. As an early stage of multi-document summarization, Radev and his colleagues [7], [8] and [9] proposed centroid-based techniques to generate a composite sentence from each cluster. Radev and Barzilay [7] and that of Barzilay et al. [9] formulated summarization as a clustering problem. To compute a similarity measure between text units, they are mapped to feature vectors that are represented by a set of single words weighted by some weighting system such as TF-IDF. Goldstein and Carbonell [10] combined query relevance with information novelty as the topic and made a major contribution to topic-driven summarization by introducing the maximal marginal relevance (MMR) measure. Mani [11] presented an information extraction framework for summarization as well as a graph-based method to find similarities and dissimilarities in pairs of documents. In this approach, it is possible to specify on maximal number of common and different sentences to control the output. This method uses these structures to actually compose abstractive summaries, rather than to extract sentences from the text.

3 Inclusion- and Exclusion-Based Summarization

This section illustrates four main parts involved in the summarizing process, (1) text preprocessing, (2) node weight assignment, (3) edge weight assignment and (4) candidate paragraphs selecting. Their details are gave in subsections as follow in Fig. 1.

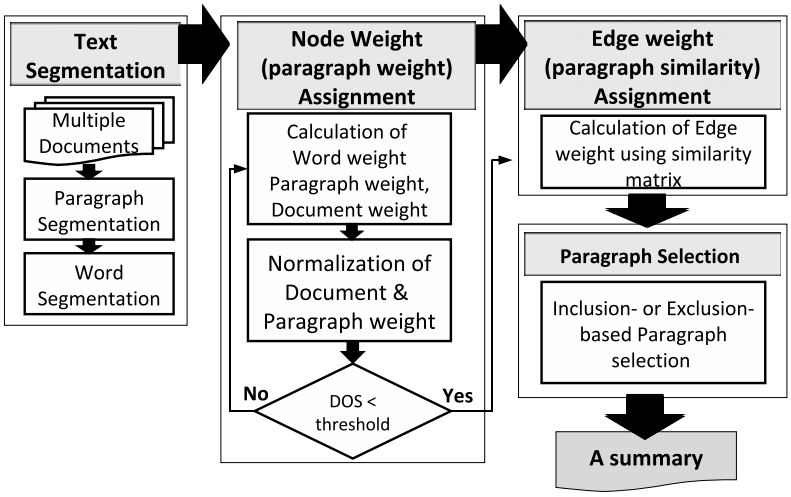

Fig. 1. The process of inclusion- and exclusion-based summarization. Thus, DOS is defined the different of summation of document weight.

3.1 Text Segmentation

As a text segmentation, the document related discovery proposed in [12] is applied to find a set of related news. For this purpose, three main types of news relations are classified based on the relevance, of new events: (1) "completely related" (CR), (2) "somehow related" (SH) and (3) "unrelated" (UR). A CR relation is detected when two new documents mention a same story. For SH relation, it is a kind of relation which has only somewhat closely related. The relation of UR is defined as a relationship of having absolutely unrelated in their events between news documents. In this work, we consider only a CR relation on Thai news articles and solve the problem on multi-document summarization. The problem of Thai language is even worse since the Thai language has no explicit sentence boundary and no clear definition. Therefore, paragraph segmentation only uses the HTML tags <p> (a new paragraph) and
 (line-break) tags. Finally, words are segmented with the maximal matching algorithm [13]. This algorithm first generates all possible segmentations for a sentence and then selects the one that contains the fewest words, which can be done efficiently by using dynamic programming technique.

3.2 Node Weight Assignment

Before assigning the edge weight, the weight of text-portion is measured by a statistical method. This method evaluates the importance of words, paragraphs and documents. Generally, TFIDF (Term Frequency times Inverse Document Frequency) is a well-known as weighting words in document. By this work we slightly modified TFIDF to find important words in a paragraph later called TFIPF (Term Frequency times Inverse Paragraph Frequency) for calculating weights. Formulations of node weight assignment can be summarized as follows.

