Computing and Informatics, Vol. 35, 2016, 1–29

THAI MULTI-DOCUMENT SUMMARIZATION: UNIT SEGMENTATION, UNIT-GRAPH FORMULATION, AND UNIT SELECTION

Nongnuch Ketui, Thanaruk Theeramunkong

School of Information, Computer, and Communication Technology
Sirindhorn International Institute of Technology
Thammasat University
Pathum Thani 12000, Thailand
e-mail: {nongnuch, thanaruk}@siit.tu.ac.th

Abstract. There have been several challenges in summarization of Thai multiple documents since Thai language itself lacks of explicit word/phrase/sentence boundaries. This paper gives definition of Thai Elementary Discourse Unit (TEDU) and then presents our three-stage summarization process. Towards implementation of this process, we propose unit segmentation using TEDUs and their derivatives, unit-graph formation using iterative unit weighting and cosine similarity, and unit selection using highest-weight priority, redundancy removal, and post-selection weight recalculation. To examine performance of the proposed methods, a number of experiments are conducted using fifty sets of Thai news articles with their manually constructed reference summary. By three common evaluation measures of ROUGE-1, ROUGE-2, and ROUGE-SU4, the results evidence that (1) our TEDU-based summarization outperforms paragraph-based summarization, (2) our iterative weighting is superior to traditional TF-IDF, (3) the highest-weight priority without centroid preference and unit redundancy consideration helps improving summary quality, and (4) post-selection weight recalculation tends to raise summarization performance under some certain circumstances.

Keywords: Thai text summarization, multi-document summarization, iterative weighting

Mathematics Subject Classification 2010: 68-T50

1 INTRODUCTION

Recently there have been an increasing number of online news articles published in any language on public news web sites, social media and web portals. Such information overload triggers a need to study and develop a system or a mechanism which contrasts and summarizes a large number of related or somehow related news articles with similar and different facts from several sources. To solve this problem, it is necessary to study an efficient and effective method to make a summary from multiple news articles. In terms of language-dependent factors, languages without word boundary markers, e.g. Chinese, Tibetan, Japanese, and Korean, trigger difficulties in recognizing or segmenting words and phrases while languages without sentence boundary markers (later called non-segmented languages), e.g. Khmer, Laos, Thai, and Vietnamese, additionally cause problems in recognizing or segmenting sentences. In a language with explicit sentence boundary, summarization usually considers sentences as units for performing summarization. However, a summarization process in a non-segmented language is inherently complicated. One can suffer from how to segment a running text into acceptable units due to high ambiguity situation. Specially for Thai language, due to lack of word/phrase/sentence boundary, summarization of Thai multiple documents has several challenges in processes.

In this work, we introduce Thai Elementary Discourse Unit (TEDU) and then present a three-stage method of Thai multi-document summarization, i.e. unit segmentation, unit-graph formulation, and unit selection for summarization. We investigate three different granularities of units; (1) TEDUs, (2) combined TEDUs, and (3) paragraphs. To evaluate our method, a number of experiments are conducted using fifty sets of Thai news articles, the model summaries of which are given. Three measures of ROUGE-1, ROUGE-2, and ROUGE-SU4 are used as the performance metrics.

Section 2 describes related work on multi-document summarization. Thai elementary discourse unit (TEDU) is defined in Section 3. Our method on Thai multi-document summarization is presented in Section 4. Experimental setup is shown in Section 5. Experimental result and discussion are presented in Section 6. Finally, conclusion and future work are in Section 7.

2 RELATED WORK

In the last decades, previous work placed interest on scientific documents and later on news articles due to their apparent structure. Various paradigms are proposed to extract salient sentences from a single document by making use of features like the skeleton of the document [28], word and phrase frequency [29], and key phrases as well as position in the text [30].

For multi-document summarization, centroid-based techniques to generate a composite sentence from each cluster are proposed by [20]. They used features to modify sentence in each cluster. Barzilay et al. [1] 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. Carbonell and Goldstein [2] 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 [14] presented an information extraction framework for summarization as well as a graph-based method to find similarities and dissimilarities in pairs of documents. As another work on summarization, Latent Semantic Analysis (LSA) was applied for generating a summary [24].

For Thai language, there have been very few works on Thai summarization since Thai texts are structurally flexible and complicated, as well as techniques and tools for basic text processing in Thai are still in their infancy stage. As an early work on Thai text summarization, Jaruskulchai and Kruengkrai [8] proposed a paragraph-based summarization which selects top-n paragraphs by formulating the importance level of a paragraph with the integration of local and global scores calculated from words in the paragraph. However, some limitations of this approach are that its processing unit is set to paragraph, duplication of units is not taken into account, and its primary target is summarizing a single document. As another work on paragraph-based summarization, Thangthai and Jaruskulchai [25] studied effect of three common parameters, i.e. stopword elimination, word segmentation, and word frequency (of six types), on paragraph selection for generating a Thai news summary in three genres of news documents, with consideration of Latent Semantic Analysis (LSA) as dimensionality reduction. However, even the authors claimed that LSA was considered but there was no comparison to evaluate the validity of the dimensionality reduction approach and the evaluation was not systematically performed. Moreover, Ketui and Theeramunkong [9] presented a more sophisticated weighting system with iterative weight calculation. They also proposed two summarization methods, called inclusion-based and exclusion-based selections to pick up a set of candidate paragraphs from multiple news documents for a summary. The methods were evaluated with a small set of politic news articles. However, summarization based on paragraph is forced to select a whole paragraph for summary, even only some parts in a paragraph are important and some are not. To avoid this restriction, two recent works [5, 22] proposed methods that select significant segments instead of paragraphs for a summary. A segment was defined as a sequence of words, delimited by stop marks (‘.’, ‘?’, ‘;’, ‘:’, ‘-’, and whitespace) after pre-processing word identification, stopword elimination [6], and stemming [19]. In this work, twenty-two content-based features and one graph-based feature were proposed for ranking segments based on their significance. Moreover, two alternative node ranking algorithms, topic sensitive PageRank algorithm [7] and Hopfield network algorithm [22] were proposed to rank the segmented text and then generate a summary for a Thai single document. However, using only punctuation marks for splitting a running text into segments and then selecting a subset of segments for summary may not be realistic since a summary generated from such subset of segments is usually not readable. As one solution, Sukvaree et al. [23] presented an EDU-based summarization approach. This approach extracted the elementary discourse units (EDUs) (i.e. the

minimal building blocks of a discourse tree connected with their relation [3]), identified their discourse coherence, and organized them into a spanning tree based on the well-known rhetorical structure theory [15]. However, the graph-based approach can specify on the maximal number of common and different sentences to control the output. This approach used these structures to actually compose abstractive summaries rather than to extract sentences from the text. In this work, unit segmentation, unit graph formulation, and unit selection constitute Thai Multi-Document summarization.

Type	Syntactic Unit	Cue or Device		Example
TEDU-1	Simple clauses	SV or SVO	SV	:[ถนน, S| *ลื่น*, V][road, S | *slip*, V]
		structure	SVO	:[ตำรวจ, S | *จับ*, V | คนร้าย, O][police, S | *arrest*, V | thief, O]
TEDU-2	Subject zero	Omission of	ϕVO	:[{ϕ} | *ตั้ง*, V | กรรมการ, O][{ϕ} | *appoint*, V | committee, O]
	anaphora clauses	sentential subject	ϕV	:[{ϕ} | *สอบสวน*, V][{ϕ} | *ask*, V | staff, O]
TEDU-3	Clauses with	Speech or cognitive	SV_AX	:[นายอภิสิทธิ์, S | ***กล่าว***, V_A | **ว่า**, X]
	attribution verbs	verbs. Eg. 'กล่าว' (say),		:[Mr.Apisit, S | ***say***, V_A | **that**, X]
		'หวัง' (hope), 'รู้สึก' (feel),	SV_A	:[ดีแทค , S | ***หวัง***, V_A]
		'คิด' (think), 'ว่า' (say/that)		:[Dtac (company), S | ***hope***, V_A]
TEDU-4	Comparative	A clause expressing	SVV_CRO	:[รายรับ, S | *ลดลง*, V | **มาก**, V_C | **กว่า**, R | รายจ่าย, O]
	clauses	comparison.		:[Income, S | *decrease*, V | **more**, V_C | **than**, R | payment, O]
		Eg. 'กว่า' (than)	SV_CRO	:[ราคาเนื้อหมู, S | ***แพง***, V_C | **กว่า**, R | เนื้อไก่, O]
				:[pork price, S | ***expensive***, V_C | **than**, R | chicken, O]
TEDU-5	Question clauses	A clause with a	WVO	:[**ใคร**, W| *ตั้ง*, V | กรรมการ, O]
		question word. Eg.		:[**Who**, W | *appoint*, V | committee, O]
		ใคร' (who),'อะไร' (what),	SVW	:[ตำรวจ, S | *จับกุม*, V | **ใคร**, W]
		'เมื่อไร' (when),'ที่ไหน' (where)		:[police, S | *arrest*, V | **whom**, W]
TEDU-6	Embedded conjunction	A clause with an	SCVO	:[ขนม, S | **เลย**, C | *ตก*, V | พื้น, O]
	clauses	embedded conjunction.		:[Snack, S | **then**, C | *fall*, V | floor, O]
		Eg. 'ก็', 'เลย',	SCV	:[ถนน, S | **จึง**, C | *ลื่น*, V]
		'จึง' (then)		:[road, S | **then**, C | *slip*, V]
TEMPP	Temporal phrases	A phrase with a	P_TDMY	:[**ในวันที่**, P_T | 1, D | มกราคม, M | 2555, Y]
		date/time. Eg.		:[**In date**, P_T | 1, D | January, M | 2012, Y]
		Eg. ในวันที่' (In date),	P_THMS$_T$	:[**เมื่อเวลา**, P_T | 10.00, HM | น., S_T]
		'เมื่อเวลา' (At time), 'น.' (o'clock)		:[**At time**, P_T | 10.00, HM | O'clock, S_T]
SPATP	Spatial phrases	A phrase with a	P_SLLLL	:[**ใน**, P_S | หมู่บ้านราชพฤกษ์, L | ถ.ติวานนท์, L | อ.เมือง, L | จ.ปทุมธานี, L]
		location/place.		:[**In**, P_S | Ratchpruek Village, L | Tiwanont Rd., L
		Eg. 'ใน' (in), 'ที่' (at),		:[Muang A., L | Patumthani, L]
		'บน' (on)	P_SLN	:[**บน**, P_S | ถนนสายรังสิต-นครนายก, L | ฟิวเจอร์พาร์ครังสิต, N]
				:[**on**, P_S | Rangsit-Nakornnayok Rd., L | Future Park Rangsit, N]
CONJP	Conjunction phrases	A phrase with a	P_CC	:[**โดย**, P_C | **หลังจากนั้น**, C]
		conjunction.		:[**By**, P_C | **afterwards**, C]
		Eg. 'โดยเบื้องต้นนั้น' (Nevertheless),	CS_C	:[**นอกจากนี้**, C | **ที่ผ่านมา**, S_C]
		'เมื่อทั้งๆที่' (Even if)		:[Moreover, C | previously, S_C]
EMBP	Embedded phrases	A phrase with a	E_P	:[**ที่/ซึ่ง/อัน**, E_P]
		relative pronoun.		:[**which/that/who**, E_P]
TEDU-LP-1	Clausal subject/object	A verb with a	P_{NC}V	:{**การ**, P_{NC} | ศึกษา, V} N: education
		nominal prefix		:{**-ion**, P_{NC} | educate, V}
		Eg. 'การ', 'ความ',	P_{NC}V	:{**ความ**, P_{NC} | รู้สึก, V} N: feeling
		({-ing}, {-ion})		:{**-ing**, P_{NC} | feel, V}
TEDU-LP-2	Synthetic nominal	A noun phrase with SV	P_{NS}V	:{**นัก**, P_{NS} |เรียน, V} N: student
	compounds	or SVO structure		:{**person**, P_{NS} | study, V}
		Eg. 'ช่าง' (man), 'เด็ก' (boy),	$P_{NS}VS_{NS}$	:{**ที่**, P_{NS}(S) | จอด, V | รถ, S_{NS}(O)} N: car park
		'นัก' (person), 'ที่' (place)		:{**place**, P_{NS}(S) | park, V | car, S_{NS}(O)}

Table 1. Six types of TEDUs, four types of COMPs, and two types of TEDU-LPs, enhanced from [3] for Thai unit definition. Here, a square bracket specifies a TEDU, a parenthesis surrounds a COMP, a brace denotes a TEDU-LP, an italicized item represents a verbal unit, and a bold-faced item displays a discourse cue.

