An EDU-Based Approach for Thai Multi-Document Summarization and Its Application

NONGNUCH KETUI, THANARUK THEERAMUNKONG, and
CHUTAMANEE ONSUWAN, Thammasat University, Thailand

Due to lack of a word/phrase/sentence boundary, summarization of Thai multiple documents has several challenges in unit segmentation, unit selection, duplication elimination, and evaluation dataset construction. In this article, we introduce Thai Elementary Discourse Units (TEDUs) and their derivatives, called Combined TEDUs (CTEDUs), and then present our three-stage method of Thai multi-document summarization, that is, unit segmentation, unit-graph formulation, and unit selection and summary generation. To examine performance of our proposed method, a number of experiments are conducted using 50 sets of Thai news articles with their manually constructed reference summaries. Based on measures of ROUGE-1, ROUGE-2, and ROUGE-SU4, the experimental results show that: (1) the TEDU-based summarization outperforms paragraph-based summarization; (2) our proposed graph-based TEDU weighting with importance-based selection achieves the best performance; and (3) unit duplication consideration and weight recalculation help improve summary quality.

Categories and Subject Descriptors: I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Text Analysis*

General Terms: Algorithms, Experimentation, Languages, Performance

Additional Key Words and Phrases: Multi-document summarization, EDU-based approach, Thai text summarization, unit selection

ACM Reference Format:
Ketui, N., Theeramunkong, T., and Onsuwan, C. 2015. An EDU-based approach for Thai multi-document summarization and its application. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 14, 1, Article 4 (January 2015), 26 pages.
DOI: http://dx.doi.org/10.1145/2641567

1. INTRODUCTION

Everyday a gigantic number of news articles are produced in various languages all over the world. Such a situation where too much information is available makes it difficult for us to find needed information and to assemble an accurate and complete record of current events from many related news articles derived from several sources with similar and different facts. To solve this problem, it is necessary to study an efficient and effective method to create a summary from multiple news articles. In an early work, Mani [1997] proposed an approach for multi-document summarization by

This work was partially supported by the National Research University Project of Thailand Office of Higher Education Commission, the National Electronics and Computer Technology Center (NECTEC) under project no. NT-B-22-KE-38-54-01 and a research grant sponsored by the Bangchak Petroleum Public Company Limited (BCP), Thailand.
Authors' addresses: N. Ketui (corresponding author), T. Theeramunkong, School of Information, Computer, and Communication Technology, Sirindhorn International Institute of Technology, Thammasat University, Thailand; C. Onsuwan, Faculty of Liberal Arts, Sirindhorn International Institute of Technology, Thammasat University, Thailand.

DOI: http://dx.doi.org/10.1145/2641567

graph search and matching under similarities and dissimilarities among document pairs. Later, Carbonell and Goldstein [1998] and Goldstein and Carbonell [1998] presented a method of combining query relevance with information novelty as topic-driven summarization with a maximal marginal relevance (MMR) measure. For this purpose, a number of works [Mani and Bloedorn 1999; McKeown and Radev 1999] explored the similarities and differences among related documents. McKeown et al. [1999] and Radev et al. [2000] proposed centroid-based techniques to generate a composite sentence from each cluster by using a number of syntactic-based features. Following techniques in Barzilay et al. [1999] and Cai and Li [2011], summarization is formulated as a clustering problem. Several works [Kuo and Chen 2008; Okazaki et al. 2005] considered a sentence ordering method that helps to extract the important information and then generate a new summary. Generally, much research focuses on extractive summarization [Alguliev et al. 2011; Aliguliyev 2009; Ferreira et al. 2013] and then updates the information for summary generation [Wang and Zhou 2012].

In Thai, there have been very few works on summarization since Thai texts are structurally flexible and complicated, and also because techniques and tools for basic text processing in Thai are still in their infancy stage. As for work on Thai text summarization, Jaruskulchai and Kruengkrai [2003] proposed a paragraph-based summarization that selects the 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 are not taken into account, and its primary target is summarizing a single document. As another work on paragraph-based summarization, Thangthai and Jaruskulchai [2004] studied effects of parameters 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. Later, Suwanno et al. [2005] proposed a compound noun extraction method and news structure (headline) consideration to enhance paragraph-based summarization with a better score ranking method. Moreover, Ketui and Theeramunkong [2010] presented a more sophisticated weighting system with iterative weight calculation. They also proposed two summarization methods, called inclusion-based and exclusion-based selection, to pick up a set of candidate paragraphs from multiple news documents for a summary. However, summarization based on paragraphs is forced to select a whole paragraph for summary, even though only some parts in a paragraph are important. To avoid this restriction, two recent works have proposed methods to select, instead of paragraphs, significant segments for a Thai single-document summarization [Chongsuntornsri and Sornil 2006; Sornil and Gree-ut 2006].

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 a subset of segments is usually not readable. As a solution, Sukvaree et al. [2007] have presented an EDU-based summarization approach that extracts Elementary Discourse Units (EDUs) [Carlson et al. 2003] from a text, then finds their discourse coherence and organizes them into a spanning tree based on the well-known rhetorical structure theory. In this work, EDUs are extracted by Thai discourse markers. The result showed that Thai discourse coherence can refer to different meanings, so-called *word sense ambiguity*. For example, some words can be both a conjunction and a marker cue. However, the graph-based approach is able to specify the maximal number of common and different sentences to control the output. Several works [Carbonell and Goldstein 1998; Erkan and Radev 2004; Mihalcea 2004] have applied graph-based approaches to order the document and produce summaries. This method uses these structures to actually compose abstractive summaries, rather than to extract sentences from the text.

In this work, we introduce Thai Elementary Discourse Units (TEDUs) and common phrases (COMPs) and then present a three-stage method of Thai multi-document summarization, that is, unit segmentation, unit graph formulation, and unit selection and summary generation for summarization. To evaluate our methods, we investigate three different granularities of units: (1) TEDU+COMP; (2) combined TEDU; and (3) paragraph. Our proposed methods can be parameterized by three factors: (1) importance-based selection; (2) redundancy avoidance; and (3) postselection weight recalculation. A number of experiments are conducted using 50 sets of Thai news documents; a summary of performance evaluations is also given. Three measures of ROUGE-1, ROUGE-2, and ROUGE-SU4 are used as performance metrics. Then, we apply a Thai multi-document summarization model to develop automatic Thai news summarization.

The rest of the article is organized as follows. Section 2 defines Thai Elementary Discourse Unit (TEDUs). Our method on Thai multi-document summarization is presented in Section 3. Methods and procedures are shown in Section 4. Experimental results and discussions are described in Section 5. A summarization engine and GUI are given in Section 6. Finally, conclusions and future work are given in Section 7.

2. THAI ELEMENTARY DISCOURSE UNITS (TEDUS)

In the past, there have been a number of works on extracting Elementary Discourse Units (EDUs) from Thai texts, such as extraction from an agriculture corpus [Charoensuk et al. 2005] and a written family law text [Sinthupoun and Sornil 2010]. Considering the definition of Thai EDUs (TEDUs) in these works and the definition of English EDUs in Carlson et al. [2003], we have defined a set of TEDUs to reflect special characteristics in the Thai language in this section. In our framework adapted from Ketui et al. [2012], 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, some word sequences look syntactically similar to a TEDU by containing a verb as their component, but they are not TEDUs. Later called a TEDU-like phrase/word (TEDU-LP), some examples of such sequences are clausal subjects/objects or synthetic nominal compounds. On the other hand, a verbal unit may be a simple verb and a verb with some auxiliary units. Moreover, in Thai texts, it is possible to have concatenated verbs, this is 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 into several units as a 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 and TEDU-LP-2) as follows. In order to detect units of such three fundamental types, namely, TEDUs, COMPs, and TEFU-LPs, a set of patterns and clues are defined for each type, as shown in Figure 1.

(1) *TEDU-1 (Simple Clauses)* refers to a simple clause composed of a subject (*S*), a verb (*V*), and an optional object (*O*), respectively. A TEDU of this type has a pattern of either *SV* or *SVO*, corresponding to a sentence with an intransitive verb or one with a transitive verb, respectively, as shown in the table.
(2) *TEDU-2 (Subject Zero-Anaphora Clauses)* refers to a clause with its subject omitted. In general, it is possible for a clause to share its subject with its preceding clause, especially in coordination. In the table, ϕ refers to a zero anaphora (some parts are omitted).

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

Fig. 1. Six types of TEDUs, four types of COMPs, and two types of TEDU-LPs, enhanced from Carlson et al. [2003] for Thai unit definition. Here, a square bracket specifies a TEDU, parentheses surround a COMP, a brace denotes a TEDU-LP, an italicized item represents a verbal unit, and a boldfaced item displays a discourse cue.

(3) *TEDU-3 (Clauses with Attribution Verb)* refers to a clause with an attribution verb (V_A). Normally it expresses a speech act or a cognitive act. In the table, *X* expresses the speech act particle and *P* means a prepositional phrase.

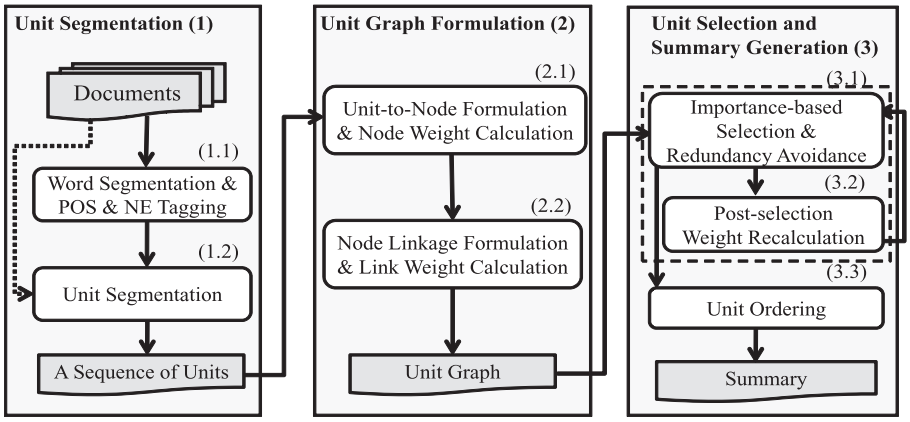


Fig. 2. Three main processes with their subprocesses in the summarization model.

(4) *TEDU-4 (Comparative Clauses)* refers to a clause with a special verb (V_C) and a comparative cue (R), such as "more (V_C) than (R)" and "higher (V_C) than (R)", as shown in the table.
(5) *TEDU-5 (Question Clauses)* refers to an interrogative sentence, that is a clause with a question word. In the table, W denotes a question word that may occur in the beginning or ending of a clause.
(6) *TEDU-6 (Embedded Conjunction Clauses)* refers to an embedded conjunction clause, namely, a clause with a conjunction (C) embedded inside a clause rather than in front of a clause. Its function is to connect two clauses, as shown in the table.
(7) *TEMPP (Temporal Phrases)* refers to an expression of date, time, or duration. It may contain a time prefix (P_T), a time suffix (S_T), day (D), month (M), year (Y), hour (H), minute (M), and second (S).
(8) *SPATP (Spatial Phrases)* refers to the location/place (L) in the clause. Sometimes, it also contains a preposition (P_S) at the beginning of a spatial phrase.
(9) *CONJP (Conjunction Phrases)* refers to a single conjunction (C), a conjunction with a prepositional prefix ($P_C C$), or a conjunction with an adverbial suffix (CS_C), as shown in the table.
(10) *EMBP (Embedded Phrases)* refers to a relative pronoun E_P, such as which or that.
(11) *TEDU-LP-1 (Clausal Subjects/Objects)* refers to a unit that has a structure of a verb leading with a nominalized prefix (P_{NC}), corresponding to a gerund form in English.
(12) *TEDU-LP-2 (Synthetic Nominal Compounds)* refers to a unit that has a structure like a TEDU, that is, SV, SVO, or VO, which can be $P_{NS}V$, $P_{NS}VS_{NS}$, and VS_{NS}, as shown in the table.