Let D be a set of news articles where $D = \{d_1, d_2, d_3, ..., d_i\}$, and $CRD = \{d_1, d_2, d_3, ..., d_m\} \subset D$ with m news articles is called a completely related news articles set. Given d_i represents the i-th order document, the document d_i has a set of paragraph p_{ij} where $\{p_{i1}, p_{i2}, p_{i3}, ..., p_{ij}\} \in d_i$ and p_{ij} is a j-th order paragraph in the document d_i. According to the paragraph p_{ij} in the document d_i has a set of word w_{ijk} where $\{w_{ij1}, w_{ij2}, w_{ij3}, ..., w_{ijk}\} \in p_{ij}$. The word weight, $W^{(t)}(w_{ik})$ is used to focus on important words appearing in both documents and paragraphs. Let t be the member of iterations and w_{ik} represents word w_k in the document d_i. Before calculating the weight of a particular word w_k, the initial word weight and paragraph weight is assigned as $W^{(0)}(w_{ik}) = \frac{N(w_{ik})}{N(w_i)}$ and $W^{(0)}(p_{ij}) = \frac{1}{P_i}$ respectively. Let $N(w_{ik})$ is the total number of a word w_k that occurs in the document d_i and $N(w_i)$ is the total number of all words in the document d_i. While the total number of paragraphs appears in the document d_i is given to P_i.

$$W_w^{(t+1)}(w_{ik}) = N(d_i) \times W^{(t)}(w_{ik}). \tag{1}$$

The frequency of a particular word w_k appears in the document d_i as $W_w^{(t+1)}(w_{ik})$. Given $N(d_i)$ is the total number of a particular word appearing in the i-th document. While the weight of a particular word w_k appears in the paragraph p_j in the document d_i as $W_p^{(t+1)}(w_{ik})$ that will be assigned as

$$W_p^{(t+1)}(w_{ik}) = P(d_i) \times \sum_{p_{ij} \in d_i} N(w_{ijk}) \times W^{(t)}(p_{ij}), \tag{2}$$

where $P(d_i)$ is the total number of paragraphs appears in i-th order document, $N(w_{ijk})$ is the total number of a particular word w_k appearing in the paragraph p_j of the document d_i.

$$TF^{(t+1)}(w_{ik}) = \sqrt{W_w^{(t+1)}(w_{ik}) \times W_p^{(t+1)}(w_{ik})}, \tag{3}$$

where $TF^{(t+1)}(w_{ik})$ is the expected frequency of a particular word w_k appearing in both of the document and paragraph as Equation 1 and 2. Let $IDF^{(t+1)}(w_k)$ be inversed document frequency represented by the following formula:

$$IDF^{(t+1)}(w_k) = \log\left(1 + \frac{\sum_{d \in D} W^{(t)}(d)}{\sum_{d_i \in D} W^{(t)}(d_i)}\right), \tag{4}$$

such that $w_k \in d_i$. Given $IDF^{(t+1)}(w_k)$ represents by the summation of document weight $W^{(t)}(d)$ in corpus divided by the expected summation of the document weight where word w_k occurs in the document d_i and in corpus D. Here, the initial document weight $W^{(0)}(d)$ is assigned as the total of documents in the corpus where $W^{(0)}(d) = \frac{1}{|D|}$. Then the word weight will be assigned with Equation 3 and 4 using the following formula:

$$W^{(t+1)}(w_{ik}) = \frac{TF^{(t+1)}(w_{ik})}{\sum_{w \in d_i} TF^{(t+1)}(w)} \times IDF^{t+1}(w_k). \tag{5}$$

The word weight $W^{(t+1)}(w_{ik})$ has the value between 0 and IDF. After assigning a word weight in the document d_i, an inverse paragraph frequency IPF is calculated as follows.

$$IPF^{(t)}(w_{ik}) = \log\left(1 + \frac{\sum_{p \in d_i} W^{(t)}(p)}{\sum_{p_j \in d_i} W^{(t)}(p_j)}\right), \tag{6}$$

such that $w_{ik} \in p_j$, where $IPF^{(t)}(w_{ik})$ is assigned with the summation of the paragraph weight $W^{(t)}(p)$ in the document d_i and the summation of the paragraph weight $W^{(t)}(p_j)$ that has a word w_i occurring in the paragraph p_j and in the document d_i. Then the paragraph weight is calculated by using the following formula:

$$W^{(t)}(p_{ij}) = \sum_{w_{ik} \in p_j} W^{(t)}(w_{ik}) \times IPF^{(t)}(w_{ik}). \tag{7}$$