3 THAI ELEMENTARY DISCOURSE UNIT (TEDU)

In the past, there has been a number of works on extracting Elementary Discourse Units (EDUs) from Thai texts, such as extraction from an agriculture corpus [4] and a written family law text [21]. The definition of Thai EDUs (TEDUs) in those works and the definition of English EDUs [3, 15] are considered. We have defined a set of TEDUs to reflect special characteristics in the Thai language in this section. In our framework, a TEDU is defined to represent a single event and usually contains a predicate (i.e. a verb). A Thai text is basically composed of continuously connected TEDUs, and sometimes there is a common phrase (COMP), such as a spatial or temporal phrase, a conjunction phrase, and an embedded phrase, between two TEDUs in order to specify relations among them. However, in Thai language, some word sequences look syntactically similar to a TEDU by containing a verb as its component, but they are not TEDUs. Some examples of such sequences, called a TEDU-like phrase/word (TEDU-LP), are clausal subjects/objects or synthetic nominal compounds. In our framework, TEDUs are composed of nominal and verbal units, mostly in the form of Subject-Verb-Object (SVO), sometimes together with some functional words as complimentary units. Here, a nominal unit, as for a subject or an object, is either a head noun, pronouns with numerals and/or classifiers, or nouns with determiners and/or adjectives. Sometimes such subjects and objects can be omitted, resulting in VO, SV or V structures. A verbal unit may be a simple verb or a verb with some auxiliary units. Moreover, in Thai texts, it is possible to have concatenated verbs; called a serial verb which signifies a relative action in order. In this work, since we principally set one verb as one TEDU, serial verbs will be broken down to several units as basic procedure. To cope with units in a Thai running text, we have defined six types of TEDUs (TEDU-1 to TEDU-6), four types of COMPs (TEMPP, SPATP, CONJP, and EMBP), and two types of TEDU-LPs (TEDU-LP-1 to TEDU-LP-2) as shown in Table 1. Moreover, TEDUs are based on English grammatically or lexical signal [3] such as clauses with attribute verbs, comparative clauses, question clauses, and conjunction phrases. In Thai, it cannot use punctuation for TEDU segmentation in order to focus on the verbal unit per TEDU.

4 THAI MULTI-DOCUMENT SUMMARIZATION

This section presents our Thai multi-document summarization model which is composed of three processes; 1) unit segmentation, 2) unit graph formulation, and 3) unit selection. In the unit segmentation (1), a Thai running text is segmented into a sequence of tractable units and tagged with part-of-speech (POSs) and named entities (NEs) (1.1). Three alternative forms, called TEDUs, combined TEDUs, and paragraphs are proposed for segmenting a Thai running text into units (1.2). In the unit graph formulation (2), a graph of units is constructed by conceptualizing a unit as a weighted node in a graph (2.1) and a relationship of two nodes is formulated as a weighted link between the nodes (2.2). The weight of a node or a link is de-

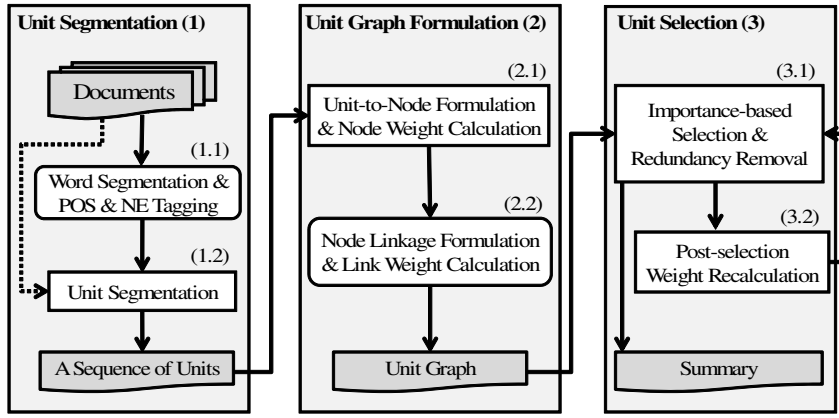

Figure 1. Three main processes with their subprocesses in the summarization model

termined by considering its importance, i.e. its contribution in the graph. In the unit selection (3), a number of important nodes and links are selected by considering importance level of nodes or links, together with redundancy among units (nodes) (3.1), and focusing on the node weight recalculation (3.2). Figure 1 displays subprocesses in the unit segmentation, the unit graph formulation, and the unit selection. Their details are shown in the next subsections.

4.1 Unit Segmentation

After we have proposed a so-called predicate-based segmentation (TEDU) in the previous section, Thai running text is split into units. Besides segmenting TEDUs, a combined TEDU (CTEDU) is composed of two TEDUs connecting them with some clues. Their processes are described in this subsection.

4.1.1 Word Segmentation and POS/NE Tagging

In the past, a number of techniques were implemented as tools for Thai word segmentation, POS tagging, and NE tagging, such as SWATH [16] and Thai E-Class [27]. Recently, the Thai E-Class has been developed as a tool for segmenting a Thai running text into words, and tagged each of them with POSs, NEs, and semantic roles in the 4W1H (Who, What, Where, When, How) format. In our approach, word segmentation and their tagging of POSs, NEs, and 4W1H are performed as a preprocessing for recognizing TEDUs, COMPs, and TEDU-LPs as shown in the next subsection.

4.1.2 Unit Segmentation

In this step, we apply a set of simple meta rule, called left-to-right longest matching [13] to uniquely determine segmentation of Thai running text into TEDU-LPs, TEDUs and COMPs. Later we use the detected TEDUs and COMPs as units for summarization. As an alternative larger unit, it is possible to combine strongly-related consecutive TEDUs into a so-called combined TEDU by a set of predefined rules on discourse markers or connectors, or even simply use a paragraph as a unit. Therefore, when there exist some clues of connection between them, two TEDUs are merged to form a combined TEDU, which is a larger unit. By the type of clues, we merge two TEDUs which have some connection clues in-between to form a combined TEDU (CTEDU). In summary, three types of units used in this work are TEDU + COMP (Thai EDU and Common Phrase), CTEDU (combined Thai EDU), and PARA (paragraph).

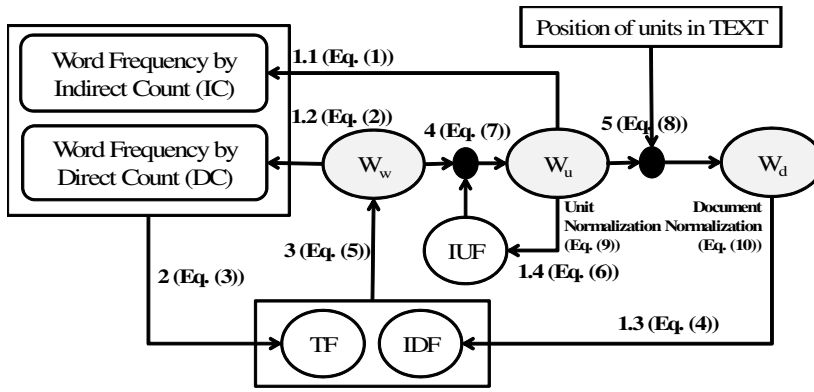

Figure 2. Iterative weighting: a calculation flow

4.2 Unit Graph Formulation

After the running text is split into tractable units; TEDU + COMP, CTEDU, or PARA, we need a model to determine importance of units and select the most suitable ones for a summary. In addition, a graph model is proposed to express units and their relations extracted from multiple documents targeted for summarization. In our approach, a node in the graph corresponds to a unit while a link in the graph expresses connections between two units.

4.2.1 Node Weight Calculation

As the most naive method, it is possible to apply TF-IDF (Term Frequency times Inverse Document Frequency) to weight words in a document and then weight a unit

by calculating the summation of weights of all words in that unit. As a more sophisticated weighting method, we have proposed a so-called iterative weighting [9] to obtain more accurate weights of units by considering importance of documents, units, and words and reflect them when we weigh units. In this method, besides IDF (inverse document frequency), a so-called inverse unit frequency (IUF) is introduced to express the frequency of units, instead of documents, that a word appears. The formulation of node weight calculation can be summarized as follows.

Let $C = \{c_1, c_2, c_3, \ldots, c_{|C|}\}$ be a whole set of $|C|$ documents, $D = \{d_1, d_2, d_3, \ldots, d_I\} \subset C$ be a set of $(I = |D|)$ related documents to be summarized, $L = \{w_1, w_2, \ldots, w_K\}$ be the set of all possible words where K is the number of possible words, and suppose that a document $d_i \in D$ consists of J_i units, i.e. $d_i = \{u_{i1}, u_{i2}, u_{i3}, \ldots, u_{iJ_i}\}$, and a unit $u_{ij} \in d_i$ can be represented by a word occurrence vector $\vec{u}_{ij} = \{N_{ij}(w_1), N_{ij}(w_2), N_{ij}(w_3), \ldots, N_{ij}(w_K)\}$, showing the occurrence frequency of each word $w_k \in L$ in the unit u_{ij}. Moreover, a document can also be expressed by a weight vector $\vec{d}_i = \{W_w(w_1, d_i), W_w(w_1, d_i), W_w(w_1, d_i), \ldots, W_w(w_K, d_i)\}$, showing the weight of each word $w_k \in L = \{w_1, w_2, \ldots, w_K\}$, in the document d_i. Note that the word weights are assigned, regardless of unit consideration. As an iterative method, all weights, including word weights $(W_w(w_k, d_i))$, unit weights $(W_u(u_{ij}))$ and document weights $(W_d(d_i))$, are associated with a time-series factor, i.e. $W_w^{(t)}(w_k, d_i), W_u^{(t)}(u_{ij})$, $W_d^{(t)}(d_i)$, where t is the t^{th} iteration.

For initialization, we can assign word weights, unit weights, and document weights to an equal value, i.e. $W_w^{(0)}(w_k, d_i) = \frac{N_i(w_k)}{\sum_k N_i(w_k)}$, $W_u^{(0)}(u_{ij}) = \frac{1}{J_i}$, and $W_d^{(0)}(d_i) = \frac{1}{I}$ (i.e. $\frac{1}{|D|}$) respectively, where $N_i(w_k)$ is the total number of the word w_k that occurs in d_i and $\sum_k N_i(w_k)$ is the total number of all words in d_i, i.e. the document size. Our proposed iterative weighting is performed by recalculating word weights, unit weights, and document weights in order, as denoted by Steps 3, 4 and 5 in Figure 2. However, some preliminary tasks, i.e. Steps 1 and 2 in Figure 2, have to be executed. The details of these steps are described below.

In the first step, we calculate the new weight for a particular word (w_k) in document d_i, the occurrence frequency of the word has to be estimated. There are two ways to calculate word frequency in a document; one is calculated indirectly from units in the document, i.e. $IC_w^{(t+1)}(w_k, d_i)$ (Step 1.1 and Equation (1)) while the other is derived directly from words in the document, i.e. $DC_w^{(t+1)}(w_k, d_i)$ (Step 1.2 in Figure 2 and Equation (2)), as shown in the following two equations.

$$IC_w^{(t+1)}(w_k, d_i) = N^u(d_i) \times \sum_{u_{ij} \in d_i} N_{ij}(w_k) \times W_u^{(t)}(u_{ij}), \tag{1}$$

$$DC_w^{(t+1)}(w_k, d_i) = N^w(d_i) \times W_w^{(t)}(w_k, d_i), \tag{2}$$

where $N^u(d_i)$ is the total number of units appearing in the document d_i, $N_{ij}(w_k)$ is the total number of a particular word w_k appearing in the unit u_j of the document d_i, $W_u^{(t)}(u_{ij})$ is the previous weight of the unit u_{ij}, $N^w(d_i)$ is the total number of words

appearing in the document d_i and $W_w^{(t)}(w_k, d_i)$ is the previous weight of the word w_k in the document d_i.

These two estimation methods are combined to predict term frequency (TF) of the word w_k in the document d_i (Step 2 in Figure 2), as shown in Equation (3).

$$TF^{(t+1)}\left(w_k, d_i\right) = \sqrt{IC_w^{(t+1)}\left(w_k, d_i\right) \times DC_w^{(t+1)}\left(w_k, d_i\right)}, \tag{3}$$

where $TF^{(t+1)}\left(w_k, d_i\right)$ is the expected frequency of a particular word w_k, calculated by geometric average of weights estimated from the two methods, i.e. Equations (1) and (2)).

