Thai News Text Summarization and Its Application

Nongnuch Ketui[†] **Thanaruk Theeramunkong**[†] **Chutamanee Onsuwan**[‡]
[†]School of Information, Computer, and Communication Technology,
Sirindhorn International Institute of Technology, Thammasat University, Thailand
[‡]Faculty of Liberal Arts, Thammasat University, Thailand
{nongnuch,thanaruk}@siit.tu.ac.th, consuwan@tu.ac.th

Abstract

Since Thai language lacks word/phrase/sentence boundaries, document summarization in Thai needs investigations in unit segmentation, unit selection, redundancy removal and evaluation dataset construction. In this work, we have proposed Thai Elementary Discourse Unit (TEDU) and a three-stage method of Thai multi-document summarization, i.e., unit segmentation, unit-graph formulation, and unit selection. A number of experiments are conducted using fifty sets of Thai news articles, the reference summaries of which are given. In terms of ROUGE-1, ROUGE-2, and ROUGE-SU4, the experimental results point out that our combined TEDU-based summarization outperforms paragraph-based summarization and that unit redundancy consideration helps improving summary quality.

Keywords: Thai Text Summarization, Multi-Document Summarization Application

1 Introduction

Nowadays a gigantic number of news articles is produced in any language all over the world. Such situation of having too much information makes it difficult for us to focus on needed information and follow the ongoing events from many related news articles from several sources with similar and different facts. To solve this problem, it is necessary to study an efficient and effective method to make a summary from multiple news articles. In this work, we introduce Thai Elementary Discourse Unit (TEDU) and common phrases (COMP) and then present a three-stage method of Thai multi-document summarization, i.e., unit segmentation, unit-graph formulation, and unit selection 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 removal, and (3) Post-selection weight recalculation. A number of experiments is conducted using fifty sets of Thai news documents; summary of performance evaluations is also given. Three measures of ROUGE-1, ROUGE-2, and ROUGE-SU4 are used as performance metrics.

In the rest, Section 2 describes the related work. Section 3 defines Thai elementary discourse unit (TEDU). A framework of Thai multi-document summarization is presented in Section 4. Methods and procedures are shown in Section 5. Experimental results are described in Section 6. Thai news summarization application is displayed in Section 7. Finally, conclusion and future work are given in Section 8.

2 Related Work

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 its infancy stage. Early works on Thai text summarization included Jaruskulchai and Kruengkrai [1] and Thangthai and Jaruskulchai [2], which proposed a paragraph-based summarization with selecting top-n paragraphs and considering of Latent Semantic Analysis (LSA), respectively. Later Suwanno et al. [3] 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 [4] 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. How-

ever, summarization based on paragraph is forced to select a whole paragraph for summary, even when some parts in a paragraph are important and some are not. To avoid this restriction, two recent works have proposed methods to select, to select segments rather than whole paragraphs for a Thai single summarization [5] and [6].

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 a solution, Sukvaree et al. [7] have presented an EDU-based summarization approach that extracts Elementary Discourse Units (EDUs) [8] from a text, then finds their discourse coherence and organizes them into a spanning tree based on the well-known rhetorical structure theory. Their work was based on a Thai agriculture corpus and the result showed that Thai discourse coherence can be refer to many different meanings so-called word-sense ambiguity. For example, some words can be both a conjunction and a marker cue. Moreover, definitions of elementary discourse units of Thai need to be established and discussed.

3 Thai Elementary Discourse Unit (TEDU)

Considering the definition of Thai EDUs (TEDUs) in those works and the definition of English EDUs by Carlson et al. [8], we have defined a set of TEDUs to reflect special characteristics in the Thai language. 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, some word sequences look syntactically similar to a TEDU by containing a verb as its 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. We define 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) [9] as follows.

1. **TEDU-1 (Simple clauses):** Refer to a simple clause, composed of a subject, a verb and an optional object, respectively.

2. **TEDU-2 (Subject zero-anaphora clauses):** Refer 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.

3. **TEDU-3 (Clauses with attribution verb):** Refer to a clause with an attribution verb. Normally it expresses a speech act or a cognitive act.

4. **TEDU-4 (Comparative clauses):** Refer to a clause with a special verb and a comparative cue, such as “more than” and “higher than”.