All words in this paragraph p_{ij} are summed up. The paragraph weight $W^{(t)}(p_{ij})$ is calculated by the summation of the multiplication of the word weight in this paragraph $W^{(t)}(w_{ik})$ and an inverse paragraph frequency $IPF^{(t)}(w_{ik})$ using Equations 5 and 6. Before calculating a document weight, we need to find the location weight W_{loc} by linear formula, $W_{loc}(p_{ij}) = -0.1 \times j + 1.1$;. Let j is the j-th paragraph. This formula is firstly generated on the random news articles and weighed by human, and then plotted graph. The location of paragraph indicates the important information in the new articles.

$$W^{(t)}(d_i) = \sum_{p_{ij} \in d_i} W^{(t)}(p_{ij}) \times W_{loc}(p_{ij}). \tag{8}$$

All paragraphs in the document d_i are summed up. The document weight $W^{(t)}(d_i)$ is calculated by the multiplication of the paragraph weight $W^{(t)}(p_{ij})$ in the paragraph p_j and the location weight $W_{loc}(p_{ij})$ that paragraph p_j occurs.

Finally, the document and paragraph weight is normalized by the two following formulas:

$$W^{(t+1)}(d_i) = \frac{W^{(t)}(d_i)}{\sum_{d \in D} W^{(t)}(d)}, \tag{9}$$

$$W^{(t+1)}(p_{ij}) = \frac{W^{(t)}(p_{ij})}{\sum_{p \in d_i} W^{(t)}(p)}. \tag{10}$$

The node weight assignment will be iterated until the difference of summation of the document weight (DOS) is less than threshold 0.5.

3.3 Edge Weight Assignment

Conceptually selecting candidate paragraph with Graph-based summarization in multiple news articles from a similarity relation graph, where a node corresponds to a paragraph in sets of news articles and an edge corresponds to an indirect

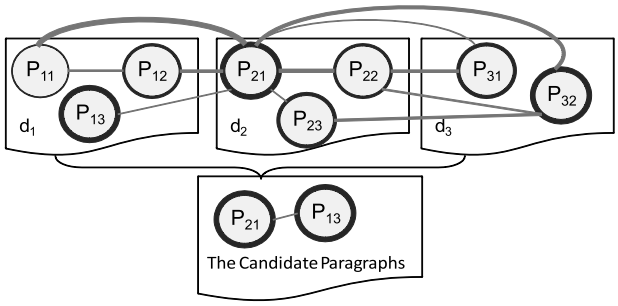

Fig. 2. An example of a similarity relation graph

relation of a paragraph to another paragraph. The edge weights display the similarity between two paragraphs in a set of completely related news articles. They are calculated by cosine similarity as the following formula:

$$sim(p_{ij}, q_{ij}) = \frac{\sum_{k=1}^{n} wp_{ijk} \times wq_{ijk}}{\sqrt{\sum_{k=1}^{n} wp_{ijk}^2} \sqrt{\sum_{k=1}^{n} wq_{ijk}^2}}, \tag{11}$$

where p_{ij} and q_{ij} are vectors of $< wp_{ij1}, wp_{ij2}, ..., wp_{ijk} >$ and $< wq_{ij1}, wq_{ij2}, ..., wq_{ijk} >$, respectively. Given wp_{ijk} and wq_{ijk} are the word weights of a word w_k occurring in the paragraph p_j in the document d_i. Then a term by paragraph matrix is generated with rows of the word weights and columns of the paragraphs.

3.4 Inclusion-Based Approach

The inclusion-based approach starts from selecting the most important text portion and then add the second most important text portion the semantics of which is not overlapped with the first selected text portion. By this approach selects the candidate paragraph from multiple news articles. Algorithm 1 shows a pseudo code of inclusion-based approach to analyze the similarity between paragraph and another paragraph. A set of node weight(paragraph weight) is given as S, and then we set the average paragraph for selecting candidate paragraphs. The average paragraph in each CR news is calculated by the ratio of number of paragraph and number of document and then multiply with 0.5. To initialize the node with a maximal weight w_{max} from the set S. We consider the neighbouring nodes which connect with the maximal node, and then find the maximum ratio of the neighbouring node weight and both of the included and non-included edge weight. Before this neighbouring node is added into the inclusion graph G, it is assigned to the maximal node. Then we iterate to consider the neighbouring nodes again until the total of nodes is equal the average of a paragraphs P_{avg}. Finally, the inclusion graph is returned to the candidate summaries.