Besides TF, we apply a generalization of the inverse document frequency (IDF) where document weights are also taken into consideration. The IDF of the word w_k is the weight that reflects the effect of the number of documents the word appears (Step 1.3 in Figure 2), as shown in Equation (4).

$$IDF^{(t+1)}(w_k) = \log\left(1 + \frac{\sum_{d_i \in D} W_d^{(t)}(d_i)}{\sum_{d_i \in D \wedge N_i(w_k) \neq 0} W_d^{(t)}(d_i)}\right), \tag{4}$$

IDF corresponds to one plus the logarithm of the summation of the weights of all documents in the corpus, divided by the summation of the weights of the documents that include the word w_k. As stated above, the initial document weight $W_d^{(0)}(d)$ is assigned as the total of documents in the corpus where $W_d^{(0)}(d) = \frac{1}{|D|}$. The weight of a word w_k in the document d_i ($W_w^{(t+1)}(w_k, d_i)$) is calculated by the multiplication of TF and IDF (Step 3 in Figure 2), as shown in Equations (3) and (4), where the TF is normalized, as shown in Equation (5).

$$W_w^{(t+1)}\left(w_k, d_i\right) = \frac{TF^{(t+1)}(w_k, d_i)}{\sum_{w \in d_i} TF^{(t+1)}\left(w, d_i\right)} \times IDF^{(t+1)}\left(w_k\right). \tag{5}$$

Note that a weight of any word $W_w^{(t+1)}\left(w_k, d_i\right)$ has a value between 0 and IDF.

To update the weight of a unit, an inverse unit frequency IUF is applied together with the updated weights of words (Step 4 in Figure 2). The IUF of a word w_k is an analogy of the IDF of a word by replacing document frequency with unit frequency (Step 1.4 in Figure 2), as shown in Equation (6).

$$IUF^{(t+1)}\left(w_k, d_i\right) = \log\left(1 + \frac{\sum_{u_{ij} \in d_i} W_u^{(t)}(u_{ij})}{\sum_{u_{ij} \in d_i \wedge N_{ij}(w_k) \neq 0} W_u^{(t)}(u_{ij})}\right), \tag{6}$$

IUF is defined as one plus the logarithm of the summation of the weights of all units in a particular document d_i, divided by the summation of the weights of the units that include the word w_k. The unnormalized weight of a unit u_{ij} in the document d_i ($W_u^{'(t+1)}(u_{ij})$) corresponds to the summation of the multiplication of the weight (i.e. Equation (5)) and the IUF (i.e. Equation (6)) of each word in the unit (Step 4 in Figure 2), as shown in Equation (7).

$$W_u^{'(t+1)}(u_{ij}) = \sum_{w_k \in L \wedge N_{ij}(w_k) \neq 0} W_w^{(t+1)}(w_k, d_i) \times IUF^{(t+1)}(w_k, d_i). \quad (7)$$

Next, the unnormalized weight of a document ($W_d^{'(t+1)}(d_i)$) is updated by considering weights and positions of the units in the document (Step 5 in Figure 2). As weighting units based on unit position, we apply a simple cut-off linear model of $W_{loc}(u_{ij}) = MAX(0.3, (J_i + 1 - j)/J_i))$. The position-based factor is added due to the assumption that a unit occurring in the beginning of a document has higher importance than the latter units. By incorporating this position-based factor into the unit weights, the unnormalized weight of a document can be calculated by the summation of weights of all units in the document, as shown in Equation (8).

$$W_d^{'(t+1)}(d_i) = \sum_{u_{ij} \in d_i} W_u^{'(t+1)}(u_{ij}) \times W_{loc}(u_{ij}). \quad (8)$$

Finally, the unit and document weights ($W_u^{(t+1)}(u_{ij})$, $W_d^{(t+1)}(d_i)$) are normalized by the two following formulas:

$$W_u^{(t+1)}(u_{ij}) = \frac{W_u^{'(t+1)}(u_{ij})}{\sum_{u \in d_i} W_u^{'(t+1)}(u)}, \quad (9)$$

$$W_d^{(t+1)}(d_i) = \frac{W_d^{'(t+1)}(d_i)}{\sum_{d \in D} W_d^{'(t+1)}(d)}. \quad (10)$$

Weights of the nodes in the graph are iteratively updated until a certain number of iteration or the difference of summation of the document weight before and after updates is less than a threshold, say 0.5 in this work.

4.2.2 Link Weight Calculation

Besides node weights, it is worth investigating relations between two units (nodes). A link weight (relation strength) between two nodes (units) describes how much two units are identical or related. Although there have been several possibilities of relations definition between units, this work simply uses a common method, namely cosine similarity. A unit $u_{ij} \in d_i$ is formulated a word-weight vector as follows. $\vec{u}_{ij} = \{F_{ij}(w_1) \times W_w(w_1, d_i), F_{ij}(w_2) \times W_w(w_2, d_i), \ldots, F_{ij}(w_K) \times W_w(w_K, d_i)\}$, where, $W_w(w_k, d_i)$ is the weight of a word w_k in the document d_i and $F_{ij}(w_k)$ is a sort of frequency function. While $W_w(w_k, d_i)$ can be derived, e.g. from Equation (5), $F_{ij}(w_k)$ can be a binary frequency, term frequency or TF-IDF. The link weight between two units, u_{ij} and $u_{i'j'}$, is defined by the cosine similarity between the vectors of those two units.

$$\text{sim}(u_{ij}, u_{i'j'}) = \frac{\vec{u}_{ij} \cdot \vec{u}_{i'j'}}{|\vec{u}_{ij}| \cdot |\vec{u}_{i'j'}|}. \quad (11)$$

The cosine similarity ranges from 0 to 1 and the highest value indicates high similarity, implying that two units are duplicates or highly related.

4.3 Unit Selection

Given the unit graph derived from the process described in the previous subsection, unit selection is a task to select a set of suitable units for constructing a summary (Process 3 in Figure 1). In the past, several works [2] applied a straightforward method to select units based on their weights. In this work, we utilize a variant of the inclusion-based summarization method proposed in [9].

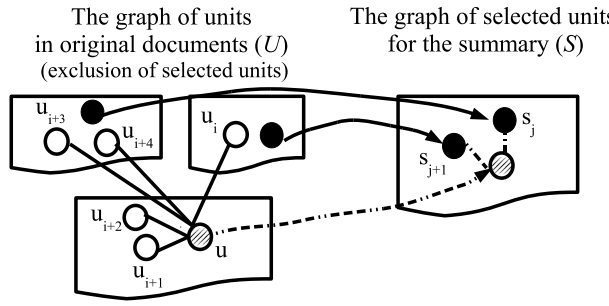

Figure 3. Overview of the inclusion-based selection. The black circles represent selected units while the shaded circle means the to-be-included unit. The left part indicates a set of original documents with their units (nodes) while the right part presents the summary graph with selected units and to-be-selected unit.

Our unit selection is an iterative process to select potential units u based on priority (weight). The unit will not be included if it is duplicated with some selected units in the summary S as illustrated in Figure 3 and Table 2. Finally, the weights of the rest of nodes u_i in the graph of units in documents G are recalculated. In the algorithm, the number of units to be selected is set (line 1). The most important nodes are repeatedly added one-by-one into the summary S and deleted from a set of unselected units U (line 2–7) until the number of selected units reaches the predefined compression rate. After the node addition, the weight of each unselected unit is recalculated (line 6). When the number of selected units satisfies the predefined compression rate, the algorithm returns the graph of summary S (line 8). In details, three basic concepts of our inclusion-based summarization approach are discussed below.

Importance-based selection: A unit with a higher weight (importance) has a higher priority to be selected. In this work, two hypotheses are investigated.

1. A high weighted unit usually includes more important words or more specific target words (TF-IDF or iterative weight).
2. A preferable unselected unit is a node with high similarity to all other unselected units. As well, the unit close to the centroid of the unselected units should be included in the summary.

Algorithm: The inclusion-based unit selection

Input: a set of units $U = \{u_1, u_2, \ldots, u_I\}$,
a set of unit weights $W = \{w_1, w_2, \ldots, w_I\}$,
a set of link weights $E = \{e_{11}, e_{12}, \ldots, e_{21}, e_{22}, \ldots, e_{II}\}$
(i.e., similarity between nodes),
a predefined compression rate cr

Output: a set of selected units for the summary S

1 : **SET** the number of units to be selected n_s to $(I \times cr)$
2 : **DO**
3 : **SELECT** the node $u_s \in U$ with the highest weight $(s = \mathbf{argmax}_i[w_i])$
(by considering three factors: importance (W), centroid (E), and redundancy (E).)
4 : **ADD** u_s into the summary S $(S = S \cup \{u_s\})$
5 : **DELETE** u_s from the set of unselected units U $(U = U - \{u_s\})$
6 : **RECALCULATE** the weight (w_i) of each unselected unit (u_i) using the unit weight W and the link weight E*(by considering its similarity (e_{ij}) to the selected unit, compared with other unselected units.)*
7 : **UNTIL** the total number of nodes is greater than or equal to n_s
8 : **RETURN** S

Table 2. Algorithm of inclusion-based unit selection

Redundancy removal: Two units with an identical content or a highly similar content should not be selected simultaneously. In other words, it is reasonable to eliminate content redundancy in order to have a good short summary.

Post-selection weight recalculation: After a selected unit is included in a summary, its selection affects possibilities of other units. Moreover, after a unit is selected, selection of probability of an unselected unit depends on the ratio of the similarity of the unselected unit against the selected unit, and the similarity of the unselected unit against the other unselected units.

Reflecting the first concepts on weighting units, the unit selection can be formulated as follows. Here, the original weighting of a unit $W_u(u)$ is modified to reflect a centroid-related factor and a redundancy-related factor as shown in Equation (12). The best unit $(\hat{u})$ can be selected by maximizing the value in the equation, where the three terms indicates original weight, centroid-related and redundancy-related factors, in order (line 4 in Table 2).

$$\hat{u} = \arg\max_u \left[W_u(u) \times \frac{\sum_{i=1, u_i \neq u}^{|U|} \mathrm{sim}(u_i, u)}{|U| - 1} \times (1 - \max_j(\mathrm{sim}(s_j, u))) \right] \tag{12}$$

In the past, TF-IDF was usually used to express such importance levels of words by using term (word) frequency and inverse document frequency. As an alternative mentioned in Section 4.2.1, iterative weighting [9] may be used. Such weight helps selecting suitable units for summarization. The second term expresses the average

similarity between the current unit and the other units. Its high value indicates the closeness to the centroid of the unselected units. The third term represents the level of content redundancy. In the extreme case, if the unit's content is similar to content of any unit in the set of selected units, the maximum is one and the term will become 0. After a selected unit is included in a summary, its selection affects possibilities that other units will be selected. As shown in Equation (13), recalculation of node weights (${W_{re}}^{(t)}(u)$) is done by decreasing the weight of an unselected node ${W_{re}}^{(t)}(u)$ by the factor of the ratio of similarity between that node (u) and the selected node s in set S and other nodes (u_i).

$$ {W_{re}}^{t}(u) = W_u^{(t)}(u) \times \left(1 - \frac{\mathrm{sim}(u,s)}{\sum_{i=1,u_i \neq u}^{|U|} \mathrm{sim}(u,u_i)}\right). \tag{13} $$

Then the new node weight $W_u^{(t+1)}(u)$ is normalized by Equation (14) (line 6 in Table 2). The steps 4–7 in the algorithm are iterated to select the next node until the compression rate reaches the predefined value.

$$ W_u^{(t+1)}(u) = \frac{{W_{re}}^{(t)}(u)}{\sum_{i=1}^{|U|} {W_{re}}^{(t)}(u_i)}. \tag{14} $$

5 EXPERIMENTAL SETUP

This section describes the dataset from a standard corpus (namely THAI-NEST), experimental settings following the proposed summarization, and evaluation method with ROUGE-based measures.

5.1 Dataset and Preprocessing

This work utilizes the THAI-NEST corpus developed in [26] which comprises 10 000 news articles in seven categories: crimes (CR), sports (SP), foreign affairs (FO), politics (PO), entertainment (EN), economics (EC), and education (ED), gathered from seventeen on-line news sources. Later, a method for discovering document relations in [11] is applied to find relation between news documents and to group highly related news documents into a data set for summarization. While most previous works focused on finding document relations judged to be either relevant or non-relevant, this work classified documents into three main types of relations: (1) completely related (CR), (2) somehow related (SH), and (3) unrelated (UR). In this work, we randomly selected 50 sets of related documents with CR and SH relations for testing our proposed graph-based summarization approach. Given each set of related documents, the documents were tagged with POSs/NEs by Thai E-Class [27].