5. **TEDU-5 (Question clauses):** Refer to an interrogative sentence, a clause with a question word.

6. **TEDU-6 (Embedded conjunction clauses):** Refer to an embedded conjunction clause, a clause with a conjunction embedded inside a clause.

7. **TEMPP (Temporal phrases):** Refer to an expression of date, time or duration.

8. **SPATP (Spatial phrases):** Refer to the location/place elements in the clause.

9. **CONJP (Conjunction phrases):** Refer to a single conjunction, a conjunction with a prepositional prefix or a conjunction with an adverbial suffix.

10. **EMBP (Embedded phrases):** Refer to a relative pronoun, such as “which” or “that”.

11. **TEDU-LP-1 (Clausal subjects/objects):** Refer to a Thai clausal subject or object with a structure of a verb preceded by a nominalized prefix.

12. **TEDU-LP-2 (Synthetic nominal compounds):** Refer to a Thai synthetic nominal compound with a structure resembles that of a TEDU.

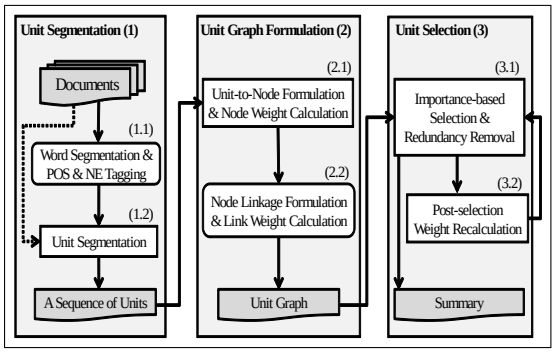

Figure 1. Three main processes with their subprocesses in the summarization model

4 Framework of Thai News Summarization

A framework of Thai multi-document summarization model is composed of three processes; unit segmentation, unit graph formulation, and unit selection. In the unit segmentation (1), a Thai running text is segmented into a sequence of tractable units and tagged with POSs and NEs (1.1). Three alternative forms, called Thai elementary discourse units (TEDUs) [9], 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 are 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 determined by considering its importance, i.e., its contribution in the graph. In the unit selection (3), a number of important nodes and links is selected by considering importance level of nodes or the 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 discussed in the next subsections.

4.1 Unit Segmentation

We have proposed a so-called predicate-based segmentation, where a Thai running text is split into units on the assumption that one unit contains only one predicate or a verb. To detect EDUs, it is necessary to perform word segmentation and POS/NE Tagging where a chunked text is segmented into words, and these words, parts-of-speech (POSs), or named entity tags (NE tags) are recognized in order to detect verbs or action words. Besides TEDUs, an alternative, namely a combined TEDU (CTEDU), is proposed on a hypothesis that two TEDUs can be merged to form a larger unit if there exist some clues of connection between them. Two processes are word segmentation and POS/NE Tagging (Process 1.1 in Figure 1) and unit segmentation (Process 1.2 in Figure 1).

4.2 Unit Graph Formulation

After a running text is split into tractable units; TEDU+COMP, CTEDU, or paragraph, we need a model to determine importance of units and select the most suitable ones for a summary. Towards this, a graph model is proposed to express units and their relations extracted from multiple targeted documents. In our approach, a node in a graph corresponds to a unit while a link in a graph expresses connections between two units (Processes 2.1 and 2.2 in Figure 1).

- **Node Weight Calculation:** As a node weight method, it is possible to apply iterative weight [4] to weigh words in a document and then weigh an unit by calculating the summation of weights of all words in that unit.

- **Link Weight Calculation:** A link weight (relation strength) between two nodes (units) is defined to describe how much two units are identical or related by the cosine similarity between the vectors of those two units.

4.3 Unit Selection

Unit selection is done to select a set of suitable units for constructing a summary (Process 3 in Figure 1). Our unit selection starts from iteratively selecting potential units based on priority (weight), by trying not to include it if the unit is duplicated with some selected units in the summary. Finally, the weights of the rest of the nodes in the graph of units in documents are recalculated. Three basic concepts of our 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 (iterative weight).
2. A preferably unselected unit is a node with high 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 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 unit is selected to include in a summary, its selection affects possibilities that other units will be selected. The implication is that after a unit is selected, selection 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.