Algorithm 1. Inclusion-based approach

Input: Number of paragraphs n_p, documents in related news n_d,
Node Weight w and Edge Weight e
Output:The candidate paragraphs in Graph G
1 : create Graph S with all node weight w and connect with edge weight e
2 : calculate the average of paragraphs per related news article $P_{avg} = \frac{n_p}{n_d} \times 0.5$
3 : set the initial node w_{max} to the maximum node weight in S
4 : **do**
5 : insert(G, w_{max})
6 : delete(S, w_{max})
7 : **foreach** all nodes $w_i \in S$ **do**
8 : **foreach** all nodes $w_j \in S$ are connected with w_{max} **do**
9 : calculate the summation of edge weights $e_{i,j}$
10: calculate the average of edge weight $avg(e_{i,j})$
11: **foreach** all nodes $w_k \in G$ are connected with w_{max} **do**
12: calculate the summation of edge weights $e_{i,k}$
13: calculate the average of edge weight $avg(e_{i,k})$
14: $w_{inew} = \frac{w_i \times avg(e_{i,j})}{avg(e_{i,k})}$
15: **if** w_{inew} is the maximum weight **then**
16: $w_{max} = w_i$
17: **until** the total of nodes is equal P_{avg}
18: return G

3.5 Exclusion-Based Approach

Exclusion-based approach finds the maximal of common paragraphs content in the related news articles and omits the maximal of similarity between two paragraphs. The summary from multiple news articles can be constructed when we omit useless paragraphs. Algorithm 2 shows a pseudo code of exclusion-based approach. A set of node weight(paragraph weight) is given as G, and then we set the average paragraph for selecting candidate paragraphs. To initialize the node with a minimal weight from the set G. We consider the neighbouring nodes which connect with the minimal node, and then find the maximum ratio of the neighbouring node weight and both of the excluded and non-excluded edge weight. Before the edge of this neighbouring node is omitted from the exclusion graph G, the neighbouring node is assigned to the minimal node. Then we iterate to consider the neighbouring nodes again until the rest total of nodes is less than or equal the average of a paragraphs P_{avg}. Finally, the rest nodes in exclusion graph are returned to the candidate summaries.

4 Experiment

4.1 Experimental Setting and Evaluation Criteria

For significant portion selection, we investigate and compare the performance of both approaches when the selected candidate paragraphs are considered as the

Algorithm 2. Exclusion-based approach

Input: Number of paragraphs n_p, documents in related news n_d,
Node Weight w, Edge Weight e and Empty Graph S
Output:The candidate paragraphs in Graph G
1 : create Graph G with all node weight w and connect with edge weight e
2 : calculate the average of paragraphs per related news article $P_{avg} = \frac{n_p}{n_d} \times 0.5$
3 : set the initial node w_{min} to the minimum node weight in G
4 : **do**
5 : delete(G, w_{min})
6 : insert(S, w_{min})
7 : **foreach** all nodes $w_i \in G$ **do**
8 : **foreach** all nodes $w_j \in G$ are connected with w_{min} **do**
9 : calculate the summation of edge weights $e_{i,j}$
10: calculate the average of edge weight $avg(e_{i,j})$
11: **foreach** all nodes $w_k \in S$ are connected with w_{min} **do**
12: calculate the summation of edge weights $e_{i,k}$
13: calculate the average of edge weight $avg(e_{i,k})$
14: $w_{inew} = \frac{e_{i,j}}{w_i \times e_{i,k}}$
15: **if** w_{inew} is the maximum weight **then**
16: $w_{min} = w_i$
17: **until** the rest nodes is equal P_{avg}
18: return G

correct answers. Precision, recall and F-measure are used in the evaluation are as follows:

$$Precision = \frac{|P_{ref} \cap P_{sys}|}{P_{ref}}. \tag{12}$$

$$Recall = \frac{|P_{ref} \cap P_{sys}|}{P_{sys}}. \tag{13}$$

$$F - measure = \frac{(\alpha + 1) \times recall \times precision}{recall + (\alpha \times precision)}. \tag{14}$$

where P_{ref} and P_{sys} denote the number of paragraphs appeared in the reference summary and in the system summary, respectively. For F-measure, the experiments use F1 (i.e., the value of α is 1).