Later a Thai running text with POSs and NEs tagging is segmented into TEDUs, COMPs, and TEDU-LPs by using 446 context-free grammar rules (CFG rules) with

a bottom-up chart parser [10] with the longest matching technique [13] to detect TEDUs in a text, together with their structures. Here, the CFG rules are built based on the syntactic categories defined by a Thai E-Class [27]. We applied three groups of CFG rules: 342 rules for TEDUs, 95 rules for COMPs, and 9 rules for TEDU-LPs. As described in Section 4.1, we have proposed to investigate three types of units, i.e. (1) TEDU + COMP, (2) CTEDU, and (3) paragraph. Conceptually, the TEDUs and COMPs can be detected after recognizing TEDU-LPs while CTEDU can be constructed by merging two related TEDUs using COMPs. The last type of units, paragraph, can be simply detected by line breaks and indents. If the text is in the HTML format, common markers are <p> or
 tags since both tags actually indicate a new paragraph in a Thai web text. To form a reference summary (i.e., model summary) for evaluation, we have asked a number of Thai language experts in the Faculty of Liberal Art, Thammasat University to manually summarize the prepared set of news articles as gold standard for evaluating system results. The summarizers were instructed to construct an abstractive-based summary with the size of 20–100 words for each of the fifty datasets. The summaries contain main contents in the original documents. Some discourse markers are added to connect clauses. These reference summaries are used for evaluating a summary obtained from the system.

Group of		Original Documents							Reference Summary	
Datasets		Size(KB)	#Words	#TEDUs	#COMPs	#CTEDUs	#PARAs	#DOCs	Size(KB)	#Words
2–3 docs/dataset	Sum	299.9	14 409	4 091	1 958	2 046	225	87	53.5	2 683
(41 datasets)	Avg	7.3	351.4	99.8	47.8	49.9	5.5	2.1	1.3	65.4
	Max	17.1	880	277	104	139	13	3	2.7	137
	Min	4.0	190	41	18	21	2	2	0.7	24
4–6 docs/dataset	Sum	105.4	5 226	1 531	690	766	73	25	10.8	550
(6 datasets)	Avg	17.6	871.0	255.2	115.0	127.6	12.2	4.2	1.8	91.7
	Max	25.7	1 373	408	176	204	15	5	2.8	149
	Min	12.0	621	162	82	81	10	4	1.2	68
7–15 docs/dataset	Sum	77.8	4 146	1 135	482	568	53	32	3.2	174
(3 datasets)	Avg	25.9	1 382.0	378.3	160.7	189.2	18.0	10.7	1.1	58.0
	Max	31.0	1 628	468	211	234	21	15	1.3	77
	Min	21.3	1 133	327	99	164	13	7	0.9	43
All datasets	Sum	483.1	23 781	6 757	3 130	3 380	351	144	67.4	3 407
(50 datasets)	Avg	9.7	475.6	135.1	62.6	67.6	7.0	2.9	1.3	68.1
	Max	31.0	1 628	468	211	234	21	15	2.8	149
	Min	4.0	190	41	18	21	2	2	0.7	24

Table 3. Characteristics of the 50 experimental datasets

Table 3 displays characteristics of 50 experimental datasets grouped by the number of documents per set, including document size, number of words, number of TEDUs, number of COMPs, number of CTEDUs, number of paragraphs, number of documents as well as the size of reference summary and number of words in reference summary. The details show sum, average, maximum, and minimum values of

Type		#Words	#Units	Type		#Words	#Units	Type		#Words	#Units
TEDU-1	Sum	3 846	1 242	TEDU-5	Sum	104	47	TEMPP	Sum	1 033	562
	Avg	76.9	24.8		Avg	2.1	0.9		Avg	20.7	11.2
	Max	302	94		Max	12	6		Max	76	49
	Min	18	6		Min	0	0		Min	2	1
TEDU-2	Sum	8 138	3 886	TEDU-6	Sum	443	124	SPATP	Sum	1 304	716
	Avg	162.8	77.7		Avg	8.9	2.5		Avg	26.1	14.3
	Max	621	282		Max	25	7		Max	108	59
	Min	47	21		Min	0	0		Min	0	0
TEDU-3	Sum	2 731	1 453	TEDU-LP-1	Sum	1 144	572	CONJP	Sum	1 504	1 412
	Avg	54.6	29.1		Avg	10	5		Avg	30.1	28.2
	Max	167	94		Max	14	7		Max	88	83
	Min	5	3		Min	6	3		Min	10	9
TEDU-4	Sum	25	5	TEDU-LP-2	Sum	792	345	EMBP	Sum	440	440
	Avg	0.5	0.1		Avg	13.8	5.5		Avg	8.8	8.8
	Max	10	2		Max	39	8		Max	31	31
	Min	0	0		Min	7	3		Min	0	0

Table 4. Numbers of words and units in 50 datasets grouped by types of TEDUs, COMPs, and TEDU-LPs

each dataset group. Our fifty datasets used for experiments contain 144 documents and 23 781 words. The size of all datasets has approximately 483.1 KB and the average size per set is 9.7 KB. The maximum size equals 31.0 KB while the minimum one is 4.0 KB. The number of TEDUs is greater than twice of CTEDUs, implying that CTEDUs are constructed from two TEDUs. At least two paragraphs appear in each set while the average number of paragraphs in each set is 7 units. Besides, we consider three groups of datasets:

1. the group with 2–3 documents/set,
2. the group with 4–6 documents/set, and
3. the group with 7–15 documents/set.

The group with 2–3 documents per set is the majority (41 out of 50 data sets) in our experiments. Not surprising, the average size of data set of '7–15 documents/set' (i.e. 25.9) is larger than that of '4–6 documents/set' (i.e. 17.6) and it is much larger than that of '2–3 documents/set' (i.e. 7.3) since there are more documents in one data set. This outlook is also true for the numbers of words, TEDUs, COMPs, CTEDUs, and paragraphs. The last two columns in Table 2 expose characteristics of the reference summary of each dataset. The summary size of a data set of '2–3 documents/set' (i.e. 1.3) is not different from that of a data set of '7–15 document/set' (i.e. 1.1), even the summary size of a data set of '4–6 documents/set' (i.e. 1.8) is slightly higher. This trend is also true for the number of words in reference summaries.

Table 4 illustrates the detail result of unit segmentation of the fifty datasets, including the numbers of words and units in 50 datasets, grouped by types of TEDUs,

COMPs, and TEDU-LPs. It also presents the sum, average, maximum, and minimum values of each type. The most frequent units are of TEDU-2 (i.e. 3 886 units). On the other hand, TEDU-4 is the least frequent type. There are only five TEDU-4 units for the fifty datasets but each unit is quite long (i.e. on average $25 \div 5 = 5$ words/units). Although we have more TEDU-3 units than TEDU-1 units, on average a TEDU-3 unit ($2\,731 \div 1\,453 = 1.9$ words/unit) is shorter than a TEDU-1 unit ($3\,846 \div 1\,242 = 3.1$ words/unit). Moreover, TEDU-4, TEDU-5, and TEDU-6 units appear in only some datasets. For COMPs, the common phrases of conjunctions (CONJP) occur the most frequently in terms of units. It usually has one word per unit and it is used to connect two TEDUs. The temporal and spatial phrases (TEMPP and SPATP) include on average two words ($1\,033 \div 562 = 1.8$ and $1\,304 \div 716 = 1.8$) while the embedded phrases (EMBP) consist of only one word. 572 TEDU-LP-1 units and 345 TEDU-LP-2 units embed in TEDUs or COMPs.

5.2 Experimental Setting

To examine performance of the proposed methods, we have conducted four experiments using 50 sets of related news documents, containing 23 781 words. The first experiment aims to investigate summarization performance according to three unit types, four summarization factors, and five compression rates. Here, the three types of units are (1) TEDU + COMP, (2) CTEDU, and (3) PARA as stated in Section 4.1.2. The four summarization factors we consider are (1) node weighting, (2) importance-based selection, (3) redundancy removal, and (4) post-selection weight recalculation. For node weighting, two options are investigated, i.e. simple TF-IDF weighting ('T') and iterative weighting ('I'). For importance-based selection, we consider the simple highest-weight priority ('H') and an extension of the highest-weight priority with centroid preference ('C'). For the redundancy removal, we compare consideration of redundancy penalty ('P') to the non-consideration version ('D'). For post-selection weight recalculation, we also compare the recalculation case ('R') with its non-recalculation one ('N'). As the notation for the combination, four characters are used. For example, the notation of 'THPR' indicates the approach that uses TF-IDF as node weighting ('T'), the simple highest-weight priority ('H') as importance-based selection and considers redundancy penalty ('P') as well as weight recalculation of remaining nodes ('R'). For the compression rate, we examine five rates of 0.1, 0.2, 0.3, 0.4, and 0.5 since a summary should not be larger than a half of the original documents. Moreover, as an optimal case, we calculate upper-bound performance (later denoted by UPB) of summarization by starting from selecting the unit that obtains the highest ROUGE score, adding it into the summary and then selecting the next unit that makes us achieve the best performance when it is added into the summary. This greedy-based selection is performed repeatedly until reaching the target compression rate. The average of the results of all sixteen methods is also provided for reference. Here, the compression rate is defined as the ratio of the number of units in a summary to the number of units in the original documents. Moreover, we compare our summariza-

tion methods with a traditional graph-based ranking model called TexRank [17]. The sentence scoring function is known as the PageRank algorithm [18]. The second experiment targets to compare performance of the sixteen combinations of the four factors. For each factor, two alternatives are investigated by varying other factors and then comparing their results. Moreover, the performances of sixteen combinations are summarized and ranked to clarify which factor combination is optimal. In the third experiment, we perform a two-tailed t-test with the significance value of 0.05 to check win/loss/tie (W/L/T) among the top-8 methods obtained the second experiment, in order to make a detailed comparison. Here, we utilize the result of the five compression rates (0.1–0.5) for performance comparison under consideration of the three ROUGE values (ROUGE-1, ROUGE-2, and ROUGE-SU4) and the top-8 methods. As the last experiment, we investigate the effect of the number documents per dataset on summarization performance by classifying the fifty dataset into three groups; '2–3 documents/dataset', '4–6 documents/dataset', and '7–15 documents/dataset'.

5.3 Evaluation Method

To evaluate a summary output from a system, we use the reference summaries as described in Section 5.1. A reference summary is an abstractive-based summary (with consideration of semantics, i.e. what, where, who, whom, why, and how) constructed by requesting Thai language experts to manually summarize a set of related news articles into 20–100 words. In this work, we utilize a standard metric, called ROUGE [12], to evaluate a system's summarization result by comparing it with its reference summary. Among various types of ROUGE, this work uses ROUGE-1 (unigram-based co-occurrence statistics), ROUGE-2 (bigram-based co-occurrence statistics), and ROUGE-SU4 (skip-bigram plus unigram-based co-occurrence statistics) which are commonly used for evaluation. Originally developed by NIST, the ROUGE we used is a new variant of ROUGEs that considers precision (ROUGE-1_P, ROUGE-2_P, ROUGE-SU4_P), recall (ROUGE-1_R, ROUGE-2_R, ROUGE-SU4_R), and F-score (ROUGE-1_F, ROUGE-2_F, ROUGE-SU4_F). Their definitions are given below.

Precision			Recall		
ROUGE–1_P	=	$\frac{\lvert S_1\rvert}{\lvert S\rvert}$	ROUGE–1_R	=	$\frac{\lvert R_1\rvert}{\lvert R\rvert}$
ROUGE–2_P	=	$\frac{\lvert S_2\rvert}{\lvert S\rvert-1}$	ROUGE–2_R	=	$\frac{\lvert R_2\rvert}{\lvert R\rvert-1}$
ROUGE–SU4_P	=	$\frac{\lvert S_{SU4}\rvert}{\lvert S\rvert+4(\lvert S\rvert-3)+3+2+1}$	ROUGE–SU4_R	=	$\frac{\lvert R_{SU4}\rvert}{\lvert R\rvert+4(\lvert R\rvert-3)+3+2+1}$

Table 5. Evaluation methods of Precision (ROUGE-1_P, ROUGE-2_P, ROUGE-SU4_P) and Recall (ROUGE-1_R, ROUGE-2_R, ROUGE-SU4_R)

Let $S = s_1 s_2 s_3 \dots s_{|S|}$ be the system summary and $R = r_1 r_2 r_3 \dots r_{|R|}$ be the reference summary where s_i is the i^{th} word in the system summary and r_j is the

j^{th} word in the reference summary. Under this definition, the unigrams of the system summary that exist in the reference summary can be defined as $S_1 = \{s_i | s_i$ appears as a word in $R\}$, the unigrams of the reference summary that exist in the system summary as $R_1 = \{r_i | r_i$ appears as a word in $S\}$, the bigrams of the system summary that exist in the reference summary as $S_2 = \{s_i s_{i+1} | s_i s_{i+1}$ appears as a word pair in $R\}$, the bigrams of the reference summary that exist in the system summary as $R_2 = \{r_i r_{i+1} | r_i r_{i+1}$ appears as a word pair in $S\}$, the four-arbitrary-gap bigram and unigram of the system summary as $S_{SU4} = \{s_i | s_i$ appears as a word in $R\} \cup \{s_i s_{i+1} | s_i s_{i+1}$ is a word pair in R with skip allowance of $4\} \cup \{s_i s_{i+2} | s_i s_{i+2}$ is a word pair in R with skip allowance of $4\} \cup \{s_i s_{i+3} | s_i s_{i+3}$ is a word pair in R with skip allowance of $4\} \cup \{s_i s_{i+4} | s_i s_{i+4}$ is a word pair in R with skip allowance of $4\}$, the four-arbitrary-gap bigram and unigram of the reference summary as $R_{SU4} = \{r_i | r_i$ appears as a word in $S\} \cup \{r_i r_{i+1} | r_i r_{i+1}$ is a word pair in S with skip allowance of $4\} \cup \{r_i r_{i+2} | r_i r_{i+2}$ is a word pair in S with skip allowance of $4\} \cup \{r_i r_{i+3} | r_i r_{i+3}$ is a word pair in S with skip allowance of $4\} \cup \{r_i r_{i+4} | r_i r_{i+4}$ is a word pair in S with skip allowance of $4\}$. Here, ROUGE-1_F equals to $\left(2 \times \frac{\text{ROUGE–N}_P \times \text{ROUGE–N}_R}{\text{ROUGE–N}_P + \text{ROUGE–N}_R}\right)$. In the same way, ROUGE-2_F and ROUGE-SU4_F can be calculated but with their precision and recall.