3. THAI MULTI-DOCUMENT SUMMARIZATION

This section presents our Thai multi-document summarization model that is composed of three processes: unit segmentation, unit graph formulation, and unit selection and summary generation. Figure 2 displays the framework, composed of three main processes and their subprocesses. In the unit segmentation (process 1 in Figure 2), a Thai running text is segmented into a sequence of tractable units and tagged with part-of-speech (POS) and named entities (NEs) (subprocess 1.1 in Figure 2). Three alternative

forms, called TEDUs, combined TEDUs, and paragraphs, are proposed for segmenting a Thai running text into units (subprocess 1.2 in Figure 2). In the unit graph formulation (process 2 in Figure 2), a graph of units is constructed by conceptualizing a unit as a weighted node in a graph (subprocess 2.1 in Figure 2) and a relationship of two nodes is formulated as a weighted link between the nodes (subprocess 2.2 in Figure 2). The weight of a node or link is determined by considering its importance, that is, its contribution in the graph. In the unit selection (process 3 in Figure 2), a number of important nodes and links are selected by considering the importance level of nodes or links, together with redundancy among units (nodes) (subprocess 3.1 in Figure 2), and focusing on the node weight recalculation (subprocess 3.2 in Figure 2). In the final step (subprocess 3.3 in Figure 2), a summary can be generated by ordering selected units under a set of predefined criteria. In this work, for sake of simplicity, the summary is generated by displaying selected units under two general conditions: (1) temporal order among documents and (2) occurrence order in each document. The details are discussed in the next sections.

3.1. Unit Segmentation

Since Thai language has no sentence boundary (or even word boundary), it is difficult to define a unit for summarization. However, paragraphs as units for Thai document summarization are too large, since a paragraph may contain heterogeneous contents and because such a paragraph-based approach does not allow to exclude some unimportant parts in a paragraph. With this issue in mind, TEDUs detection is useful, but requires some forms of word segmentation and POS/NE tagging. Towards this issue of TEDUs detection, it is necessary to perform word segmentation and POS/NE tagging. Besides TEDUs, an alternative, namely a combined TEDU, is proposed under the hypothesis that two TEDUs can be merged to form a larger unit if there exist some clues of connection between them. In the rest of this section, word segmentation and POS/NE tagging (process 1.1 in Figure 2) as well as unit segmentation (process 1.2 in Figure 2) are described.

— *Word Segmentation and POS/NE Tagging.* In the past, a number of techniques have been implemented as tools for Thai word segmenetation, part-of-speech tagging (POS tagging), and named entity tagging (NE tagging), such as SWATH [Meknavin et al. 1997] and Thai E-Class [Tongtep and Theeramunkong 2013]. Recently, the Thai E-Class has been developed as a tool for segmenting a Thai running text into words and for tagging each word with either NEs, POS, and/or a semantic role from five types or 4W1H (What, Where, Who, When, and How). In this work, word segmentation and tagging of NEs, POS, and 4W1H are performed as preprocess for recognizing TEDUs, COMPs, and TEDU-LPs.
— *Unit Segmentation.* For this step, we apply a set of simple metarules, called left-to-right longest matching [Maier 1978] to uniquely determine the segmentation of a Thai running text into TEDU-LPs, TEDUs, and COMPs, using the segmented words with tags (i.e., NEs, POS, and 4W1Hs). 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 (CTEDU) by a set of predefined rules on discourse markers or connectors, or even to simply use a paragraph as a unit. Therefore, when there exist some clues (e.g., “when”, “since”, “in a city”, “On December 1, 2013”, etc.) of connection between them, two TEDUs are merged to form CTEDU.
In summary, the three types of units used in this work are TEDU+COMP, CTEDU, and paragraph (PARA).

3.2. Unit Graph Formulation

After a running text is split into tractable units, namely, TEDU+COMP, CTEDU, or paragraph, we need a model to determine the importance of units and to select the most suitable ones for a summary. Towards this end, a graph model is proposed to express units and their relations extracted from multiple targeted documents for summarization. In our approach, a node in a graph corresponds to a unit while a link in a graph expresses relation connections between two units (processes 2.1 and 2.2 in Figure 2).

— *Node Weight Calculation.* As a well-known method, it is possible to apply TF/IDF (Term Frequency-Inverse Document Frequency) to weigh words in a document and then weigh a unit by calculating the summation of weights of all words in this unit.

— *Link Weight Calculation.* Besides node weights, it is worth investigating the relations between two units (nodes). A link weight (relation strength) between two nodes (units) is defined to describe the degree to which two units are identical or related. While there are several measures for defining the link weight, such as cosine similarity [Singhal 2001], Jaccard similarity coefficient [Jaccard 1901], and Euclidean distance [Deza and Deza 2009], this work uses cosine similarity [Singhal 2001] due to its simplicity and practicality. Here, the link weight between two units, say u_{ij} and $u_{i'j'}$, is defined by the cosine similarity between the vectors of these two units as follows.

$$sim(u_{ij}, u_{i'j'}) = \frac{\vec{u}_{ij} \cdot \vec{u}_{i'j'}}{|\vec{u}_{ij}| \cdot |\vec{u}_{i'j'}|}. \quad (1)$$

Normally cosine similarity ranges from 0 to 1 and a higher value indicates higher similarity, implying that two units are redundant or highly related.

3.3. Unit Selection and Summary Generation

Given the unit graph derived from the process described in the previous section, unit selection and summary generation are tasks to select a set of suitable units for constructing a summary (process 3 in Figure 2) and to generate a summary using the selected units. In the past, several works [Goldstein and Carbonell 1998] 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 Ketui and Theeramunkong [2010]. Our unit selection starts from iteratively selecting potential units based on priority (weight), but trying not to include a unit if it is duplicated with some selected units in the summary S. Finally, the weights of the rest of the nodes u_i in the graph of units in documents G are recalculated. As shown in Algorithm 1, the number of units to be selected is set. The most important nodes are repeatedly added one-by-one into the summary S and then these nodes are deleted from a set of unselected units U until the number of selected units reaches a predefined compression rate. During the node addition, the weight of each unselected unit is recalculated. When the number of selected units satisfies the predefined compression rate, the algorithm will return the graph of summary S. Three basic concepts of our inclusion-based summarization approach (subprocesses 3.1 and 3.2 in Figure 2) are discussed next.

— *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).

ALGORITHM 1: The inclusion-based unit selection

Input: a set of units $U = \{u_1, u_2, ..., u_I\}$,
a set of unit weights $W = \{w_1, w_2, ..., w_I\}$,
a set of link weights $E = \{e_{11}, e_{12}, ..., e_{21}, e_{22}, ..., e_{II}\}$ *(i.e., similarity among nodes)*,
a predefined compression rate cr.
Output: a set of selected units for the summary S.
Set the number of units to be selected n_s to $(I \times cr)$;
repeat
select the node $u_s \in U$ with the highest weight $(s = \underset{i}{\textbf{argmax}}[w_i])$;
(by considering three factors: importance (W), centroid (E), and redundancy (E).);
add u_s into the summary S $(S = S \cup \{u_s\})$.;
delete u_s from the set of unselected units U $(U = U - \{u_s\})$.;
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 against other unselected units.);
until *the total number of nodes is greater than or equal to* n_s;
Result: S;

(2) A preferably unselected unit is a node with higher degree of similarity to all other unselected units. In other words, the unit close to the centroid of the unselected units should be included in the summary.

— *Redundancy avoidance.* Two units with identical or 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.

— *Postselection weight recalculation.* After a unit is selected to include in a summary, its selection affects the possibility that other units will be selected. The implication is that, after a unit is selected, the selection probability of an unselected unit depends on the ratio of similarity of the unselected unit against the selected unit, and the similarity of the unselected unit against the other unselected units.

According to the first concept on weighting units, the unit selection can be formulated as follows (here, the original weighting of a unit $W(u)$ is modified to reflect a centroid-related and a redundancy-related factor, as shown in Eq. (2)). The best unit $(\hat{u})$ can be selected by maximizing the value in the equation, where the three terms indicate original weight, centroid-related, and redundancy-related factors, respectively. Let u indicate the current unit. The unselected unit is represented by u_i that occurs in the whole unselected set $|U|$, where i is the i-th unselected unit while s_j displays the selected unit in a summmary, where j is the j-th selected unit. When more than one unit has the same highest score, the most preceding one in the latest document will be selected.

$$\hat{u} = \arg\max_{u} \left[W(u) \times \frac{\sum_{i=1, u_i \neq u}^{|U|} sim(u_i, u)}{|U| - 1} \times (1 - \max_{j}(sim(s_j, u))) \right]. \tag{2}$$

For the first term, TF/IDF can be used to express such importance levels of words by using term (word) frequency and inverse document frequency. The second term expresses the average similarity between the current unit and other units; a high value indicates the closeness to the centroid of the unselected units. The third term represents the level of content redundancy. As an extreme case, if the unit has similar

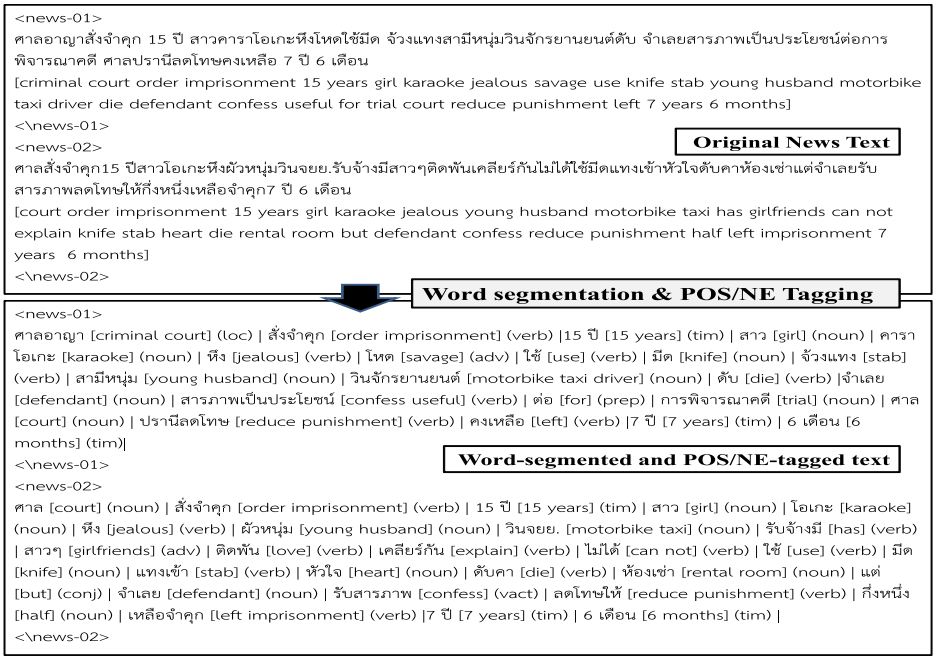


Fig. 3. Examples of word segmentation and POS/NE tagging. English word-by-word glosses are provided in square brackets.

content to that in the set of selected units, the maximum is 1 and the term will become 0. As the third concept, after a unit is selected to include in a summary, its selection affects the possibility that other units will be selected. As shown in Eq. (3), 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 this node (u) and the selected node s_j in set S and other nodes (u_i).

$$W_{re}{}^t(u) = W^t(u) \times \left(1 - \frac{sim(u, s_j)}{\sum_{i=1, u_i \neq u}^{|U|} sim(u, u_i)}\right). \tag{3}$$

Then the new node weight $W^{t+1}(u)$ is normalized by $W^{t+1}(u) = \frac{W_{re}{}^t(u)}{\sum_{i=1}^{|U|} W_{re}{}^t(u_i)}$. The steps (select, add, delete, and recalculate) in the algorithm are iterated to select the next node until the compression rate reaches a predefined value. After the unit selection, a summary can be generated by ordering selected units under a set of predefined criteria. Later the summary can be revised manually for readability improvement. However, in this work, for sake of simplicity, we simply generate a summary by ordering the selected units in the temporal relations among documents and the ordinal relations in the original documents.