In addition, Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [14] is used for the evaluation. ROUGE calculation is based on various statistical metrics by counting overlapping units such as n-grams, word sequences, and word pairs between systems which generate summaries correlating with those extracted by human evaluations. ROUGE-N is included in the summarization evaluation package, distributed by National Institute of Standards and Technology (NIST). ROUGE-N is an N-gram recall between an automatic summary and a set of manual summaries, which is calculated as:

Table 1. The output paragraphs selection from human and system summaries

Set of CR News	Number of paragraph in document	Number of word in paragraph	Number of output paragraphs	Reference summaries	System summaries	
					Inc.-based approach	Exc.-based approach
1	3(3,2,3)	1(61,85,54) 2(132,85,54) 3(27,118,54)	2	(2-1, 2-2)	(2-1, 2-2)	(2-1, 3-2)
2	3(6,9,8)	4(67,167,187,160,138,105) 5(30,127,37,113, 70,126,29,146) 6(11,73,180,115,70, 160,144,104)	4	(6-3, 5-5, 4-3, 4-4)	(4-3, 6-3, 4-2, 4-4)	(4-3, 4-2, 6-3, 5-5)
3	3(7,4,5)	7(44,56,95,99,100,92,57) 8(218,149,59,62) 9(9,65,204,148,79)	3	(9-3, 8-2, 9-5)	(8-1, 9-4, 8-2)	(8-1, 8-2, 9-3)
4	3(4,5,2)	10(11,10,175,36) 11(25,50,75,72,45) 12(114,59)	2	(12-1, 12-2)	(12-1, 10-3)	(12-1, 10-3)
5	3(7,4,5)	13(44,56,95,99,100,91,57) 14(218,149,59,61) 15(9,65,204,148,79)	3	(15-3, 14-2, 15-5)	(14-1, 15-4, 14-2)	(14-1, 14-2, 15-3)
6	3(4,3,4)	16(12,22,141,37) 17(38,123,42) 18(30,95,47,39)	2	(16-3, 17-3)	(17-2, 16-3)	(17-2, 16-3)
7	3(8,18,14)	19(7,51,80,201,222 ,157,133,122) 20(37,56,15,72,32,44,57, 50,66,66,38,66,81,42, 44,46,56,65) 21(117,91,134,11,78,69, 58,122,75,134,12,86,95,33)	7	(19-4, 19-5, 19-6, 19-7, 19-8, 21-1, 19-2)	(19-4, 21-1, 19-5,19-8, 19-6,21-10, 21-3)	(21-1, 19-4, 19-5, 19-8, 19-6, 21-3, 19-3)
8	3(7,7,10)	22(11,150,59,312, 95,162,117) 23(72,143,32,229, 118,159,203) 24(40,88,84,85,101,103, 161,67,104,49)	4	(22-4, 23-5, 23-6, 23-7)	(22-4, 23-7, 22-2, 23-4)	(22-4, 23-7, 22-2, 23-4)
9	3(6,10,10)	25(34,94,95,79,156,86) 26(34,70,46,78,61, 47,45,83,73,35) 27(26,70,46,78,60, 47,45,83,73,35)	4	(25-2, 25-3, 25-5, 25-6)	(25-5, 26-6, 27-9, 25-2)	(25-5, 27-9, 25-2, 26-6)
10	3(5,4,4)	28(27,100,20,39,88) 29(12,133,39,88) 30(5,47,142,42)	2	(30-3, 28-5)	(30-3, 29-4)	(30-3, 29-2)
11	4(3,5,4,4)	31(144,68,78) 32(27,103,71,54,28) 33(44,128,71,53) 34(56,118,71,54)	3	(31-1, 31-2, 31-3)	(31-1, 32-3, 34-2)	(33-2, 31-1, 32-3)
12	3(5,6,7)	35(57,208,179,101,78) 36(9,36,146,197,98,88) 37(14,5,210,85, 101,129,115)	3	(37-3, 35-3, 35-4)	(35-2, 36-4, 37-3)	(36-4, 35-3, 37-3)
13	4(4,3,4,3)	38(29,76,67,57) 39(104,69,83) 40(27,72,63,92) 41(90,62,98)	2	(39-1, 40-3)	(39-3, 41-1)	(41-1, 41-3)
14	3(4,2,4)	42(29,64,52,63) 43(150,113) 44(29,64,52,63)	2	(43-1, 43-2)	(43-1, 42-4)	(43-1, 43-2)
15	3(4,3,2)	45(14,49,158,88) 46(36,160,87) 47(161,86)	2	(47-1, 45-4)	(47-1, 46-2)	(47-1, 46-2)