6 EXPERIMENTAL RESULT AND DISCUSSION

6.1 Performance Investigation on Three Unit Types, Four Summarization Factors, and Five Compression Rates

In this experiment, we investigate and compare the performance of three types of units, four summarization factors, and five compression rates (0.1–0.5). The results are shown in Table 6-8. The superscripts are given to the highest and the second highest values in each method, respectively. Table 6 shows that TEDU + COMP with IHPR gets the highest ROUGE-1 performance at the compression rate of 0.4. For the unit type of TEDU-COMP, the methods with the simple TF-IDF (i.e. 'T***' in the 'Factor' column), achieve the best ROUGE-1 performance at the compression rate of 0.5 while those with the iterative weighting (i.e. 'I***' in the 'Factor' column) are superior at the compression of 0.4. The best performance is 0.2982, obtained when the iterative weighting, the redundancy removal, and weight recalculation (i.e. IHPR) are considered at the compression rate of 0.4. The PageRank (henceforth PR) is used as our baseline. At the compression rate of 0.4, PR achieves the highest R-1 of 0.2439. Its performance is lower than our method. In the same way, for the unit type of CTEDU, IHPR also gains the highest ROUGE-1 performance of 0.3194 at the same compression rate but higher than the case of TEDU + COMP. TCDN, TCDR, and TCPN achieve the lowest performance of R-1 at the compression rate 0.1. Their performance is lower than the baseline (PR). For the unit type of PARA (paragraph), IHPN is superior to other methods with ROUGE-1 performance of 0.3184. IHPR is slightly lower at the ROUGE-1 of 0.3116. In total, the iterative weighting, the

UT	Method	Compression Rate					Method	Compression Rate				
		0.1	0.2	0.3	0.4	0.5		0.1	0.2	0.3	0.4	0.5
TEDU	THDN	0.1775	0.2141	0.2432	0.2632^{2}	0.2752^{1}	IHDN	0.2217	0.2427	0.2547^{1}	0.2476^{2}	0.2472
+COMP	THDR	0.1835	0.2226	0.2509	0.2729^{2}	0.2846^{1}	IHDR	0.2375	0.2727	0.2891^{1}	0.2851^{2}	0.2797
	THPN	0.1845	0.2158	0.2461	0.2675^{2}	0.2835^{1}	IHPN	0.2402	0.2768	0.2892^{2}	0.2952^{1}	0.2878
	THPR	0.1814	0.2254	0.2537	0.2739^{2}	0.2865^{1}	**IHPR**	0.2437	0.2843	0.2968	$\mathbf{0.2982}^{1}$	0.2973^{2}
	TCDN	0.1530	0.2108	0.2430	0.2619^{2}	0.2788^{1}	ICDN	0.1542	0.1930	0.2047	0.2179^{2}	0.2225^{1}
	TCDR	0.1636	0.2207	0.2509	0.2746^{2}	0.2881^{1}	ICDR	0.1798	0.2190	0.2376	0.2419^{1}	0.2411^{2}
	TCPN	0.1580	0.2167	0.2475	0.2687^{2}	0.2876^{1}	ICPN	0.1943	0.2287	0.2400^{2}	0.2456^{1}	0.2385
	TCPR	0.1653	0.2247	0.2563	0.2781^{2}	0.2900^{1}	ICPR	0.2082	0.2449	0.2563^{1}	0.2486^{2}	0.2464
	AVG	0.1904	0.2321	0.2537	0.2650^{2}	$\mathbf{0.2709}^{1}$	UPB	0.4023	$\mathbf{0.4042}^{1}$	0.4039^{2}	0.3945	0.3647
	PR	0.2036	0.2312	0.2424	$\mathbf{0.2439}^{1}$	0.2433^{2}	PR	0.2036	0.2312	0.2424	$\mathbf{0.2439}^{1}$	0.2433^{2}
CTEDU	THDN	0.1915	0.2468	0.2752	0.2791^{2}	0.2846^{1}	IHDN	0.2400	0.2856^{1}	0.2778	0.2779^{2}	0.2672
	THDR	0.1887	0.2471	0.2734	0.2858^{1}	0.2851^{2}	IHDR	0.2309	0.2835	0.2991^{1}	0.2961^{2}	0.2860
	THPN	0.1937	0.2405	0.2735	0.2832^{2}	0.2879^{1}	IHPN	0.2233	0.2880	0.3111^{2}	0.3122^{1}	0.2986
	THPR	0.1858	0.2436	0.2723	0.2907^{1}	0.2883^{2}	**IHPR**	0.2212	0.2824	0.3186^{2}	$\mathbf{0.3194}^{1}$	0.3044
	TCDN	0.1609	0.2259	0.2522	0.2632^{2}	0.2690^{1}	ICDN	0.1779	0.2253	0.2445	0.2450^{2}	0.2489^{1}
	TCDR	0.1692	0.2446	0.2645^{2}	0.2766^{1}	0.2766^{1}	ICDR	0.1814	0.2475	0.2640^{1}	0.2635^{2}	0.2632
	TCPN	0.1680	0.2339	0.2584	0.2659^{2}	0.2711^{1}	ICPN	0.2002	0.2678	0.2718^{2}	0.2722^{1}	0.2638
	TCPR	0.1819	0.2527	0.2715	0.2803^{2}	0.2828^{1}	ICPR	0.2033	0.2763	0.2842^{1}	0.2836^{2}	0.2752
	AVG	0.1949	0.2557	0.2758	$\mathbf{0.2809}^{1}$	0.2783^{2}	UPB	0.4744	$\mathbf{0.5564}^{1}$	0.5409^{2}	0.4990	0.4518
	PR	0.1884	0.2352	0.2521	$\mathbf{0.2599}^{1}$	0.2542^{2}	PR	0.1884	0.2352	0.2521	$\mathbf{0.2599}^{1}$	0.2542^{2}
PARA	THDN	0.0463	0.1424	0.2239	0.2588^{2}	0.2832^{1}	IHDN	0.0606	0.2042	0.2631	0.2828^{2}	0.2871^{1}
	THDR	0.0449	0.1554	0.2325	0.2668^{2}	0.2795^{1}	IHDR	0.0505	0.1912	0.2474	0.2762^{2}	0.2951^{1}
	THPN	0.0449	0.1556	0.2295	0.2644^{2}	0.2833^{1}	**IHPN**	0.0505	0.1968	0.2532	0.2942^{2}	$\mathbf{0.3184}^{1}$
	THPR	0.0449	0.1565	0.2309	0.2648^{2}	0.2802^{1}	IHPR	0.0505	0.1953	0.2517	0.2906^{2}	0.3116^{1}
	TCDN	0.0464	0.1484	0.2209	0.2510^{2}	0.2587^{1}	ICDN	0.0517	0.1750	0.2263	0.2559^{2}	0.2673^{1}
	TCDR	0.0464	0.1466	0.2285	0.2637^{2}	0.2719^{1}	ICDR	0.0516	0.1805	0.2314	0.2581^{2}	0.2762^{1}
	TCPN	0.0464	0.1467	0.2210	0.2508^{2}	0.2591^{1}	ICPN	0.0517	0.1842	0.2450	0.2828^{2}	0.2902^{1}
	TCPR	0.0464	0.1448	0.2275	0.2635^{2}	0.2766^{1}	ICPR	0.0516	0.1840	0.2444	0.2786^{2}	0.3005^{1}
	AVG	0.0491	0.1692	0.2361	0.2689^{2}	$\mathbf{0.2837}^{1}$	UPB	0.0879	0.2685	0.3643	$\mathbf{0.4080}^{1}$	0.4038^{2}
	PR	0.0349	0.1581	0.2244	0.2564^{2}	$\mathbf{0.2639}^{1}$	PR	0.0349	0.1581	0.2244	0.2564^{2}	$\mathbf{0.2639}^{1}$

Table 6. ROUGE-1_F performance of three unit types (UT column), four summarization factors (Method column), five compression rates (ranking from 0.1 to 0.5). When the original text is used directly as the summary (compression rate = 1.0), the ROUGE-1 performance is 0.2264.

highest-weight priority, and the redundancy removal (i.e. IHP*) are effective to improve the ROUGE-1 performance of summarization. The performance of all methods is higher than the baseline method (PR) at the whole range of compression rates. However, focused on ROUGE-1 performance, on average, the performance rank for unit types is CTEDU > TEDU + COMP > PARA (i.e. 0.2571, 0.2424, and 0.2014, respectively). For the upper-bound (UPB) and the average (AVG) performances, the same performance trend is obtained, i.e. CTEDU > TEDU + COMP > PARA. As shown in Table 7, the ROUGE-2 results are similar to the ROUGE-1 as the iterative weighting, the highest-weight priority and the redundancy removal (i.e. IHP*) obtains high performance for all types of units. At the compression rate of 0.1, R-2 values of TCDN, TCDR, TCPN, and TCPR are lower than our baseline. While the weight recalculation is helpful in cases of TEDU + COMP, it is not so effective for CTEDU and PARA. The IHPR achieves the highest ROUGE-2 of 0.1250 (compres-