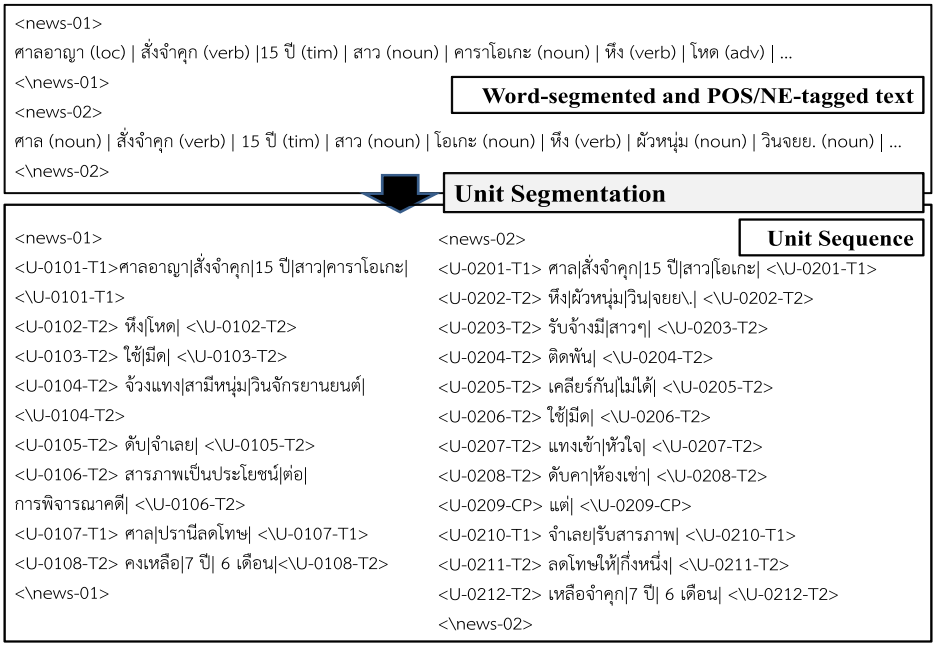


Fig. 4. Examples of unit (TEDU+COMP segmentation).

3.4. An Example

For clarity, Figures 3 to 5 are provided to illustrate a trace of three processes with two news text samples. Figure 3 shows two original news texts (<news-01> and <news-02>) with their glosses in the upper part and the corresponding word-segmented and POS/NE-tagged texts in the lower part. In Figure 3, segmentation is expressed by a vertical bar while the POS or NE type of each segmented word is given inside parentheses. English word-by-word glosses are provided in square brackets. Figure 4 displays the process of forming a sequence of one or more segmented words into units (either TEDUs or COMPs). Note that, since each TEDU-LP becomes a part of an EDU, it is not included in the result. In the lower part of Figure 4, there are eight TEDUs for <news-01> and 11 TEDUs with one COMP for <news-02>. Figure 5 illustrates the process of unit graph formulation, unit selection, and summary generation. In the middle part of Figure 5, three sample nodes (i.e., units) (NW-0101, NW-0102, and NW-0212) and two linking edges (EW-0101-0102 and EW-0101-0212) are shown with their weights (i.e., 1.459, 0.622, and 1.219 for the nodes and 0.101 and 0.688 for the edges). Finally in the lower part, a summary (at the compression rate of 0.4) is generated by selecting a set of nodes (TEDUs or COMPs) with a criterion (such as higher weights) and listing them under two general conditions, that is, the temporal order among documents and the occurrence order in each document. Here, two different dashed boxes show the resulting TEDUs or COMPs of <news-01> and <news-02>, respectively.

4. METHOD AND PROCEDURE

This section describes the dataset, evaluation method, and experimental settings. As for the evaluation criteria, a number of ROUGE-based measures are used to define the differences between manual summary and system summary. To evaluate the performance of our methods, four experiments are conducted.

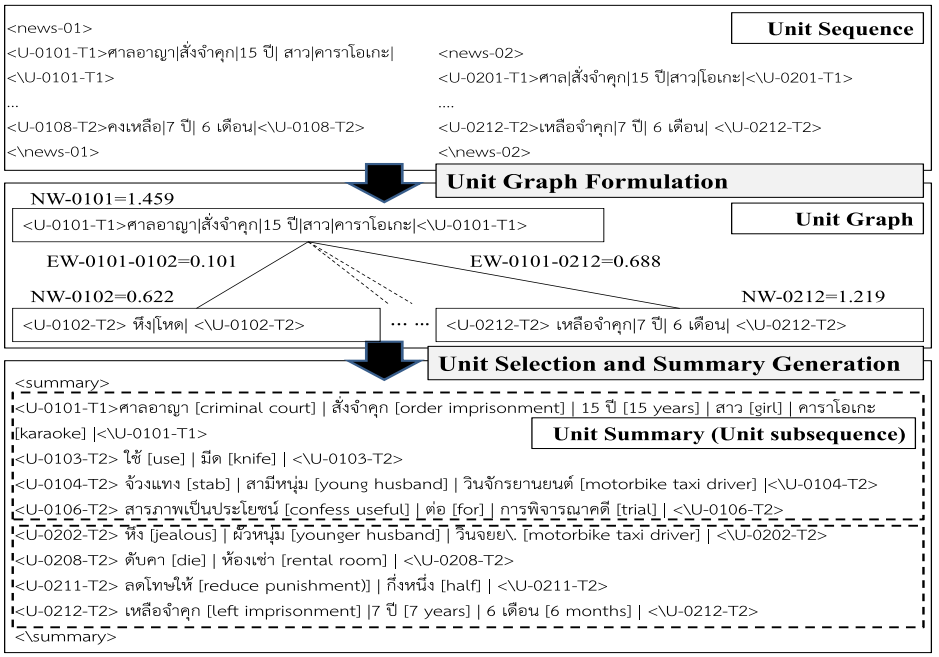


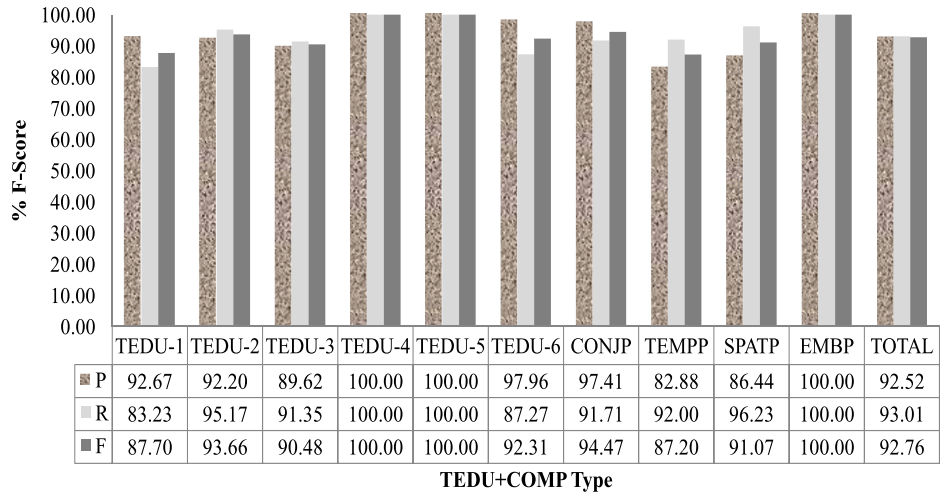
Fig. 5. Examples of unit graph formulation, unit selection, and summary generation.

4.1. Dataset and Preprocessing

This work utilizes the THAI-NEST corpus developed in Theeramunkong et al. [2010] comprised of 10,000 news articles in seven categories: crimes, sports, foreign affairs, politics, entertainment, economics, and education. These are from Thai news agencies such as *Daily News*[1], *Thairath*[2], and *Matichon*[3]. Later, a method for discovering document relations in Kittiphattanabawon et al. [2010] is applied to find relations among news documents and to group highly related news documents into a dataset for summarization. In this work, we randomly select 50 sets of related documents with completely related (CR) and somehow related (SH) relations for testing our proposed graph-based summarization approach. Each set of related documents are tagged with NEs and POS by Thai E-Class [Tongtep and Theeramunkong 2013].

Later our Thai running text with POSs and NEs tagging are segmented into TEDUs, COMPs, and TEDU-LPs by using 446 context-free grammar rules (CFG rules) with chart parsing [Ketui et al. 2012; 2013]. We applied three groups of CFG rules, namely, 342 rules for TEDUs, 95 rules for COMPs, and 9 rules for TEDU-LPs. Conceptually, the TEDUs and COMPs can be detected after recognizing TEDU-LPs while a CTEDU can be constructed by merging two related TEDUs using COMPs. Utilizing results in Ketui et al. [2013], the longest matching technique was found to outperform the maximum matching technique. The performance of the longest matching is displayed as a graph in Figure 6. Here, a pattern column shows the precision value, a gray column illustrates the recall value, and a black one displays the F-score value. The last type of units, paragraph, can be simply detected by line breaks and indents. If the text is in

[1]www.dailynews.co.th
[2]www.thairath.co.th
[3]www.matichon.co.th



	TEDU-1	TEDU-2	TEDU-3	TEDU-4	TEDU-5	TEDU-6	CONJP	TEMPP	SPATP	EMBP	TOTAL
P	92.67	92.20	89.62	100.00	100.00	97.96	97.41	82.88	86.44	100.00	92.52
R	83.23	95.17	91.35	100.00	100.00	87.27	91.71	92.00	96.23	100.00	93.01
F	87.70	93.66	90.48	100.00	100.00	92.31	94.47	87.20	91.07	100.00	92.76

Fig. 6. Precision (P), Recall (R), and F-Score (F) of Longest Matching (LM) in TEDU detection.

Table I. Characteristics of the 50 Experimental Datasets.