$$ROUGE-N = \frac{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)}. \quad (15)$$

4.2 Experimental Results

The experiment is to compare two alternative approaches; inclusion-based and exclusion-based approaches. Table 1 showed the selected paragraphs from manual and both approaches summaries. The detail of sets of CR news are also displayed the number of documents, paragraphs and words. For example, the first set of CR news has three documents; the document no.1 contains three paragraphs and each of them consists of 61,85 and 54 words, respectively.

From Table 1, the number of words in paragraph of the inclusion-based approach was less than the exclusion-based approach in five sets and greater than the exclusion-based approach in four sets. The number of output paragraphs are calculated from the average of paragraphs in related news and multiply with 0.5. Our approaches have to find the different and common of content, so that the output summaries are selected at least two paragraphs.

In addition, Table 2 showed the comparison of f-measure of both approaches, both approaches could achieve the similar performance in eights sets of CR news. The inclusion-based approach performed better than the exclusion-based

Table 2. F-measure Comparison between inclusion-based and exclusion-based approaches

Set of CR News	The number of output paragraphs	Inc.-based approach			Exc.-based approach		
		P(%)	R(%)	F(%)	P(%)	R(%)	F(%)
1	2	100.00	100.00	100.00	63.31	82.20	71.53
2	4	89.38	84.52	86.88	89.38	84.52	86.88
3	3	74.90	88.51	81.14	67.95	89.20	77.14
4	2	60.33	83.91	70.19	60.33	83.91	70.19
5	3	74.90	88.51	81.14	67.95	89.20	77.14
6	2	64.34	89.73	74.94	64.34	89.73	74.94
7	7	90.43	97.33	93.75	90.12	90.30	90.21
8	4	77.74	85.46	81.42	77.74	85.46	81.42
9	4	96.71	86.64	91.40	96.71	86.64	91.40
10	2	76.49	93.85	84.28	57.14	76.92	65.57
11	3	71.91	73.38	72.64	70.59	73.72	72.12
12	3	59.09	81.29	68.43	59.09	81.29	68.43
13	2	75.19	79.05	77.07	75.19	79.05	77.07
14	2	89.80	66.42	76.36	100.00	100.00	100.00
15	2	62.87	76.89	69.18	62.87	76.89	69.18
Average		77.60	85.03	80.59	73.51	84.60	78.21

approach in six sets while the exclusion-based approach was good measures in only one set. In the inclusion-based approach, we found that precision measure achieve 77.60% and recall measure achieve up 85.03%. While the exclusion-based approach has 73.51% in precision measure and 84.60% in recall measure. The inclusion-based approach can achieve up to 80.59% and the exclusion-based approach can achieve up to 78.21%. As a conclusion, the performance of the inclusion-based approach is better than the exclusion-based approach.