UT	Method	Compression Rate					Method	Compression Rate				
		0.1	0.2	0.3	0.4	0.5		0.1	0.2	0.3	0.4	0.5
TEDU	THDN	0.0865	0.0896	0.1018	0.1066^{2}	0.1135^{1}	IHDN	0.0998	0.1016	0.1059	0.1068^{2}	0.1088^{1}
+COMP	THDR	0.0803	0.0911	0.1012	0.1069^{2}	0.1166^{1}	IHDR	0.1030	0.1111	0.1159^{1}	0.1156^{2}	0.1143
	THPN	0.0866	0.0915	0.1015	0.1081^{1}	0.1159^{1}	IHPN	0.1112	0.1190	0.1210	0.1233^{1}	0.1223^{2}
	THPR	0.0767	0.0917	0.1017	0.1083^{2}	0.1181^{1}	**IHPR**	0.1145	0.1243	**0.1250**^{1}	0.1249^{2}	0.1249^{2}
	TCDN	0.0520	0.0771	0.0993	0.1069^{2}	0.1156^{1}	ICDN	0.0582	0.0868	0.0892	0.0975^{2}	0.1022^{1}
	TCDR	0.0543	0.0783	0.0957	0.1098^{2}	0.1184^{1}	ICDR	0.0699	0.0907	0.1026	0.1036^{2}	0.1061^{1}
	TCPN	0.0539	0.0834	0.1003	0.1100^{2}	0.1199^{1}	ICPN	0.0816	0.0912	0.1043	0.1067^{1}	0.1049^{2}
	TCPR	0.0550	0.0826	0.0998	0.1114^{2}	0.1191^{1}	ICPR	0.0880	0.1035	0.1082^{1}	0.1054	0.1064^{2}
	AVG	0.0795	0.0946	0.1046	0.1095^{2}	**0.1142**^{1}	UPB	0.2773	**0.3369**^{1}	0.3270^{2}	0.3173	0.3007
	PR	0.0666	0.0779	0.0910	0.0972^{2}	**0.0989**^{1}	PR	0.0666	0.0779	0.0910	0.0972^{2}	**0.0989**^{1}
CTEDU	THDN	0.0712	0.0854	0.0939	0.0943^{2}	0.0952^{1}	IHDN	0.0884	0.1041^{1}	0.0969	0.0976^{2}	0.0973
	THDR	0.0666	0.0849	0.0945^{2}	0.0961^{1}	0.0936	IHDR	0.0853	0.0974	0.1003^{1}	0.0991	0.0993^{2}
	THPN	0.0703	0.0810	0.0937	0.0957^{2}	0.0971^{1}	**IHPN**	0.0838	0.1009	0.1092^{2}	**0.1095**^{1}	0.1069
	THPR	0.0601	0.0848	0.0940	0.0977^{1}	0.0958^{2}	IHPR	0.0774	0.0974	0.1069	0.1088^{1}	0.1075^{2}
	TCDN	0.0528	0.0761	0.0879	0.0916^{2}	0.0919^{1}	ICDN	0.0660	0.0728	0.0851	0.0890^{2}	0.0925^{1}
	TCDR	0.0551	0.0808	0.0879	0.0945^{1}	0.0932^{2}	ICDR	0.0639	0.0825	0.0868	0.0926^{2}	0.0948^{1}
	TCPN	0.0580	0.0798	0.0914	0.0915^{2}	0.0927^{1}	ICPN	0.0671	0.0938	0.0961	0.0982^{1}	0.0963^{2}
	TCPR	0.0659	0.0827	0.0925	0.0948^{2}	0.0956^{1}	ICPR	0.0661	0.0967	0.0994	0.0999^{1}	0.0985^{2}
	AVG	0.0686	0.0876	0.0948	**0.0969**^{1}	0.0968^{2}	UPB	0.2585	**0.2701**^{1}	0.2608^{2}	0.2462	0.2274
	PR	0.0594	0.0783	0.0803	0.0862^{2}	**0.0864**^{1}	PR	0.0594	0.0783	0.0803	0.0862^{2}	**0.0864**^{1}
PARA	THDN	0.0203	0.0428	0.0681	0.0792^{2}	0.0904^{1}	IHDN	0.0319	0.0761	0.0944^{2}	0.0974^{1}	0.0974^{1}
	THDR	0.0143	0.0497	0.0712	0.0869^{2}	0.0920^{1}	IHDR	0.0184	0.0681	0.0836	0.0933^{2}	0.1033^{1}
	THPN	0.0143	0.0497	0.0703	0.0843^{2}	0.0920^{1}	**IHPN**	0.0184	0.0691	0.0837	0.0992^{2}	**0.1124**^{1}
	THPR	0.0143	0.0498	0.0706	0.0854^{2}	0.0917^{1}	IHPR	0.0184	0.0692	0.0835	0.0971^{2}	0.1067^{1}
	TCDN	0.0145	0.0415	0.0642	0.0840^{2}	0.0867^{1}	ICDN	0.0186	0.0538	0.0684	0.0824^{2}	0.0884^{1}
	TCDR	0.0145	0.0394	0.0737	0.0863^{2}	0.0905^{1}	ICDR	0.0185	0.0560	0.0705	0.0828^{2}	0.0966^{1}
	TCPN	0.0145	0.0397	0.0639	0.0835^{2}	0.0865^{1}	ICPN	0.0186	0.0557	0.0828	0.0964^{2}	0.1000^{1}
	TAPR	0.0145	0.0383	0.0727	0.0855^{2}	0.0918^{1}	ICPR	0.0185	0.0557	0.0821	0.0951^{2}	0.1047^{1}
	AVG	0.0177	0.0534	0.0752	0.0887^{2}	**0.0957**^{1}	UPB	0.0490	0.1127	0.1486	**0.1621**^{1}	0.1574^{2}
	PR	0.0054	0.0421	0.0627	0.0823^{2}	**0.0868**^{1}	PR	0.0054	0.0421	0.0627	0.0823^{2}	**0.0868**^{1}

Table 7. ROUGE-2_F performance of three unit types (UT column), four summarization factors (Method column), five compression rates (ranking from 0.1 to 0.5). When the original text is used directly as the summary (compression rate = 1.0), the ROUGE-2 performance is 0.0978.

sion rate = 0.3) for the TEDU + COMP unit type while the IHPN obtains the best ROUGE-2 of 0.1095 (compression rate = 0.4) and 0.1124 (compression rate = 0.5) for the unit type of CTEDU and PARA, respectively. Our baseline (PR) got the highest performance at the compression rate of 0.5 (for the unit type of CTEDU and PARA, i.e. 0.0864 and 0.0868, respectively). The average performance ranked by unit types is TEDU + COMP (0.1005) > CTEDU (0.0889) > PARA (0.0661). For the upper-bound (UPB) and the average (AVG) performances, the same performance trend is obtained, i.e. TEDU + COMP > CTEDU > PARA. As shown in Table 8, the R-SU4 results are similar to the R-2 and R-1 as the iterative weighting, the highest-weight priority and the redundancy removal (i.e. IHP*) obtain high performance for all types of units. Similar to R-2 performance, the weight recalculation is helpful for TEDU + COMP but not for CTEDU and PARA. The IHPR achieves the highest R-SU4 of 0.1310 (compression rate = 0.5) for the TEDU + COMP unit

UT	Method	Compression Rate					Method	Compression Rate				
		0.1	0.2	0.3	0.4	0.5		0.1	0.2	0.3	0.4	0.5
TEDU	THDN	0.0885	0.0924	0.1062	0.1115^2	0.1189^1	IHDN	0.1031	0.1065	0.1112	0.1120^2	0.1140^1
+COMP	THDR	0.0824	0.0938	0.1049	0.1116^2	0.1221^1	IHDR	0.1069	0.1170	0.1221^1	0.1214^2	0.1199
	THPN	0.0889	0.0941	0.1054	0.1127^2	0.1218^1	IHPN	0.1159	0.1248	0.1266	0.1292^1	0.1284^2
	THPR	0.0798	0.0942	0.1054	0.1131^2	0.1238^1	**IHPR**	0.1196	0.1297	0.1305	0.1308^2	**0.1310**^{1}
	TCDN	0.0547	0.0815	0.1037	0.1120^2	0.1213^1	ICDN	0.0606	0.0903	0.0923	0.1016^2	0.1069^1
	TCDR	0.0581	0.0831	0.1004	0.1154^2	0.1243^1	ICDR	0.0743	0.0935	0.1069	0.1084^2	0.1109^1
	TCPN	0.0574	0.0880	0.1048	0.1156^2	0.1261^1	ICPN	0.0863	0.0949	0.1088	0.1120^1	0.1099^2
	TCPR	0.0589	0.0871	0.1050	0.1172^2	0.1250^1	ICPR	0.0927	0.1074	0.1134^1	0.1102	0.1114^2
	AVG	0.0830	0.0986	0.1092	0.1147^2	**0.1197**^{1}	UPB	0.4152	**0.4171**^{1}	0.4169^2	0.4066	0.3743
	PR	0.0683	0.0795	0.0943	0.0993^2	**0.1021**^{1}	PR	0.0683	0.0795	0.0943	0.0993^2	**0.1021**^{1}
CTEDU	THDN	0.0797	0.0950	0.1069	0.1086^2	0.1099^1	IHDN	0.0993	0.1194^1	0.1104	0.1124^2	0.1115
	THDR	0.0733	0.0954	0.1063	0.1091^2	0.1079^1	IHDR	0.0938	0.1135	0.1166^1	0.1161^2	0.1154
	THPN	0.0767	0.0918	0.1081	0.1102^2	0.1121^1	**IHPN**	0.0930	0.1166	0.1285^2	**0.1290**^{1}	0.1238
	THPR	0.0684	0.0954	0.1062	0.1124^1	0.1103^2	IHPR	0.0877	0.1153	0.1267^2	0.1274^1	0.1248
	TCDN	0.0571	0.0825	0.0964	0.1038^2	0.1045^1	ICDN	0.0710	0.0800	0.0940	0.0992^2	0.1031^1
	TCDR	0.0599	0.0886	0.1005	0.1093^1	0.1082^2	ICDR	0.0689	0.0905	0.0980	0.1036^2	0.1072^1
	TCPN	0.0627	0.0884	0.0996	0.1043^2	0.1065^1	ICPN	0.0751	0.1049	0.1074	0.1110^1	0.1104^2
	TCPR	0.0706	0.0916	0.1063	0.1093^2	0.1115^1	ICPR	0.0745	0.1074	0.1137^2	0.1152^1	0.1136
	AVG	0.0757	0.0985	0.1079^2	**0.1113**^{1}	**0.1113**^{1}	UPB	0.3013^2	**0.3131**^{1}	0.2977	0.2745	0.2465
	PR	0.0636	0.0851	0.0890^2	**0.0983**^{1}	0.0975^2	PR	0.0636	0.0851	0.0890^2	**0.0983**^{1}	0.0975^2
PARA	THDN	0.0227	0.0579	0.0916	0.1062^2	0.1234^1	IHDN	0.0402	0.0983	0.1221^2	0.1284^1	0.1284^1
	THDR	0.0185	0.0698	0.1010	0.1184^2	0.1228^1	IHDR	0.0238	0.0862	0.1108	0.1245^2	0.1363^1
	THPN	0.0185	0.0698	0.0993	0.1142^2	0.1244^1	**IHPN**	0.0238	0.0916	0.1150	0.1384^2	**0.1531**^{1}
	THPR	0.0185	0.0704	0.0996	0.1169^2	0.1233^1	IHPR	0.0238	0.0916	0.1142	0.1357^2	0.1478^1
	TCDN	0.0186	0.0599	0.0902	0.1082^2	0.1091^1	ICDN	0.0240	0.0700	0.0920	0.1058^2	0.1115^1
	TCDR	0.0186	0.0576	0.0989	0.1122^2	0.1176^1	ICDR	0.0239	0.0727	0.0939	0.1078^2	0.1210^1
	TCPN	0.0186	0.0591	0.0904	0.1081^2	0.1100^1	ICPN	0.0240	0.0768	0.1089	0.1269^2	0.1300^1
	TCPR	0.0186	0.0566	0.0983	0.1123^2	0.1207^1	ICPR	0.0239	0.0768	0.1085	0.1274^2	0.1423^1
	AVG	0.0225	0.0728	0.1022	0.1182^2	**0.1264**^{1}	UPB	0.0669	0.1476	0.1902^2	**0.2003**^{1}	0.1855
	PR	0.0074	0.0653	0.0908	0.1069^2	**0.1145**^{1}	PR	0.0074	0.0653	0.0908	0.1069^2	**0.1145**^{1}

Table 8. ROUGE-SU4$_F$ performance of three unit types (UT column), four summarization factors (Method column), five compression rates (ranking from 0.1 to 0.5). When the original text is used directly as the summary (compression rate = 1.0), the ROUGE-SU4 performance is 0.1083.

type, compared to IHPNs R-SU4 of 0.1284. For CTEDU and PARA, the IHPN obtains the best R-SU4 of 0.1290 (compression rate = 0.4) and 0.1531 (compression rate = 0.5), compared to IHPRs R-SU4 of 0.1248 and 0.1478, respectively. Moreover, R-SU4 values of PR are lower than our proposed method at the whole compression rate except TCDN, TCDR, and TCPN (at the compression of 0.1 and 0.2). The average performance ranked by unit types is TEDU + COMP (0.1051) > CTEDU (0.1009) > PARA (0.0884). The same trend (TEDU+COMP > CTEDU > PARA) is obtained for the upper bound (UBP) and the average (AVG) performances.