Group of dataset		Original documents							Reference summary	
		Size(KB)	#Words	#TEDUs	#COMPs	#CTEDUs	#Paragraphs	#Documents	Size(KB)	#Words
All datasets (50 datasets)	Sum	483.1	23,781	6,757	3,130	3,380	351	144	67.4	3,407
	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

HTML format, common markers are <p> or
 tags. To create a reference summary (i.e., model summary) for evaluation, a number of Thai language experts at the Faculty of Liberal Arts, Thammasat University manually constructed an abstractive-based summary with a size of 50–100 words for each of the 50 datasets as the gold standard for evaluating system results. The summaries contain main contents in the original documents, including What, Where, Who, When, and How (4W1H). Some discourse markers are added to connect clauses. These reference summaries will be used for evaluating a summary obtained from a system.

Table I displays characteristics of the 50 experimental datasets grouped by number of documents per set, including document size, number of words, number of TEDUs, number of COMPs, number of CTEDUs, number of paragraphs, and number of documents as well as the size of reference summary and number of words in the reference summary. The details show summation, average, maximum, and minimum by dataset. Our 50 datasets used for experiments contain 144 documents and 23,781 words. The size of all datasets is approximately 483.1KB and the average size per set is 9.7KB. The maximum size equals 31.0KB while the minimum is 4.0KB. The number of TEDUs is greater than twice that of CTEDUs, implying 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. This tendency is also true for the numbers of words, TEDUs, COMPs, CTEDUs, and paragraphs. The last two columns in Table I show characteristics of the reference summary of all datasets. The average size of a dataset of the reference summary is 1.3KB. The maximum size equals 2.8KB, while the minimum is 0.7KB. The number of words of the reference summary is less than 7× the original dataset. This trend is also true for the number of words in the reference summaries. Table II illustrates the detailed results of unit segmentation of the 50 datasets,

Table II. Number of Words and Units in 50 Datasets, Grouped by Types of TEDUs, COMPs, and TEDU-LPs

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

including number of words and units grouped by types of TEDUs, COMPs, and TEDU-LPs. Considering all units, there are 23,781 words and 9,887 TEDUs+COMPs in total. In detail, the table presents the summation, average, maximum, and minimum of each type. The most frequent units are those 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 50 datasets but each unit is quite long (i.e., on average 25÷5 = 5 words/units). Although we have more TEDU-3 units than TEDU-1 units, on average, a TEDU-3 unit (2,731÷1,453 = 1.9 words/unit) is shorter than a TEDU-1 unit (3,846÷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. They usually have one word per unit and are used to connect two TEDUs. The temporal and spatial phrases (TEMPP and SPATP) include, on average, two words (1,033÷562 = 1.8 and 1,304÷716 = 1.8) while the embedded phrases (EMBP) consist of only one word. Moreover, 572 TEDU-LP-1 units and 345 TEDU-LP-2 units are embedded in TEDUs or COMPs.

4.2. Experimental Setting and Evaluation Method

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, three 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 3.1). The three summarization factors we considered are: (1) importance-based selection; (2) redundancy avoidance; and (3) postselection weight recalculation. For importance-based selection, we consider the simple highest-weight priority (H) and an extension of the highest-weight priority with centroid preference (C). As for redundancy avoidance, we compare the consideration of redundancy penalty (P) to the nonconsideration version (D). For postselection weight recalculation, we also compare the recalculation case (R) with its nonrecalculation one (N). 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 half of the original documents. Here, the compression rate is defined as the ratio of number of units in a summary to that of units in the original documents. Moreover, as an optimal case, we calculate upper bound performance (UPB) of summarization by starting from selecting the unit that obtains the highest ROUGE, adding it into the summary,

and then selecting the next unit that achieves the best performance when added into the summary. This greedy-based selection is performed repeatedly until reaching the target compression rate. The average of results of all eight methods is also provided for reference. Moreover, we compare our summarization methods with a traditional graph-based ranking model called TexRank [Mihalcea 2004]. The sentence scoring function is known as the PageRank algorithm [Page et al. 1998]. The second experiment compares performance of the combinations of the three factors. For each factor, two alternatives are investigated by varying other factors and then comparing their results. Moreover, the performance of the combinations is summarized and ranked to clarify which factor combination is optimal. The third experiment performs a two-tailed t-test with significance value 0.05 to check win/loss/tie (W/L/T) among all eight methods in order to make a detailed comparison. Lastly, the fourth experiment investigates the quality of generated summaries by humans. For this task, six Thai linguists from Faculty of Liberal Arts of Thammasat University are requested to read the generated texts of 50 datasets and evaluate their readability. Since this task is labor intensive, we have evaluated only two systems that obtain the best ROUGH performance.

To evaluate a summary output from a system, we use the reference summaries as described in Section 4.1. For each set of related news articles, the reference summary is constructed by requesting Thai linguists to read the articles and then manually make their abstractive summary. In this work, we utilize ROUGE [Lin 2004] to evaluate a system's summarization result by comparing it with its reference summary. Among various types of ROUGE, this work uses R-1 (unigram-based co-occurrence statistics), R-2 (bigram-based co-occurrence statistics), and R-SU4 (skip-bigram plus unigram-based co-occurrence statistics) that are commonly used for evaluation. Originally developed by NIST, the ROUGE is a variant of ROUGEs that consider precision (R-1_P, R-2_P, R-SU4_P), recall (R-1_R, R-2_R, R-SU4_R) and F-score (R-1_F, R-2_F, R-SU4_F). Here, we utilize the result of the five compression rates (0.1–0.5) for performance comparison under consideration of the three ROUGE values (R-1, R-2, R-SU4) across the eight methods.

5. EXPERIMENTAL RESULTS AND DISCUSSION

This section reports four experimental results: (1) investigation of summarization performance on three unit types, three summarization factors, and five compression rates; (2) performance comparison of the eight combinations of the three factors with pairwise comparison and overall ranking; (3) win/loss/tie (W/L/T) checking among the eight methods by a two-tailed t-test with significance value 0.05; and (4) quality assessment by humans. To obtain insight into the results, error analysis is made and reported at the end of this section.

5.1. Performance Investigation on Three Unit Types, Three Summarization Factors, and Five Compression Rates

The results in three F-score ROUGEs (R-1_F, R-2_F, and R-SU4_F) are presented in Table III. This section provides observations on the results based on unit types. As for TEDU+COMP, CPR outperforms other methods with R-1 at 29.00 at compression rate 0.5, while CPN is superior with the best R-2 of 11.99 and R-SU4 of 12.61 at compression rate 0.2. Compared to UPB (the optimal case), around 20%–30% performance is achieved. The PageRank (henceforth PRK) is used as our baseline. At compression rate 0.5, PRK achieves the highest R-1 of 28.54 while obtaining the highest R-2 and R-SU4 of 9.89 and 10.21. Its performance is lower than our method. As for CTEDU (the second unit type), HPR gets the best R-1, R-2, and R-SU4 (29.07, 9.77, and 11.24) at compression rate 0.4. The performance of the eight methods tends to be high at compression rate 0.4 while the highest values of UPB are at compression rate 0.2

Table III. ROUGE-1_F, ROUGE-2_F, and ROUGE-$SU4_F$ Performance of Three Unit Types (UT column), Three Summarization Factors (method column), Five Compression Rates (ranking from 0.1 to 0.5)

UT	Method	Compression Rate														
		0.1			0.2			0.3			0.4			0.5		
		R-1	R-2	R-SU4	R-1	R-2	R-SU4	R-1	R-2	R-SU4	R-1	R-2	R-SU4	R-1	R-2	R-SU4
TE	HDN	17.75	8.65	8.85	21.41	8.96	9.24	24.32	10.18	10.62	26.32^2	10.66^2	11.15^2	27.52^1	11.35^1	11.89^1
	HDR	18.35	8.03	8.24	22.26	9.11	9.38	25.09	10.12	10.49	27.29^2	10.69^2	11.16^2	28.46^1	11.66^1	12.21^1
	HPN	18.45	8.66	8.89	21.58	9.15	9.41	24.61	10.15	10.54	26.75^2	10.81^2	11.27^2	28.35^1	11.59^1	12.18^1
	HPR	18.14	7.67	7.98	22.54	9.17	9.42	25.37	10.17	10.54	27.39^2	10.83^2	11.31^2	28.65^1	11.81^1	12.38^1
	CDN	15.30	5.20	5.47	21.08	7.71	8.15	24.30	9.93	10.37	26.19^2	10.69^2	11.20^2	27.88^1	11.56^1	12.13^1
	CDR	16.36	5.43	5.81	22.07	7.83	8.31	25.09	9.57	10.04	27.46^2	10.98^2	11.54^2	28.81^1	11.84^1	12.43^1
	CPN	15.80	5.39	5.74	21.67	8.34	8.80	24.75	10.03	10.48	26.87^2	11.00^2	11.56^2	28.76^1	$\mathbf{11.99}^1$	$\mathbf{12.61}^1$
	CPR	16.53	5.50	5.89	22.47	8.26	8.71	25.63	9.98	10.50	27.81^2	11.14^2	11.72^2	$\mathbf{29.00}^1$	11.91^1	12.50^1
	AVG	17.09	6.81	7.11	21.88	8.57	8.93	24.90	10.02	10.45	27.01^2	10.85^2	11.36^2	$\mathbf{28.43}^1$	$\mathbf{11.71}^1$	$\mathbf{12.29}^1$
	PRK	20.36	6.66	6.83	23.12	7.79	7.95	24.24	9.10	9.43	24.39	9.72	9.93	28.54	9.89	10.21
	UPB	40.23	27.73	41.52	$\mathbf{40.42}^1$	$\mathbf{33.69}^1$	$\mathbf{41.71}^1$	40.39^2	32.70^2	41.69^2	39.45	31.73	40.66	36.47	30.07	37.43
CT	HDN	19.15	7.12	7.97	24.68	8.54	9.50	27.52	9.39	10.69	27.91^2	9.43^2	10.86^2	28.46^1	9.52^1	10.99^1
	HDR	18.87	6.66	7.33	24.71	8.49	9.54	27.34	9.45^2	10.63	28.58^1	9.61^1	10.91^1	28.51^2	9.36	10.79^2
	HPN	19.37	7.03	7.67	24.05	8.10	9.18	27.35	9.37	10.81	28.32^2	9.57^2	11.02^2	28.79^1	9.71^1	11.21^1
	HPR	18.58	6.01	6.84	24.36	8.48	9.54	27.23	9.40	10.62	$\mathbf{29.07}^1$	$\mathbf{9.77}^1$	$\mathbf{11.24}^1$	28.83^2	9.58^2	11.03^2
	CDN	16.09	5.28	5.71	22.59	7.61	8.25	25.22	8.79	9.64	26.32^2	9.16^2	10.38^2	26.90^1	9.19^1	10.45^1
	CDR	16.92	5.51	5.99	24.46	8.08	8.86	26.45^2	8.79	10.05	27.66^1	9.45^1	10.93^1	27.66^1	9.32^2	10.82^2
	CPN	16.80	5.80	6.27	23.39	7.98	8.84	25.84	9.14	9.96	26.59^2	9.15^2	10.43^2	27.11^1	9.27^1	10.65^1
	CPR	18.19	6.59	7.06	25.27	8.27	9.16	27.15	9.25	10.63	28.03^2	9.48^2	10.93^2	28.28^1	9.56^1	11.15^1
	AVG	18.00	6.25	6.85	24.19	8.19	9.11	26.76	9.20	10.38	27.81^2	$\mathbf{9.45}^1$	10.84^2	$\mathbf{28.07}^1$	9.44^2	$\mathbf{10.89}^1$
	PRK	18.84	5.94	6.36	23.52	7.83	8.51	25.21	8.03	8.90	25.99	8.62	9.83	25.42	8.64	9.75
	UPB	47.44	25.85	30.13	$\mathbf{55.64}^1$	$\mathbf{27.01}^1$	$\mathbf{31.31}^1$	54.09^2	26.08^2	29.77^2	49.90	24.62	27.45	45.18	22.74	24.65
PA	HDN	4.63	2.03	2.27	14.24	4.28	5.79	22.39	6.81	9.16	25.88^2	7.92^2	10.62^2	28.32^1	9.04^1	12.34^1
	HDR	4.49	1.43	1.85	15.54	4.97	6.98	23.25	7.12	10.10	26.68^2	8.69^2	11.84^2	27.95^1	$\mathbf{9.20}^1$	12.28^1
	HPN	4.49	1.43	1.85	15.56	4.97	6.98	22.95	7.03	9.93	26.44^2	8.43^2	11.42^2	$\mathbf{28.33}^1$	$\mathbf{9.20}^1$	$\mathbf{12.44}^1$
	HPR	4.49	1.43	1.85	15.65	4.98	7.04	23.09	7.06	9.96	26.48^2	8.54^2	11.69^2	28.02^1	9.17^1	12.33^1
	CDN	4.64	1.45	1.86	14.84	4.15	5.99	22.09	6.42	9.02	25.10^2	8.40^2	10.82^2	25.87^1	8.67^1	10.91^1
	CDR	4.64	1.45	1.86	14.66	3.94	5.76	22.85	7.37	9.89	26.37^2	8.63^2	11.22^2	27.19^1	9.05^1	11.76^1
	CPN	4.64	1.45	1.86	14.67	3.97	5.91	22.10	6.39	9.04	25.08^2	8.35^2	10.81^2	25.91^1	8.65^1	11.00^1
	CPR	4.64	1.45	1.86	14.48	3.83	5.66	22.75	7.27	9.83	26.35^2	8.55^2	11.23^2	27.66^1	9.18^1	12.07^1
	AVG	4.58	1.52	1.91	14.96	4.39	6.27	22.68	6.93	9.62	26.05^2	8.44^2	11.21^2	$\mathbf{27.40}^1$	$\mathbf{9.02}^1$	$\mathbf{11.89}^1$
	PRK	3.49	0.54	0.74	15.81	4.21	6.53	22.44	6.27	9.08	25.64	8.23	10.69	26.39	8.68	11.45
	UPB	8.79	4.90	6.69	26.85	11.27	14.76	36.43	14.86	19.02^2	$\mathbf{40.80}^1$	$\mathbf{16.21}^1$	$\mathbf{20.03}^1$	40.38^2	15.74^2	18.55