5 Conclusion and Future Work

This paper presented a graph-based summarization method to use the sets of Thai political news articles as resources for extracting the candidate summary. Four main steps of summarization in this work are as follows. First, preprocessing is performed to extract the sets of news article relations and remove all unnecessary significant symbols. Then, text-portion segmentation is used to separate an original article into paragraph and word unit as sources for assigning node weight. Second, node weight assignment is used to find an appropriate weight of a word, a paragraph and a document. Third, edge weight assignment find similarity between two paragraphs represented an edge of nodes. Finally, in the process of portion selection, we use two alternative methods; inclusion-based approach and exclusion-based approach. In the first approach, the most important contents are added. We considers the differential contents as finding the maximal of average weight between node weight and edge weight, and including the summary. In the second approach, exclusion-based approach focuses on the common contents as finding the minimal of average weight between node weight and edge weight, and excluding the summary. By comparing the result of inclusion-based and exclusion-based approach, the evaluation use ROUGE-N [14] compare with manual evaluation Using fifteen sets of Thai political news articles obtained from an allnews database and classified in completely related type of news relations based on the relevance of news events as a process to discover news article relation [12], two proposed methods were shown to be a powerful way to select the candidate paragraphs and comparison to human judgement. By experiment, the result shows that the inclusion-based approach outperforms the exclusion-based one with approximately 2% performance gap (80.59 to 78.21%). As a future work, we plan to improve the inclusion- and exclusion-based approach by iteration of node and edge weight after insertion or deletion nodes. Towards this work, we need to consider the performance of selecting the summary between two approaches in multiple news articles.

Acknowledgments. This work was supported by the National Research University Project of Thailand Office of Higher Education Commission, as well as the National Electronics and Computer Technology Center (NECTEC) under Project Number NT-B-22-KE-38-54-01. We would like to thank to all members at KINDML laboratory at Sirindhorn International Institute of Technology for fruitful discussion and comments, and Dr.Suriyawut Ketui on his kind helps in human evaluation.

References

1. Luhn, H.P.: The automatic creation of literature abstracts. IBM J. Res. Dev. 2(2), 159–165 (1958)
2. Baxendale, P.B.: Machine-made index for technical literature: an experiment. IBM J. Res. Dev. 2(4), 354–361 (1958)
3. Edmundson, H.P.: New methods in automatic extracting. J. ACM 16(2), 264–285 (1969)
4. Goldstein, J., Kantrowitz, M., Mittal, V., Carbonell, J.: Summarizing text documents: Sentence selection and evaluation metrics. In: Research and Development in Information Retrieval, pp. 121–128 (1999)
5. Mani, I., Bloedorn, E.: Summarizing similarities and differences among related documents. Information Retrieval 1, 35–67 (1999), 10.1023/A:1009930203452
6. Radev, D.R., Hovy, E., McKeown, K.: Introduction to the special issue on summarization (2002)
7. Radev, D.R., Jing, H., Budzikowska, M.: Centroid-based summarization of multiple documents: Sentence extraction, utility-based evaluation, and user studies (2000)
8. McKeown, K., Klavans, J., Hatzivassiloglou, V., Barzilay, R., Eskin, E.: Towards multidocument summarization by reformulation: Progress and prospects. In: AAAI/IAAI, pp. 453–460 (1999)
9. Barzilay, R., McKeown, K.R., Elhadad, M.: Information fusion in the context of multi-document summarization (1999)
10. Goldstein, J., Carbonell, J.: Summarization (1) using mmr for diversity - based reranking and (2) evaluating summaries. In: Proceedings of a Workshop on Held at Baltimore, Maryland, pp. 181–195. Association for Computational Linguistics, Morristown (1996)
11. Mani, I.: Multi-document summarization by graph search and matching. In: Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI 1997), pp. 622–628. AAAI (1997)
12. Kittiphattanabawon, N., Theeramunkong, T., Nantajeewarawat, E.: Exploration of document relation quality with consideration of term representation basis, term weighting and association measure. In: Chen, H., Chau, M., Li, S.-h., Urs, S., Srinivasa, S., Wang, G.A. (eds.) PAISI 2010. LNCS, vol. 6122, pp. 126–139. Springer, Heidelberg (2010)
13. Meknavin, S., Charoenpornsawat, P., Kijsirikul, B.: Feature-based Thai word segmentation. In: Proceedings of the Natural Language Processing Pacific Rim Symposium 1997 (1997)
14. Lin, C.-Y.: Rouge: A package for automatic evaluation of summaries. In: Proc. ACL Workshop on Text Summarization Branches Out, p. 10 (2004)