To conclude, for the unit type of TEDU+COMP, the method of iterative weighting, the highest-weight priority, redundancy removal, and weight recalculation (i.e. IHPR), achieves the highest performance in all ROUGE measures. For the unit type of CTEDU, the method of the iterative weighting, the highest-weight priority

Rank No.	Method	ROUGE-1			ROUGE-2			ROUGE-SU4			AVG
		P	R	F	P	R	F	P	R	F	F
1	IHPR	0.2119	0.4222	**0.2644**	0.0744	0.1882	0.0991	0.0882	0.2010	**0.1158**	**0.1598**
2	IHPN	0.2087	0.4274	0.2624	0.0739	0.1938	**0.0993**	0.0876	0.2066	**0.1158**	0.1592
3	IHDR	0.2005	0.4252	0.2547	0.0684	0.1929	0.0939	0.0805	0.2017	0.1083	0.1523
4	IHDN	0.1913	0.4141	0.2440	0.0682	0.1960	0.0936	0.0806	0.2036	0.1078	0.1485
5	ICPR	0.1914	0.3908	0.2391	0.0658	0.1769	0.0886	0.0776	0.1864	0.1026	0.1434
6	ICPN	0.1850	0.3828	0.2318	0.0635	0.1755	0.0862	0.0746	0.1828	0.0992	0.1391
7	THPR	0.1900	0.3713	0.2319	0.0647	0.1556	0.0827	0.0755	0.1649	0.0958	0.1368
8	THDR	0.1895	0.3722	0.2316	0.0648	0.1572	0.0831	0.0754	0.1657	0.0958	0.1368
9	THPN	0.1881	0.3729	0.2303	0.0649	0.1599	0.0835	0.0758	0.1689	0.0965	0.1368
10	THDN	0.1835	0.3699	0.2270	0.0635	0.1600	0.0826	0.0736	0.1670	0.0946	0.1347
11	TCPR	0.1886	0.3676	0.2295	0.0627	0.1549	0.0801	0.0732	0.1621	0.0926	0.1341
12	ICDR	0.1783	0.3651	0.2225	0.0601	0.1639	0.0812	0.0697	0.1679	0.0921	0.1319
13	TCDR	0.1860	0.3601	0.2258	0.0613	0.1504	0.0782	0.0714	0.1569	0.0902	0.1314
14	TCPN	0.1815	0.3517	0.2200	0.0611	0.1511	0.0779	0.0708	0.1560	0.0893	0.1291
15	TCDN	0.1791	0.3416	0.2163	0.0600	0.1450	0.0761	0.0692	0.1489	0.0869	0.1264
16	ICDN	0.1676	0.3376	0.2073	0.0575	0.1526	0.0767	0.0665	0.1559	0.0868	0.1236

Table 9. Ranking methods by the average of ROUGE-1, ROUGE-2, and ROUGE-SU4

and the redundancy removal (i.e. IHPN) is superior to other methods for R-2 and R-SU4, whereas IHPR is best for R-1. For the simple unit type of PARA, IHPN is the optimal for all ROUGE performances. On average, the performance rank of unit types is TEDU + COMP > CTEDU > PARA. For the upper-bound (UPB) and the average (AVG) performances, the same performance trend is obtained, i.e. TEDU + COMP > CTEDU > PARA. The unit type of TEDU + COMP obtains better performance than the others since it includes short keywords and allows us to flexibly select units for summarization. However, from the results in Tables 6–8, we still cannot figure out which method is the best. To emphasize and contrast their performances, the average ROUGE-based (R-1, R-2, and R-SU4) precision, recall, and F-score over all three unit types and all compression rates (i.e. 0.1 to 0.5) for each method are calculated. Table 9 shows the results in the order of the average F-score from R-1, R-2, and R-SU4. Here, the compression rates used for evaluation are from 0.1 to 0.5. In this table, we found that IHPR achieves the best performance for R-1 but is slightly inferior to IHPN for R-2. Both IHPR and IHPN achieve the highest F-score for R-SU4. Moreover, the average F-score of three ROUGE performances for IHPR and IHPN are comparable. The result comes to the same conclusion as done in Tables 6–8, that the combination of the iterative weighting, highest-weight priority and redundancy removal make us gain the highest performance. However, we cannot conclude performance comparison among methods by using only average performance. In the conclusion, in the next two sections, we show the results of factor-specific comparison with t-test in Section 6.2 and top-8 method comparison with t-test in Section 6.3.

Factor	Method (M1, M2)	AVG(ROUGE$_F$) (M1, M2)	t-test	W/L/T	Factor	Method (M1, M2)	AVG(ROUGE$_F$) (M1, M2)	t-test	W/L/T
I/T	ICPN, TCPN	0.1391, 0.1291	0.00	W	C/H	ICPN, IHPN	0.1391, 0.1592	0.00	L
	ICDN, TCDN	0.1236, 0.1264	0.23	T		ICDN, IHDN	0.1236, 0.1485	0.00	L
	IHPN, THPN	0.1592, 0.1368	0.00	W		ICPR, IHPR	0.1434, 0.1598	0.00	L
	ICPR, TCPR	0.1434, 0.1341	0.00	W		ICDR, IHDR	0.1319, 0.1523	0.00	L
	ICDR, TCDR	0.1319, 0.1314	0.80	T		TCPN, THPN	0.1291, 0.1368	0.00	L
	IHPR, THPR	0.1598, 0.1368	0.00	W		TCDN, THDN	0.1264, 0.1347	0.00	L
	IHDR, THDR	0.1523, 0.1368	0.00	W		TCPR, THPR	0.1341, 0.1368	0.10	T
	IHDN, THDN	0.1485, 0.1347	0.00	W		TCDR, THDR	0.1314, 0.1368	0.00	L
Overall				**6/0/2**	**Overall**				**0/7/1**
P/D	ICPN, ICDN	0.1391, 0.1236	0.00	W	R/N	ICPR, ICPN	0.1434, 0.1391	0.00	W
	IHPN, IHDN	0.1592, 0.1485	0.00	W		ICDR, ICDN	0.1319, 0.1236	0.00	W
	ICPR, ICDR	0.1434, 0.1319	0.00	W		IHPR, IHPN	0.1598, 0.1592	0.39	T
	IHPR, IHDR	0.1598, 0.1523	0.00	W		IHDR, IHDN	0.1523, 0.1485	0.04	W
	TCPN, TCDN	0.1291, 0.1264	0.00	W		TCPR, TCPN	0.1341, 0.1291	0.00	W
	THPN, THDN	0.1368, 0.1347	0.20	T		TCDR, TCDN	0.1314, 0.1264	0.00	W
	TCPR, TCDR	0.1341, 0.1314	0.00	W		THPR, THPN	0.1368, 0.1368	0.94	T
	THPR, THDR	0.1368, 0.1368	0.99	T		THDR, THDN	0.1368, 0.1347	0.22	T
Overall				**6/0/2**	**Overall**				**5/0/3**

Table 10. Performance comparison among three factors, using a two-tailed t-test with 5 % significance

6.2 Effect of Four Factors on Summarization Performance

This section explores the effect of four factors on the performance. For each factor, two alternatives, i.e. (I vs. T), (P vs. D), (C vs. H), and (R vs. N), are investigated by varying other factors and then comparing their results. Table 10 demonstrates the performance comparison using a two-tailed t-test with 5 per cent significance. Each comparison contains 2 250 cases (3 units $\times$ 50 datasets $\times$ 5 compression rates $\times$ 3 ROUGEs). The averages of R-1$_F$, R-2$_F$, and R-SU4$_F$ are used to find the number of wins, losses, and ties (W/L/T).

Table 10 shows that the iterative weighting outperforms the simple TF-IDF with 6 wins and 2 ties. This win comes from the fact that the iterative weighting revises weights of words, units, and documents after each unit selection while the simple TF does nothing. For the importance-based selection, the selection based on the highest-weight priority is powerful enough without consideration of centroid preference. The probable reason may be triggered by the fact that the centroid may not be a good candidate for summary, but can be instead of selecting decentralized information for summary. For redundancy removal, we can obtain six wins and two ties when we consider that redundant parts should not be included into a summary. Especially when we summarize multiple documents that have a high potential of content overlap, the redundancy removal process is highly recommended. For the last factor, the post-selection weight recalculation obtains five wins and three ties. This result implies that we should reduce the weight of a node that is similar to the previously selected nodes in order not to be included in the summary.

Method	ROUGE	IHPR	IHPN	IHDR	IHDN	ICPR	ICPN	THPR	THDR	Overall	Score	Rank
IHPR	R-1	–	1/0/4	3/0/2	3/0/2	5/0/0	5/0/0	5/0/0	5/0/0	27/0/8		
	R-2	–	0/0/5	3/0/2	2/0/3	5/0/0	5/0/0	5/0/0	5/0/0	25/0/10		
	R-SU4	–	0/0/5	4/0/1	2/0/3	5/0/0	5/0/0	5/0/0	5/0/0	26/0/9		
	Overall	–	**1/0/14**	**10/0/5**	**7/0/8**	**15/0/0**	**15/0/0**	**15/0/0**	**15/0/0**	**78/0/27**	**261**	**1**
IHPN	R-1	0/1/4	–	2/0/3	3/0/2	5/0/0	5/0/0	5/0/0	5/0/0	25/1/9		
	R-2	0/0/5	–	3/0/2	2/0/3	5/0/0	5/0/0	5/0/0	5/0/0	25/0/10		
	R-SU4	0/0/5	–	4/0/1	2/0/3	5/0/0	5/0/0	5/0/0	5/0/0	26/0/9		
	Overall	**0/1/14**	–	**9/0/6**	**7/0/8**	**15/0/0**	**15/0/0**	**15/0/0**	**15/0/0**	**76/1/28**	**256**	**2**
IHDR	R-1	0/3/2	0/2/3	–	3/0/2	5/0/0	5/0/0	3/0/2	3/0/2	19/5/11		
	R-2	0/3/2	0/3/2	–	1/0/4	1/0/4	3/0/2	3/0/2	3/0/2	11/6/18		
	R-SU4	0/4/1	0/4/1	–	1/0/4	1/0/4	3/0/2	3/0/2	3/0/2	11/8/16		
	Overall	**0/10/5**	**0/9/6**	–	**5/0/10**	**7/0/8**	**11/0/4**	**9/0/6**	**9/0/6**	**41/19/45**	**168**	**3**
IHDN	R-1	0/3/2	0/3/2	0/3/2	–	1/0/4	3/0/2	2/1/2	2/1/2	8/11/16		
	R-2	0/2/3	0/2/3	0/1/4	–	1/0/4	2/0/3	3/0/2	3/0/2	9/5/21		
	R-SU4	0/2/3	0/2/3	0/1/4	–	1/0/4	2/0/3	2/0/3	2/0/3	7/5/23		
	Overall	**0/7/8**	**0/7/8**	**0/5/10**	–	**3/0/12**	**7/0/8**	**7/1/7**	**7/1/7**	**24/21/60**	**132**	**4**
ICPR	R-1	0/5/0	0/5/0	0/5/0	0/1/4	–	4/0/1	2/0/3	2/0/3	8/16/11		
	R-2	0/5/0	0/5/0	0/1/4	0/1/4	–	1/0/4	0/0/5	0/0/5	1/12/22		
	R-SU4	0/5/0	0/5/0	0/1/4	0/1/4	–	3/0/2	0/0/5	0/0/5	3/12/20		
	Overall	**0/15/0**	**0/15/0**	**0/7/8**	**0/3/12**	–	**8/0/7**	**2/0/13**	**2/0/13**	**12/40/53**	**89**	**5**
ICPN	R-1	0/5/0	0/5/0	0/5/0	0/3/2	0/4/1	–	1/1/3	1/1/3	2/24/9		
	R-2	0/5/0	0/5/0	0/3/2	0/2/3	0/1/4	–	0/0/5	0/0/5	0/16/19		
	R-SU4	0/5/0	0/5/0	0/3/2	0/2/3	0/3/2	–	0/0/5	0/0/5	0/18/17		
	Overall	**0/15/0**	**0/15/0**	**0/11/4**	**0/7/8**	**0/8/7**	–	**1/1/13**	**1/1/13**	**2/58/45**	**51**	**8**
THPR	R-1	0/5/0	0/5/0	0/3/2	1/2/2	0/2/3	1/1/3	–	1/0/4	3/18/14		
	R-2	0/5/0	0/5/0	0/3/2	0/3/2	0/0/5	0/0/5	–	1/0/4	1/16/18		
	R-SU4	0/5/0	0/5/0	0/3/2	0/2/3	0/0/5	0/0/5	–	1/0/4	1/15/19		
	Overall	**0/15/0**	**0/15/0**	**0/9/6**	**1/7/4**	**0/2/13**	**1/1/13**	–	**3/0/12**	**5/49/51**	**66**	**6**
THDR	R-1	0/5/0	0/5/0	0/3/2	1/2/2	0/2/3	1/1/3	0/1/4	–	2/19/14		
	R-2	0/5/0	0/5/0	0/3/2	0/3/2	0/0/5	0/0/5	0/1/4	–	0/17/18		
	R-SU4	0/5/0	0/5/0	0/3/2	0/2/3	0/0/5	0/0/5	0/1/4	–	0/16/19		
	Overall	**0/15/0**	**0/15/0**	**0/9/6**	**1/7/7**	**0/2/13**	**1/1/13**	**0/3/12**	–	**2/52/51**	**57**	**7**

Table 11. Performance comparison among 8 methods with a two-tailed t-test at 5 % significance

6.3 Overall Comparison of Top-8 Methods

In this experiment, we rank the average ROUGE_F of sixteen methods and perform the two-tailed paired t-test at 5 per cent significance level on the top-8 methods. Table 11 shows the results when each of the top-8 methods is investigated by considering 105 comparisons (5 compression rates $\times$ 3 ROUGE values $\times$ 7 methods). Following a standard schema, we set the scores of a win, a loss, and a tie to 3, 0, and 1, respectively. For each win, loss or tie, we derive the result from 150 cases (50 datasets $\times$ 3 unit types). The row header and the column header

indicate two methods to be compared. The value in each cell contains numbers of wins, losses, and ties. In this table, we noticed that IHPR and IHPN achieve best and the second best performances in terms of R-1, R-2, and R-SU4 at the compression rates of 0.1 to 0.5. IHDR performs well in terms of R-1 while IHDN gains high performance of R-2. ICPR wins ICPN, THPR, and THDR in terms of R-1, R-2, and R-SU4. For the worst methods (out of 8), ICPN and THDR obtain only two wins. ICPN is located at the sixth rank in Table 9 while it is placed in the eighth rank in Table 11. Even this ICPN obtains better average performance than THPR and THDR (see Table 9) but in details, it wins THPR and THDR only for R-1 with a large gap. Therefore, when we consider win/loss/tie (see Table 11), ICPN loses THPR and THDR. It is possible that both the extension of the highest-weight priority with centroid preference and the non-recalculation decrease the summarization performance. In conclusion, the rank order among the eight methods is IHPR > IHPN > IHDR > IHDN > ICPR > THPR > THDR > ICPN. The combination of the iterative weighting, the highest-weight priority, and the redundancy removal (i.e. 'IHP*'), improves the summarization performance.