Here, superscripts are given to the highest and second highest ROUGE, indicating the compression rate that achieves the highest ROUGE for each method of each unit type. Here, TE represents TEDU+COMP, CT stands for CTEDU, and PA means PARA (paragraph).

and then drop at compression rate 0.3. The average of R-1 and R-SU4 reaches highest levels at compression rate 0.5 (28.07 and 10.89) and the second highest values at compression 0.4 (27.81 and 10.84). The average of R-1, R-2, and R-SU4 over all methods is higher than the baseline method (PRK) at the whole range of compression rates. Considering a larger unit called paragraph (PARA), R-1, R-2, and R-SU4 are low at compression rate 0.1 because only in some datasets could we select the unit to generate a summary. It is possible that the multiplication of the total number of paragraphs and the compression rate is less than 1. Moreover, R-1 values of the eight methods are slightly higher at around 11% from compression rate 0.1 to 0.2 and then approximately increase to 7% and 3% at compression rate 0.3 and 0.4, respectively. HPN gets the highest performance of R-1, R-2, and R-SU4 at compression rate 0.5 while achieving the second highest at compression rate 0.4. On average, R-1, R-2, and R-SU4 become the highest at compression rate 0.5 for PARA, while PARA with PRK gets low performance of R-2 and R-SU4. To compare the performance of three units using eight different methods at compression rates between 0.1 and 0.5, the average of R-1 values of CTEDU (24.97) is greater than TEDU+COMP (23.86) and PARA (19.93), that is, CTEDU > TEDU+COMP > PARA. Unlike R-1, the ranking of average of R-2 values is TEDU+COMP > CTEDU > PARA (9.59, 8.51, and 6.06). R-SU4 follows the same pattern as R-2 (10.03, 9.61, and 8.18).

To conclude, we found that both the highest and second highest ROUGE values are at compression rates between 0.4 and 0.5. The performance of TEDU+COMP is

Table IV. Ranking Methods by the Average of ROUGE-1, ROUGE-2, and ROUGE-SU4

Rank No.	Method	ROUGE-1			ROUGE-2			ROUGE-SU4			AVG
		P	R	F	P	R	F	P	R	F	F
1	HPR	19.00	37.13	**23.19**	06.47	15.56	08.27	07.55	16.49	09.58	**13.68**
2	HDR	18.95	37.22	23.16	06.48	15.72	**08.31**	07.54	16.57	09.58	**13.68**
3	HPN	18.81	37.29	23.03	06.49	15.99	**08.35**	07.58	16.89	09.65	**13.68**
4	HDN	18.35	36.99	22.70	06.35	16.00	08.26	07.36	16.70	09.46	13.47
5	CPR	18.86	36.76	22.95	06.27	15.49	08.01	07.32	16.21	09.26	13.41
6	CDR	18.60	36.01	22.58	06.13	15.04	07.82	07.14	15.69	09.02	13.14
7	CPN	18.15	35.17	22.00	06.11	15.11	07.79	07.08	15.60	08.93	12.91
8	CDN	17.91	34.16	21.63	06.00	14.50	07.61	06.92	14.89	08.69	12.64
	PRK	16.95	47.99	22.81	06.02	19.09	08.30	06.86	20.41	09.50	13.54

Here, the compression rates used for evaluation are from 0.1 to 0.5.

generally greater than CTEDU and paragraph. It suggests that the length of TEDU+COMP is suitable for summarization since TEDU+COMP might be short and contains important information (keywords) that overlaps with the reference summary. CTEDU comes from TEDU and COMP and also obtains a high performance. Summarization using paragraph units may not be effective, since a paragraph may include heterogeneous content, where some should be in summary and some not. Our proposed methods can achieve 30%–50% of UPB (optimal case). The performance of our methods is relatively higher than the baseline graph-based approach (PRK). However, the results in Table III show that the performance of the proposed methods obtains a similar value. To distinguish them, we calculate the average F-score of three ROUGE values (R-1_F, R-2_F, and R-SU4_F) and rerank these scores as shown in Table IV. In this table, we see that HPR has the best performance for R-1 but is lower than HDR and HPN for R-2. Both HPR and HDR achieve the highest F-score for R-SU4. Moreover, there is no difference on the average F-score of three ROUGE values for HPR, HDR, and HPN but most methods outperform PRK. CDN achieves the lowest value. From this result, we cannot determine which method certainly wins out. Therefore, we perform another experiment in Section 5.3 and explore the effect of three factors on summarization performance in Section 5.2.

5.2. Effect of Three Factors on Summarization Performance

This section explores the effect of three factors on the methods. For each factor, two alternatives are investigated by varying other factors and then comparing their results. Table V demonstrates the performance comparison with a two-tailed t-test with 5% significance among three factors. Each comparison contains 150 cases (three units $\times$ 50 datasets) and applies a t-test score. The average of ROUGE-1_F, ROUGE-2_F, and ROUGE-SU4_F is used to find the number of wins, losses, and ties (W/L/Ts).

Table V shows that the simple highest-weight priority (H**) tends to outperform the highest-weight priority with centroid preference (C**). One possible reason is that the centroid preference (that selects the common unit from unselected units) may suggest a unit that overlaps with the currently selected units for the summary. When comparing the methods with (*P*) and without redundancy avoidance (*D*), we found that *P* outperforms *D* with two wins and two ties. For example, CPR beats CDR by 0.27 (13.41 versus 13.14) with 5% significance. This indicates the effectiveness of redundancy removal in multi-document summarization. The redundancy can be found when we form a unit graph in our second step, showing the effectiveness of the graph formulation. As for the last factor, the versions with postselection weight recalculation

Table V. Performance Comparison among Three Factors, Using a Two-Tailed T-Test with 5% Significance

Factor	Method (M1, M2)	AVG(ROUGE$_F$) (M1, M2)	t-test	W/L/T
H** vs. C**	HPN, CPN	13.68, 12.91	0.00	W
	HDN, CDN	13.47, 12.64	0.00	W
	HPR, CPR	13.68, 13.41	0.10	T
	HDR, CDR	13.68, 13.14	0.00	W
P vs. *D*	CPN, CDN	12.91, 12.64	0.00	W
	HPN, HDN	13.68, 13.47	0.20	T
	CPR, CDR	13.41, 13.14	0.00	W
	HPR, HDR	13.68, 13.68	0.99	T
**R vs. **N	CPR, CPN	13.41, 12.91	0.00	W
	CDR, CDN	13.14, 12.64	0.00	W
	HPR, HPN	13.68, 13.68	0.94	T
	HDR, HDN	13.68, 13.47	0.22	T

(**R) outperform the non-recalculation version (**N) with two wins and two ties. The weight recalculation will adjust the weight of a node to a more suitable value.

In conclusion, three factors, namely, the simple highest-weight priority, the redundancy avoidance, and the postselection weight recalculation, have positive effect on the performance of the summarization method. With a suitable combination, the performance can be improved. In more detail, the performance of our proposed method is investigated by varying the three factors and comparing them with a two-tailed t-test with 5% significance, as discussed in the next section.

5.3. Overall Comparison of Proposed Methods

In this experiment, we rank the average ROUGE$_F$ values of our eight methods and perform a two-tailed paired t-test at 5% significance level. Table VI shows that each method is tested with 105 pairs (5 compression rates $\times$ 3 ROUGE values $\times$ 7 methods). Following a normal schema, we set the scores of a win, loss, and tie to 3, 0, and 1, respectively. There are 150 cases from 50 datasets and three types of units. The row and column headers show the method names, and the cell values represent the number of wins, losses, and ties. The value in each cell shows the number of compression rates where the method in the row is superior (win), inferior (loss), and comparable (tie) on the column.

For each evaluation metric, we noticed that HPR performs better than the other methods, except for HPN, in terms of R-1 and R-SU4. This method only loses to HPN in terms of R-2. HDR has 23 wins, comparable to CPR, but its number of losses is lower than CPR. Specifically, for HPN there are 19 wins, 3 losses, and 83 ties. CDN never won, indicating that the highest-weight priority with centroid preference does not perform well. Furthermore, the performance of CPR increases, since the effect of the highest-weight priority with centroid preference, the redundancy avoidance, and the postselection weight recalculation can help improve the performance. Although the previous result showed this method is at the fifth rank (see Table V), we found that the average ROUGE values at the initial compression rate are very low and the values slightly increase at compression rates 0.4–0.5. In addition to CPR, even if they have the centroid-related consideration and either the redundancy avoidance or the postselection weight recalculation, this is not enough to get high accuracy. Since the centroid of the unselected units is considered the common information as the important unit, redundancy avoidance is necessary. Including the postselection weight recalculation may reduce the weight of unselected

Table VI. Performance Comparison among 8 Methods with a Two-Tailed T-Test at 5% Significance

Method	ROUGE	HPR	HDR	HPN	HDN	CPR	CDR	CPN	CDN	Overall	Score	Rank
HPR	R-1	-	1/0/4	2/0/3	2/0/3	0/0/5	1/0/4	3/0/2	5/0/0	14/0/21		
	R-2	-	1/0/4	0/1/4	0/0/5	0/0/5	1/0/4	0/0/5	2/0/3	4/1/30		
	R-SU4	-	1/0/4	0/0/5	0/0/5	0/0/5	2/0/3	1/0/4	3/0/2	7/0/23		
	Overall	-	**3/0/12**	**2/1/12**	**2/0/13**	**0/0/15**	**4/0/11**	**4/0/11**	**10/0/5**	**25/1/79**	**154**	**1**
HDR	R-1	0/1/4	-	1/0/4	2/0/3	0/0/5	1/0/4	4/0/1	5/0/0	13/1/21		
	R-2	0/1/4	-	0/0/5	0/0/5	1/0/4	1/0/4	1/0/4	1/0/4	4/1/30		
	R-SU4	0/1/4	-	0/0/5	0/0/5	1/0/4	1/0/4	1/0/4	3/0/2	6/1/28		
	Overall	**0/3/12**	-	**1/0/14**	**2/0/13**	**2/0/13**	**3/0/12**	**6/0/9**	**9/0/6**	**23/3/79**	**148**	**2**
HPN	R-1	0/2/3	0/1/4	-	0/0/5	0/0/5	1/0/4	3/0/2	4/0/1	8/3/24		
	R-2	1/0/4	0/0/5	-	0/0/5	1/0/4	1/0/4	1/0/4	1/0/4	5/0/30		
	R-SU4	0/0/5	0/0/5	-	0/0/5	1/0/4	1/0/4	1/0/4	3/0/2	6/0/29		
	Overall	**1/2/12**	**0/1/14**	-	**0/0/15**	**2/0/13**	**3/0/12**	**5/0/10**	**8/0/7**	**19/3/83**	**140**	**4**
HDN	R-1	0/2/3	0/2/3	0/0/5	-	0/0/5	0/0/5	1/0/4	2/0/3	3/4/28		
	R-2	0/0/5	0/0/5	0/0/5	-	0/0/5	1/0/4	1/0/4	1/0/4	3/0/32		
	R-SU4	0/0/5	0/0/5	0/0/5	-	0/0/5	1/0/4	1/0/4	2/0/3	4/0/31		
	Overall	**0/2/13**	**0/2/13**	**0/0/15**	-	**0/0/15**	**2/0/13**	**3/0/12**	**5/0/10**	**10/4/91**	**121**	**5**
CPR	R-1	0/0/5	0/0/5	0/0/5	0/0/5	-	4/0/1	5/0/0	5/0/0	14/0/21		
	R-2	0/0/5	0/1/4	0/1/4	0/0/5	-	2/0/3	0/0/5	1/0/4	3/2/30		
	R-SU4	0/0/5	0/1/4	0/1/4	0/0/5	-	2/0/3	2/0/3	2/0/3	6/2/27		
	Overall	**0/0/15**	**0/2/13**	**0/2/13**	**0/0/15**	-	**8/0/7**	**7/0/8**	**8/0/7**	**23/4/78**	**147**	**3**
CDR	R-1	0/1/4	0/1/4	0/1/4	0/0/5	0/4/1	-	3/0/2	5/0/0	8/7/20		
	R-2	0/1/4	0/1/4	0/1/4	0/1/4	0/2/3	-	0/0/5	0/0/5	0/6/29		
	R-SU4	0/2/3	0/1/4	0/1/4	0/1/4	0/2/3	-	1/0/4	2/0/3	3/7/25		
	Overall	**0/4/11**	**0/3/12**	**0/3/12**	**0/2/13**	**0/8/7**	-	**4/0/11**	**7/0/8**	**11/20/74**	**107**	**6**
CPN	R-1	0/3/2	0/4/1	0/3/2	0/1/4	0/5/0	0/3/2	-	4/0/1	4/19/12		
	R-2	0/0/5	0/1/4	0/1/4	0/1/4	0/0/5	0/0/5	-	1/0/4	1/3/31		
	R-SU4	0/1/4	0/1/4	0/1/4	0/1/4	0/2/3	0/1/4	-	2/0/3	2/7/26		
	Overall	**0/4/11**	**0/6/9**	**0/5/10**	**0/3/12**	**0/7/8**	**0/4/11**	-	**7/0/8**	**7/29/69**	**90**	**7**
CDN	R-1	0/5/0	0/5/0	0/4/1	0/2/3	0/5/0	0/5/0	0/4/1	-	0/30/5		
	R-2	0/2/3	0/1/4	0/1/4	0/1/4	0/1/4	0/0/5	0/1/4	-	0/7/28		
	R-SU4	0/3/2	0/3/2	0/3/2	0/2/3	0/2/3	0/2/3	0/1/4	-	0/16/19		
	Overall	**0/10/5**	**0/9/6**	**0/8/7**	**0/5/10**	**0/8/7**	**0/7/8**	**0/6/9**	-	**0/53/52**	**52**	**8**

Table VII. Text Quality Measures of Two Summarization Methods at Compression Rate 0.4 (Scoring Ranging from 1 to 5)

Aspects of Text Quality	HPR				CPR			
	H_1	H_2	H_3	AVG	H_4	H_5	H_6	AVG
Grammaticality	3.42	3.36	3.83	3.54	1.69	2.33	2.09	2.04
Non-Redundancy	3.83	3.57	3.68	3.69	1.79	2.63	2.28	2.23
Referential clarity	4.11	3.81	3.75	3.89	2.33	3.10	2.51	2.65
Readability and Coherence	3.28	3.64	3.38	3.43	1.65	2.52	2.43	2.20
AVG	3.66	3.60	3.66	3.64	1.87	2.65	2.33	2.28

unit, similar to the previously selected units. Besides this, we found that postselection weight recalculation is a part of top-3 methods, therefore this factor is very important for summarization tasks. The ranking of the eight methods is HPR > HDR > CPR > HPN > HDN > CDR > CPN > CDN. Note that the ranking of the methods in Tables IV and VI are not the same. We conclude that HPR has the best performance at compression rates between 0.1 and 0.5 and obtains the best performance for CTEDU, while CPN and CPR work well with TEDU+COMP.

5.4. Quality Assessment by Humans

In this experiment, we evaluate the text summarization quality. The evaluation is conducted by asking six Thai linguists (H_1–H_6) to rate news summaries in four aspects. We select HPR and CPR for quality assessment by humans since they are the best methods for CTEDU and TEDU+COMP, respectively. Moreover, by this means, we can evaluate simultaneously the effect of unit types on summary quality. Table VII displays the average scores of text quality measures of CTEDU with HPR and TEDU+COMP

with CPR summarization methods. Four aspects of text quality evaluated include the following.

(1) *Grammaticality.* Text summarization should not contain the non-textual items, punctuation errors, or incorrect words.
(2) *Nonredundancy.* Text summarization should not contain redundancy content.
(3) *Reference Clarity.* Text summarization should contain the nouns and pronouns clearly referred to in the summary.
(4) *Readability and Coherence.* Text summarization should have a good summary or be easy to read.

The Thai linguists gave rating scores of 1–5 (1 = fail; 2 = poor; 3 = average; 4 = good; 5 = very good) to each summary from 50 datasets of Thai news articles. The results show that three aspects of text quality (grammaticality 3.54, nonredundancy 3.69, readability and coherence 3.43) of CTEDU with HPR get average ratings, and good rating (3.75) for referential clarity. On the other hand, for TEDU+COMP with CPR, three aspects of text quality (2.04, 2.23, and 2.20, respectively) are quite poor while referential clarity is close to average (2.65). For CTEDU with HPR, the average of four aspects of text quality (3.64) is greater than TEDU+COMP with CPR (2.28). Furthermore, one additional observation is that CTEDU seems to gain better summary quality than TEDU+COMP since the larger units, that is, CTEDUs obtaining conjunctive cohesion, are more understandable than TEDU+COMP units. According to the linguists' suggestions, the unit ordering should be improved because the disordered summary may make readers misunderstand the content. In this work, the summary is easily generated by listing the selected units in the temporal relations among documents and the ordinal relations in the original documents. However, a summary is likely to obtain better performance if semantic components are considered when it is automatically generated. This is left as our future work.

5.5. Error Analysis

In all, we ranked the 50 datasets based on their performance (average F-score over all ROUGEs and compression rates). Some observations can be drawn. First, datasets with a very low ratio of document size between the reference summary and the original seem to obtain low performance in summarization. This implies the difficulty in selecting suitable phrases/words for a summary due to the large number of candidates. Second, datasets with heterogeneous contents or somehow related (SH) relations tend to get low performance. For instance, while one of the datasets consists of 15 documents, only five documents are strongly related and the others trivially so. The task of summarizing heterogeneous content documents is difficult. Third, datasets that occupy documents with a small number of paragraphs tend to get low performance for paragraph-based units (PARA). When unit selection is done based on the paragraph, it is highly possible to include words/phrases that are not suitable to be included in the summary since the paragraph-based method will select the whole paragraph as one unit. Lastly, datasets with improper TEDU+COMP or CTEDU segmentation may trigger lower performance in ROUGE-2 and ROUGE-SU4. Selection on missegmented units is not efficient, resulting in low performance.

6. SUMMARIZATION ENGINE AND GUI

Up to now, most applications for multi-document summarization have usually been related to the domain of news including both public and commercial sites. For

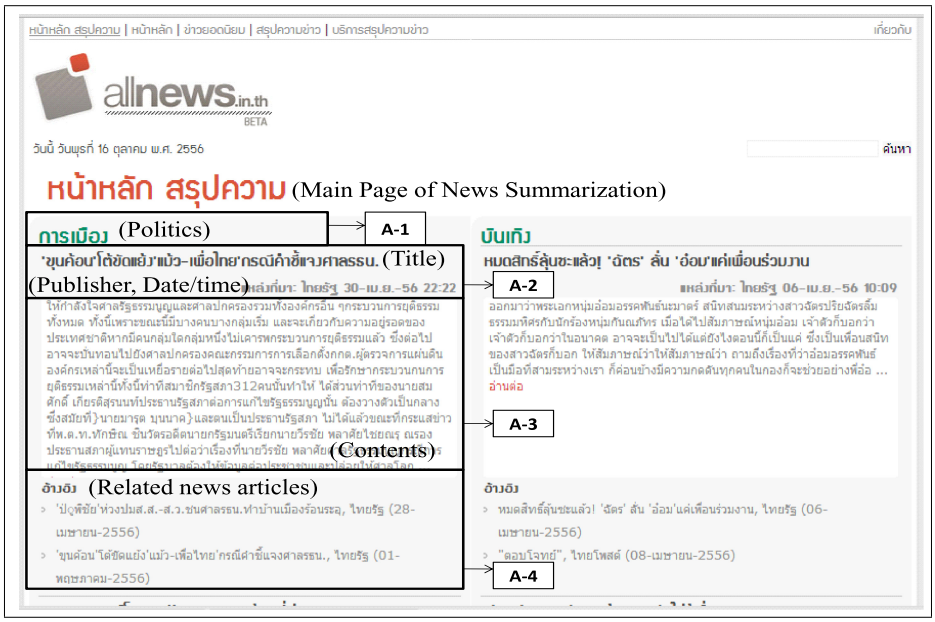


Fig. 7. Main page of Thai automatic summarization application.

example, Google News[4] and Columbia NewsBlaster[5] are public and commercial sites while News In Essence[6] is a commercial one. In this work, we have developed a Thai automatic summarization application that is part of the AllNews system[7] with a concept of unit segmentation, unit graph formulation, and unit selection. For our implementation, the following software and hardware requirements are used. The system was developed using C++ language and shell script. GCC (GNU Compiler Collection) version 4.6.3, which is an integrated distribution of compilers for several major programming languages, is used. We edited all programs with Eclipse version Juno. We installed Java and Perl language programming for supporting unit segmentation and evaluation methods. The Web interface of the system was constructed by PHP (Personal Home Page) programming. A MySQL database was used for storing original texts and the summarization results. The system ran on Ubuntu 10.04, Intel Xeon 3.3 MHz, 8GB RAM, and 1TB HDD.