Group of Dataset	ROUGE-1			ROUGE-2			ROUGE-SU4		
	P	R	F	P	R	F	P	R	F
2–3 docs/dataset	0.1976	0.3555	0.2386	0.0642	0.1504	0.0826	0.0758	0.1543	0.0957
4–6 docs/dataset	0.1913	0.5350	0.2649	0.0953	0.3083	0.1367	0.1048	0.3388	0.1510
7–15 docs/dataset	0.0630	0.3974	0.1034	0.0099	0.1131	0.0167	0.0149	0.1260	0.0253
All datasets	0.1888	0.3795	0.2336	0.0647	0.1671	0.0852	0.0756	0.1748	0.0981

Table 12. Average ROUGE-based (ROUGE-1, ROUGE-2, and ROUGE-SU4) precision (P), recall (R), and F-score (F) summarized over all three unit types and all compression rates (i.e. 0.1 to 0.5) for each method, classified by dataset size. Here, datasets are grouped into small-sized (2–3 documents per set), middle-sized (4–6 documents per set), and large-sized (7–15 documents per set) datasets.

6.4 Effect of Number of Documents on Summarization

This experiment aims to investigate summarization performance by taking the dataset size into consideration. Here, three groups of datasets are considered, (1) the group of small-sized datasets (2–3 documents per dataset), (2) the group of middle-sized datasets (4–6 documents per dataset), and (3) the group of large-sized datasets (7–15 documents per dataset). Table 12 displays R-1, R-2, and R-SU4 when each cell value is calculated by averaging performances over three types of units and five compression rates. It seems that summarization of a middle-sized dataset (4–6 documents per set) can be done efficiently with the highest performance of F-score-based R-1, R-2, and R-SU4. When we focus on the overall performance, summarization of small-sized datasets obtains lower performance than that of middle-sized datasets.

Two potential reasons are (1) a small-sized dataset has quite uniform contents with a lot of overlaps between original documents and then the summary is not efficiently generated, and (2) a small-sized dataset has a smaller number of paragraphs and then it is hard to obtain good performance for paragraph-based summarization. Summarization of a larger dataset seems more difficult than that of a smaller one since a larger dataset contains various contents and a lot of candidates to be selected as a part of the summary, resulting in low performance.

7 CONCLUSION AND FUTURE WORK

This paper introduced a definition of Thai Elementary Discourse Unit (TEDU) and then presented a three-stage method of Thai multi-document summarization, i.e. unit segmentation, unit graph formulation, and unit selection for summarization.

We investigated three different units: TEDU + COMP, CTEDU, and paragraph.

These units are represented as nodes and their relationships are formed as links among units with weights.

In this work, four factors considered are the node weighting, the importance-based selection, the redundancy removal, and the post-selection weight recalculation.

Using fifty sets of Thai news articles, we showed that TEDU + COMP with the IHPR method (the method of iterative weighting, the highest-weight priority without centroid preference, redundancy removal, and post-selection weight recalculation) yielded the best performance in all ROUGEs evaluations. For CTEDU, IHPR performed the best for R-1 whereas IHPN (comparable to IHPR) gains high R-2 and R-SU4. For the simple unit type of PARA, IHPN was the optimal for all ROUGE performances. On average, the performance rank of unit types is TEDU + COMP > CTEDU > PARA. The unit type of TEDU + COMP obtains better performance than the others since it includes short keywords and allows us to flexibly select units for summarization.

In future works, we will analyze the relation between TEDUs in order to form a more suitable set of combined TEDU with consideration of semantic. Moreover, we plan to investigate more semantic-based unit weighting and selection, including consideration of semantics of conjunctions or discourse markers in order to improve our 'combined EDU (CTEDU)'. Finally, it would be useful to explore our approach on a larger dataset.

Acknowledgement

This work was partially supported by the National Research University Project of Thailand OCE of Higher Education Commission, the National Electronics and Computer Technology Center (NECTEC) under Project Number NT-B-22-KE-38-54-01, and a Research Grant sponsored by the Bangchak Petroleum Public Company Limited (BCP), Thailand.

REFERENCES

[1] Barzilay, R.–McKeown, K. R.–Elhadad, M.: Information Fusion in the Context of Multi-Document Summarization. Proceedings of the 37th Annual Meeting of the ACL, 1999, pp. 550–557.

[2] Carbonell, J.—Goldstein, J.: The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. Research and Development in Information Retrieval, 1998, pp. 335–336.

[3] Carlson, L.—Marcu, D.—Okurowski, M. E.: Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. Current Directions in Discourse and Dialogue, 2003, pp. 85–112.

[4] Charoensuk, J.—Sukvaree, T.—Kawtrakul, A.: Elementary Discourse Unit Segmentation for Thai Using Discourse Cues and Syntactic Information. Proceedings of the 2nd International Conference on Corpus Linguistics (NCSEC2005), 2005.

[5] Chongsuntornsri, A.–Sornil, O.: An Automatic Thai Text Summarization Using Topic Sensitive Pagerank. Proceedings of International Symposium on Communications and Information Technologies (ISCIT '06), 2006, pp. 547–552.

[6] Frakes, W. B.—Baeza-Yates, R. A.: Information Retrieval. Data Structures & Algorithms. Prentice-Hall, Editors 1992.

[7] Haveliwala, T.: Topic-Sensitive Pagerank. Proceedings of the 11th International Conference on World Wide Web, 2000, pp. 517–526.

[8] Jaruskulchai, C.–Kruengkrai, C.: A Practical Text Summarizer by Paragraph Extraction for Thai. Proceedings of the Sixth International Workshop on Information Retrieval with Asian Languages (AsianIR '03), Vol. 11, 2003, pp. 9–16.

[9] Ketui, N.—Theeramunkong, T.: Inclusion-Based and Exclusion-Based Approaches in Graph-Based Multiple News Summarization. Knowledge, Information and Creativity Support Systems, LNCS, Vol. 6746, 2010, pp. 91–102.

[10] Ketui, N.—Theeramunkong, T.—Onsuwan, C.: Thai Elementary Discourse Unit Analysis and Syntactic-Based Segmentation. Information. An International Interdisciplinary Journal (Information-Tokyo), Vol. 16, 2013, No. 10, pp. 7423–7436.

[11] Kittiphattanabawon, N.—Theeramunkong, T.—Nantajeewarawat, E.: News Relation Discovery Based on Association Rule Mining with Combining Factors. IEICE Transactions on Information and Systems, Vol. E94-D, 2011, No. 3, pp. 404–415.

[12] Lin, C.-Y.: Rouge: A Package for Automatic Evaluation of Summaries. Proceedings of ACL Workshop on Text Summarization, 2004, pp. 74–81.

[13] Maier, D.: The Complexity of Some Problems on Subsequences and Supersequences. Journal of Association for Computing Machinery, Vol. 25, 1978, No. 2, pp. 322–336.

[14] Mani, I.: Multi-Document Summarization by Graph Search and Matching. Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-97), 1997, pp. 622–628.

[15] Mann, W. C.—Thompson, S. A.: Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. Text, Vol. 8, 1988, No. 3, pp. 243–281.

[16] Meknavin, S.—Charoenpornsawat, P.—Kijsirikul, B.: Feature-Based Thai Word Segmentation. Proceedings of the Natural Language Processing Pacific Rim Symposium (NLPRS '97), 1997.

[17] Mihalcea, R.: Graph-Based Ranking Algorithms for Sentence Extraction, Applied to Text Summarization. Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions (ACLDEMO '04), Association for Computational Linguistics, Stroudsburg, PA, USA, 2004.

[18] Page, L.—Brin, S.—Motwani, R.—Winograd, T.: The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford University, 1998.

[19] Porter, M. F.: An Algorithm for Suffix Stripping. Program, Vol. 14, 1980, No. 30, pp. 130–137.

[20] Radev, D. R.—Jing, H.—Budzikowska, M.: Centroid-Based Summarization of Multiple Documents: Sentence Extraction, Utility-Based Evaluation, and User Studies. Proceedings of NAACL-ANLP 2000 Workshop on Automatic Summarization, 2000, pp. 21–30.

[21] Sinthupoun, S.—Sornil, O.: Thai Rhetorical Structure Analysis. International Journal of Computer Science and Information Security (IJCSIS), Vol. 7, 2010, No. 1, pp. 95–105.

[22] Sornil, O.—Gree-Ut, K.: An Automatic Text Summarization Approach Using Content-Based and Graph-Based Characteristics. Proceedings of IEEE Conference on Cybernetics and Intelligent Systems, 2006, pp. 1–6.

[23] Sukvaree, T.—Charoensuk, J.—Wattanamethanont, M.—Kawtrakul, A.: RST Based Text Summarization with Ontology Driven in Agriculture Domain. Proceedings of Workshop on Supporting Multilinguality in Agricultural Information Access, 2004.

[24] Steinberger, J.—Jezek, K.: Evaluation Measures for Text Summarization. Computing and Informatics, Vol. 28, 2009, pp. 251–275.

[25] Thangthai, A.—Jaruskulchai, C.: Impact Parameter on LSA Performance for Thai Text Summarization. Proceedings of the 43rd Kasetsart University Annual Conference: Veterinary Medicine, Science (Vichakarn '43), 2004, pp. 331–339.

[26] Theeramunkong, T.—Boriboon, M.—Haruechaiyasak, C.—Kittiphattanabawon, N.—Kosawat, K.—Onsuwan, C.—Siriwat, I.—Suwanapong, T.—Tongtep, N.: Thai-Nest: A Framework for Thai Named Entity Tagging Specification and Tools. Proceedings of the 2nd International Conference on Corpus Linguistics (CILC '10), University of A. Coruna, Spain, 2010, pp. 895–908.

[27] Tongtep, N.—Theeramunkong, T.: Multi-Stage Automatic NE and POS Annotation Using Pattern-Based and Statistical-Based Techniques for Thai Corpus Construction. IEICE Transaction on Information and Systems, Vol. E96-D, 2013, No. 10, pp. 2245–2256.

[28] Wan, X.—Xiao, J.: Exploiting Neighborhood Knowledge for Single Document Summarization and Keyphrase Extraction. ACM Transactions on Information Systems (TOIS), Vol. 28, 2010, No. 2, pp. 8:1–8:34.

[29] Wu, H.-C.—Luk, R.-W.-P.—Wong, K.-F.—Kwok, K.-L.: Interpreting TF-IDF Term Weights as Making Relevance Decisions. ACM Transactions on Information

Systems (TOIS), Vol. 26, 2008, No. 3, pp. 13:1–13:37.

[30] Yeh, J.-Y.—Ke, H.-R.—Yang, W.-P.—Meng, I.-H.: Text Summarization Using a Trainable Summarizer and Latent Semantic Analysis. Information Processing and Management, Vol. 41, 2005, No. 1, pp. 75–95.

Nongnuch Ketui is currently a student in Ph.D. program, School of Information, Computer, and Communication Technology, Sirindhorn International Institute of Technology, Thammasat University, Thailand. Her research interests are natural language processing and data mining.

Thanaruk Theeramunkong is Professor at the School of Information, Computer, and Communication Technology, Sirindhorn International Institute of Technology, Thammasat University, Thailand. His research interests include human language technology, data/text mining, artificial intelligence, database technology, and service sciences.