In the system of Thai automatic summarization, the original text can be gathered from online news articles. Figures 7 to 10 show user interface (UI) snapshots when the system generates a summary from related news articles. Figures 11 and 12 display the manual input. Details and descriptions for Figures 7 through 12 are as follows.

(1) Figure 7 displays of Thai news articles from eight categories (crimes, sports, foreign affairs, politics, entertainment, economics, general news, and education) (A-1); news title, publisher, date/time (A-2); and contents (A-3). Each topic is shown and ordered by date and time in ascending order. Related news articles are displayed (A-4), including the publisher and date/time. The related news articles are found by

[4]http://news.google.com
[5]http://newsblaster.cs.columbia.edu
[6]http://NewsInEssence.com
[7]http://203.131.209.100/c/summarizedPageMain.php

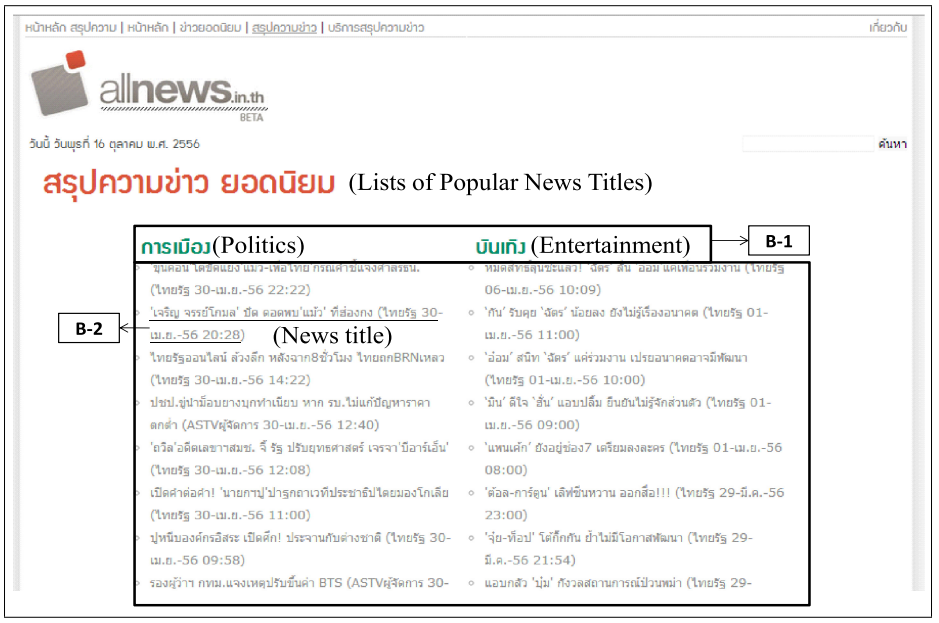


Fig. 8. A page of Thai automatic summarization application showing lists of popular news titles.

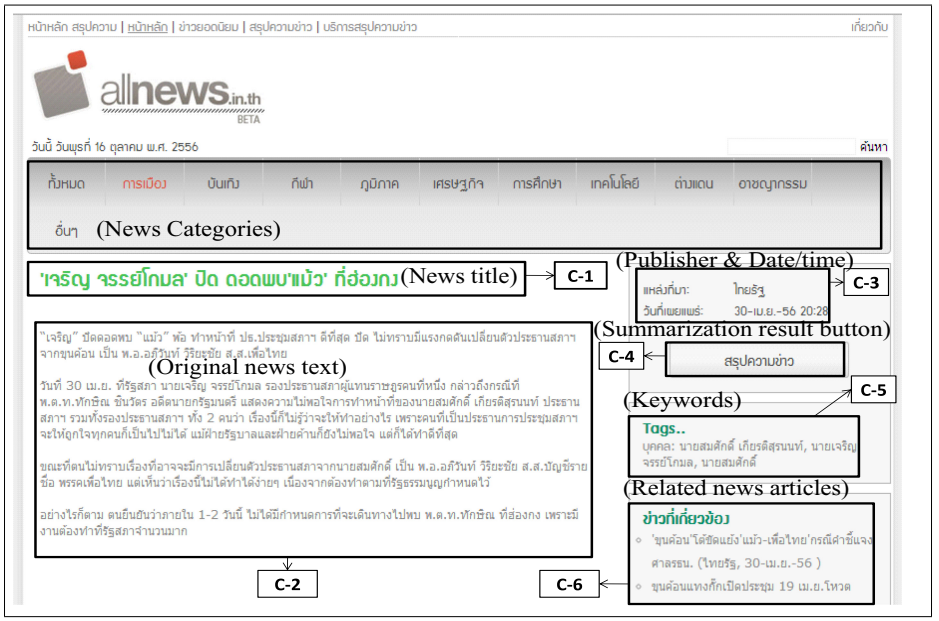


Fig. 9. A content page of Thai automatic summarization application prior to summarization process.

the NewsRelation() function. An association rule mining (ARM) technique is used to find the mined news. This function is executed at all times and uses original news articles from the main database within a span of 30 days.

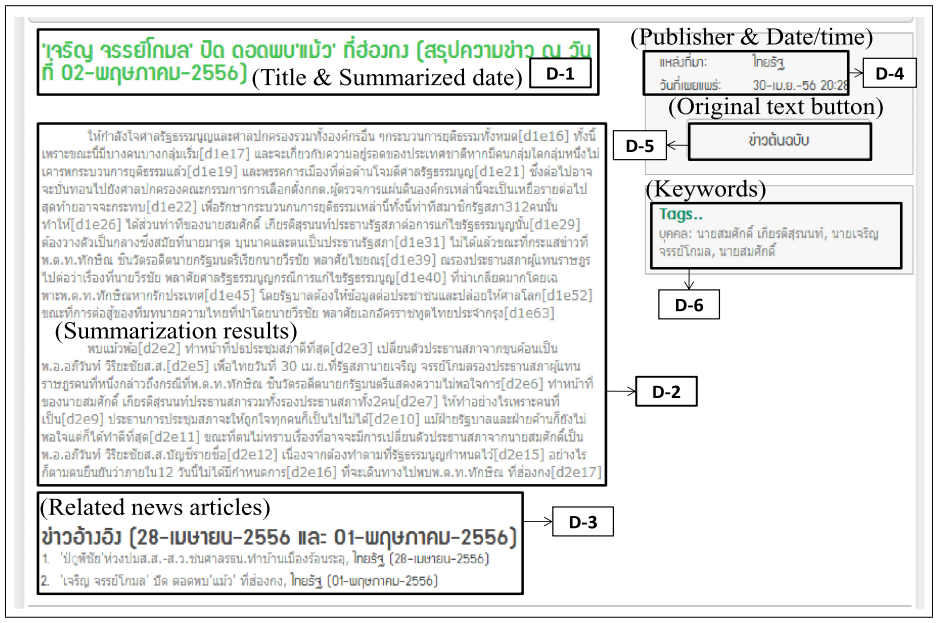


Fig. 10. A page of Thai automatic summarization application showing summarization results.

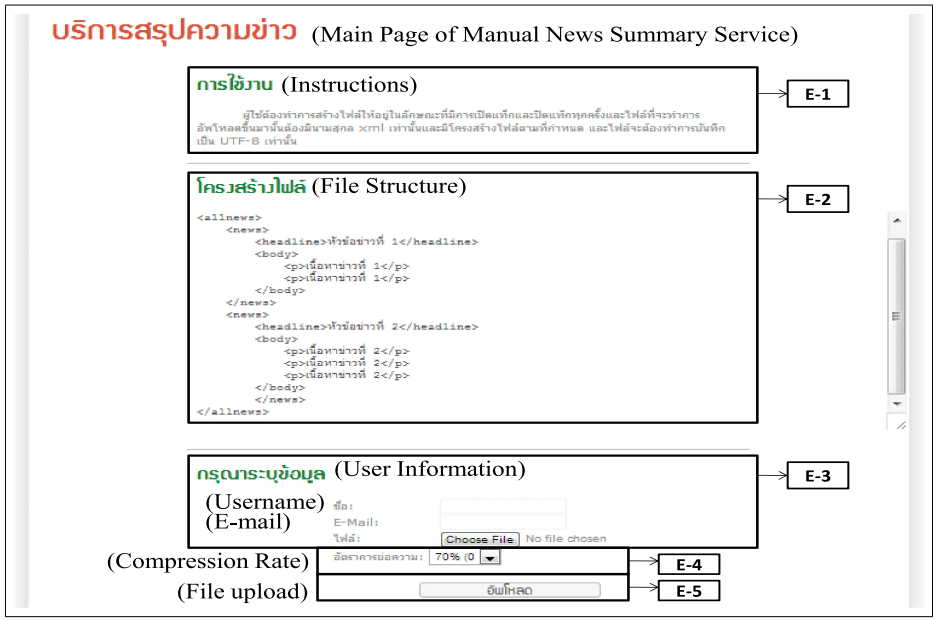


Fig. 11. Main page of Thai automatic summarization application showing manual service.

(2) Figure 8 shows current popular news titles (B-2) grouped into the eight categories, that is, crimes, sports, foreign affairs, politics, entertainment, economics, general news, and education (B-1). Popular news titles of each category are retrieved from

Fig. 12. A page of Thai automatic summarization application showing summarization results.

the database within seven days by the ShowPopularTopic() function. The news title is automatically linked to its summary results as shown in Figure 9.

(3) Figure 9 Illustrates the news title (C-1), original news text (C-2), publisher and date/time (C-3), summarization result button (C-4), keywords (C-5), and related news articles (C-6). These details are shown when applying the ShowContentByTopic() function. To search for keywords or named entities, the NETagging() function is used to segment a running text into a tractable unit (word), tag with part-of-speech, and to extract relevant words for display. Related news articles of the NewsRelation() function and POS/NE of NETagging() function can be set at the input of the TextSummarizer() function.

(4) Figure 10 demonstrates the title and summarized date (D-1), summarized results (D-2), and its related news articles (D-3). The summary is retrieved by the ShowSummaryByTopic() function. In this application, the number of related news articles is set to 2–5 articles for each summarization processing. The TextSummarizer() function reads the words with POS/NE tagging from related news articles and detects the Combined TEDU (CTEDU). The output of this function (CTEDU) is the input unit for a CoreSummarizer() function. The TextSummarizer() function is also a unit segmentation of the Thai automatic summarization framework and composed of the NewsRelation() function, NETagging() function, and UnitDetection() function. To obtain the summary results, the CoreSummarizer() works through multiple documents with an inclusion-based approach and sets the size of summarization at compression rate 0.3 in order to display the results as the Combined TEDU (CTEDU).

(5) Figure 11 shows instructions for an input of original news text (E-1). The formatted input should be set in XML format since the NETagging() function needs to use a tag for word segmentation (E-2). This page is shown when applying the SummaryByManual() function according to the manual input when a user uploads an input into the system and fills in the user information (i.e., Name and Email) (E-3).

Users can also select among different levels of compression rate (E-4). Within the next three days the system will send the news summarization to the user since the response time depends on the number of summary processes in a row.

(6) Figure 12 depicts of summarization results with the manual input (F-1). The ShowSummaryByManual() function is applied to process the following CoreSummarizer() function.

7. CONCLUSION AND FUTURE WORK

This article provided a definition of the Thai Elementary Discourse Unit (TEDU) and then presented our three-stage method of Thai multi-document summarization, that is, unit segmentation, unit graph formulation, and unit selection and summary generation. In the unit segmentation process, we investigated three different units: TEDU+COMP, CTEDU, and paragraph. For TEDU+COMP, a Thai running text is segmented into TEDUs, COMPs, and TEDU-LPs. The unit graph formulation represents units as weighted nodes in a graph and their relationships as weighted links among nodes, according to their importance and contribution in the graph. The unit selection is performed to include important nodes and links by considering redundancy among units (nodes) and content difference among units. After the unit selection, a summary is generated by ordering the selected units in the temporal relations among documents and the ordinal relations in the original documents. Three factors that have been considered include importance-based selection, redundancy avoidance, and postselection weight recalculation. The 50 sets of Thai news articles are used for evaluation. The results show that TEDU+COMP yields the best performance in terms of R-2 and R-SU4 while CTEDU is superior in terms of R-1. When the average ROUGE F-score is used for ranking, HPR and HDR have superior performance. In other words, it is highly effective to select units based on their weights with consideration for redundancy avoidance and weight recalculation.

As for future work, we will analyze the relations among TEDUs in order to form combined TEDUs with considerations of semantics. Moreover, we plan to improve initialized node weights and investigate more semantic-based unit selection, including consideration of discourse structure via conjunctions or discourse markers, as well as TEDUs and COMPs. Moreover, it is worth exploring a query-based approach where a query is provided to find related documents before summarization. An investigation on a larger dataset will help us evaluate our method in a more practical way. Finally, we plan to make some improvements on our news summarization application, such as an easy-to-use GUI, a mobile and real-time version, and an XML-compatible implementation.

REFERENCES

Alguliev, R. M., Aliguliyev, R. M., Hajirahimova, M. S., and Mehdiyev, C. A. 2011. Mcmr: Maximum coverage and minimum redundant text summarization model. *Expert Syst. Appl. 38*, 12, 14514–14522.

Aliguliyev, R. M. 2009. A new sentence similarity measure and sentence based extractive technique for automatic text summarization. *Expert Syst. Appl. 36*, 4, 7764–7772.

Barzilay, R., McKeown, K. R., and Elhadad, M. 1999. Information fusion in the context of multi-document summarization. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*. 550–557.

Cai, X. and Li, W. 2011. A spectral analysis approach to document summarization: Clustering and ranking sentences simultaneously. *Inf. Sci. 181*, 18, 3816–3827.

Carbonell, J. and Goldstein, J. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98)*, 335–336.

Carlson, L., Marcu, D., and Okurowski, M. E. 2003. Building a discourse-tagged corpus in the frame-work of rhetorical structure theory. In *Proceedings of the 2nd SIGDIAL Workshop on Discourse and Dialogue (SIGDIAL'03)*.

Charoensuk, J., Sukvaree, T., and Kawtrakul, A. 2005. Elementary discourse unit segmentation for thai using discourse cues and syntactic information. In *Proceedings of the 6th Symposium on Natural Language Processing (SNLP'05)*.

Chongsuntornsri, A. and Sornil, O. 2006. An automatic thai text summarization using topic sensitive pagerank. In *Proceedings of the International Symposium on Communications and Information Technologies (ISCIT '06)*. 547–552.

Deza, M. M. and Deza, E. 2009. *Encyclopedia of Distances*. Springer.

Erkan, G. and Radev, D. R. 2004. Lexpagerank: Prestige in multi-document text summarization. http://clair.si.umich.edu/~radev/papers/emnlp04pos.pdf.

Ferreira, R., Cabral, L. D. S., Lins, R. D., Silva, G. P., Freitas, F., Cavalcanti, G. D., Lima, R., Simske, S. J., and Favaro, L. 2013. Assessing sentence scoring techniques for extractive text summarization. *Expert Syst. Appl. 40*, 14, 5755–5764.

Goldstein, J. and Carbonell, J. 1998. Summarization: (1) using mmr for diversity - based reranking and (2) evaluating summaries. In *Proceedings of the Workshop on Tipster Text Program (TIPSTER'98)*. Association for Computational Linguistics, 181–195.

Jaccard, P. 1901. Etude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin del la Societe Vaudoise des Sciences Naturelles 37*, 547–579.

Jaruskulchai, C. and Kruengkrai, C. 2003. A practical text summarizer by paragraph extraction for thai. In *Proceedings of the 6th International Workshop on Information Retrieval with Asian Languages (AsianIR'03)*. 9–16.

Ketui, N. and Theeramunkong, T. 2010. Inclusion-based and exclusion-based approaches in graph-based multiple news summarization. In *Proceedings of the 5th International Conference on Knowledge, Information and Creativity Support Systems (KICSS'10)*, Lecture Notes in Computer Science, vol. 6746, Springer, 91–102.

Ketui, N., Theeramunkong, T., and Onsuwan, C. 2012. A rule-based method for thai elementary discourse unit segmentation (ted-seg). In *Proceedings of the 7th International Conference on Knowledge, Information and Creativity Support Systems (KICSS'12)*, IEEE Computer Society. 195–202.

Ketui, N., Theeramunkong, T., and Onsuwan, C. 2013. Thai elementary discourse unit analysis and syntactic-based segmentation. *Inf.-Ann. Int. Interdiscipl. J. 16*, 10, 7423–7436.

Kittiphattanabawon, N., Theeramunkong, T., and Nantajeewarawat, E. 2010. Exploration of document relation quality with consideration of term representation basis, term weighting and association measure. In *Proceedings of the Pacific Asia Workshop on Intelligence and Security Informatics (PAISI'10)*, Lecture Notes in Computer Science, vol. 6122, Springer, 126–139.

Kuo, J.-J. and Chen, H.-H. 2008. Multidocument summary generation: Using informative and event words. *ACM Trans. Asian Lang. Inform. Process. 7*, 1, 3:1–3:23.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Proceeding of the ACL Workshop on Text Summarization Branches Out (WAS'04)*. 74–81.

Maier, D. 1978. The complexity of some problems on subsequences and supersequences. *J. ACM 25*, 2, 322–336.

Mani, I. 1997. Multi-document summarization by graph search and matching. In *Proceedings of the 14th National Conference on Artificial Intelligence and the 9th Conference on Innovative Applicatins of Artificial Intelligence (AAAI/IAAI'97)*, 622–628.

Mani, I. and Bloedorn, E. 1999. Summarizing similarities and differences among related documents. *Inf. Retriev. 1*, 35–67.

McKeown, K., Klavans, J., Hatzivassiloglou, V., Barzilay, R., and Eskin, E. 1999. Towards multi-document summarization by reformulation: Progress and prospects. In *Proceedings of the 16th National Conference on Artificial Intelligence and the 11th Innovative Applications of Artificial Intelligence Conference (AAAI/IAAI'99)*. 453–460.

McKeown, K. and Radev, D. 1999. Generating summaries of multiple news articles. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99)*. 74–82.

Meknavin, S., Charoenpornsawat, P., and Kijsirikul, B. 1997. Feature-based thai word segmentation. In *Proceedings of the Natural Language Processing Pacific Rim Symposium (NLPRS'97)*.

Mihalcea, R. 2004. Graph-based ranking algorithms for sentence extraction, applied to text summarization. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions (ACLdemo'04)*. Association for Computational Linguistics.

Okazaki, N., Matsuo, Y., and Ishizuka, M. 2005. Improving chronological ordering of sentences extracted from multiple newspaper articles. *ACM Trans. Asian Lang. Inform. Process. 4*, 3, 321–339.

Page, L., Brin, S., Motwani, R., and Winograd, T. 1998. The pagerank citation ranking: Bringing order to the web. http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf

Radev, D. R., Jing, H., and Budzikowska, M. 2000. Centroid-based summarization of multiple documents: Sentence extraction, utility-based evaluation, and user studies. In *Proceedings of the NAACL-ANLP Workshop on Automatic Summarization (NAACL-ANLP-AutoSum'00)*, 21–30. Association for Computational Linguistics. 21–30.

Singhal, A. 2001. Modern information retrieval: A brief overview. *Bull. IEEE Comput. Soc. Technic. Committee Data Engin. 24*, 4, 35–43.

Sinthupoun, S. and Sornil, O. 2010. Thai rhetorical structure analysis. *Int. J. Comput. Sci. Inf. Secur. 7*, 1, 95–105.

Sornil, O. and Gree-ut, K. 2006. An automatic text summarization approach using content-based and graph-based characteristics. In *Proceedings of the IEEE Conference on Cybernetics and Intelligent Systems (ICCIS'06)*, 1–6.

Sukvaree, T., Kawtrakul, A., and Caelen, J. 2007. Thai text coherence structuring with coordinating and subordinating relations for text summarization. In *Proceedings of the 6th International and Interdisciplinary Conference on Modeling and Using Context (CONTEXT'07)*, 453–466.

Suwanno, N., Suzuki, Y., and Yamazaki, H. 2005. Extracting thai compound nouns for paragraph extraction in thai text. In *Proceedings of the IEEE International Conference on Natural Language Processing and Knowledge Engineering (NLP/KE'05)*, 657–662.

Thangthai, A. and Jaruskulchai, C. 2004. Impact parameter on lsa performance for thai text summarization. In *Proceedings of the 43rd Kasetsart University Annual Conference: Veterinary Medicine, Science (Vichakarn'04)*. 331–339.

Theeramunkong, T., Boriboon, M., Haruechaiyasak, C., Kittiphattanabawon, N., Kosawat, K., Onsuwan, C., Siriwat, I., Suwanapong, T., and Tongtep, N. 2010. Thai-nest: A framework for thai named entity tagging specification and tools. In *Proceedings of the 2nd International Conference on Corpus Linguistics (CILC'10)*, 895–908.

Tongtep, N. and Theeramunkong, T. 2013. Multi-stage automatic ne and pos annotation using pattern-based and atatistical-based techniques for thai corpus construction. *IEICE Trans. Inf. Syst. E96-D*, 10, 2245–2256.

Wang, H. and Zhou, G. 2012. Toward a unified framework for standard and update multi-document summarization. *ACM Trans. Asian Lang. Inform. Process. 11*, 2, 5:1–5:18.

Received February 2014; revised July 2014; accepted July 2014