5 Methods and Procedures

This work utilizes the THAI-NEST corpus developed in [10]. Fifty sets of related documents with completely-related and somehow-related relations [11] are selected for testing our proposed graph-based summarization approach. Given each set of related documents, the documents were tagged with name entities (NEs), parts of speech (POS) by Thai E-Class [12]. We applied three groups of context free grammar rules; 342 rules for TEDUs, 95 rules for COMPs, and 9 rules for TEDU-LPs [9]. To create a reference summary for evaluation, we have asked a number of Thai language experts at Thammasat University to read the fifty datasets of news articles and construct a summary of 50-100 words for each dataset. We use ROUGE-1, ROUGE-2, and ROUGE-SU4 [13]. The focused three summarization factors are importance-based selection, redundancy removal, and post-selection weight recalculation. For importance-based selection, we apply maximum weight. For redundancy removal, we compare consideration of redundancy penalty ('P') to non-consideration version ('D'). For post-selection weight recalculation, we also compare the recalculation case ('R') with its non-recalculation one ('N'). We examine five compression 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.

6 Experimental Results

Table 1 shows that the maximum of R-1 values of CTEDU (31.94) is greater than TEDU+COMP (29.82) and PARA (29.42) i.e., CTEDU $>$ TEDU+COMP $>$ PARA. Unlike R-1, the ranking of maximum of R-2 values is TEDU+COMP $>$ CTEDU $>$ PARA (12.50, 10.92, and 9.44). While the ranking of maximum of R-SU4 is PARA $>$ TEDU+COMP $>$ CTEDU (15.31, 13.10, and 12.48). To consider the average of ROUGE values, we found that the average of R-1 values follows the same pattern of the maximum of R-1 values i.e., CTEDU $>$ TEDU+COMP $>$ PARA (28.12, 26.94, and 21.86). Unlike R-1, the ranking of average of R-2 and R-SU4 are the same i.e., TEDU+COMP $>$ CTEDU $>$ PARA. For R-2, the ranking number is 11.47, 9.87, and 7.61 while R-SU4 is 12.00, 11.41, and 10.17). The average ROUGE-based (ROUGE-1, ROUGE-2, and ROUGE-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 2 shows the results in the order of the average f-score from ROUGE-1, ROUGE-2, and ROUGE-SU4. We found that PR is the best performance for ROUGE-1 but is slightly inferior to PN for ROUGE-2. Both PR and PN achieve the highest f-score for ROUGE-SU4. Moreover, the average f-score of three ROUGE values for PR, PN, DR, and DN do not differ.

7 Thai News Summarization Application

Up to now, most applications for multi-document summarization usually have accessed to the news domain, including public and commercial sites. For example, Google News[1] and Columbia NewsBlaster[2] are public sites while News In Essence[3] are a commercial one. For Thai, we have been developing a Thai automatic

[1] http://news.google.com
[2] http://newsblaster.cs.columbia.edu
[3] http://NewsInEssence.com

Table 1. 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	DN	22.17	9.98	10.31	24.27	10.16	10.65	25.47	10.59	11.12	24.76	10.68	11.20	24.72	10.88	11.40
	DR	23.75	10.30	10.69	27.27	11.11	11.70	28.91	11.59	12.21	28.51	11.56	12.14	27.97	11.43	11.99
	PN	24.02	11.12	11.59	27.68	11.90	12.48	28.92	12.10	12.66	29.52	12.33	12.92	28.78	12.23	12.84
	PR	24.37	11.45	11.96	28.43	12.43	12.97	29.68	**12.50**	13.05	**29.82**	12.49	13.08	29.73	12.49	**13.10**
	UPB	40.23	27.73	41.52	40.42	33.69	41.71	40.39	32.70	41.69	39.45	31.73	40.66	36.47	30.07	37.43
CT	DN	24.00	8.84	9.93	28.56	10.41	11.94	27.78	9.69	11.04	27.79	9.76	11.24	26.72	9.73	11.15
	DR	23.09	8.53	9.38	28.35	9.74	11.35	29.91	10.03	11.66	29.61	9.91	11.61	28.60	9.93	11.54
	PN	22.33	8.38	9.30	28.80	10.09	11.66	31.11	**10.92**	12.85	31.22	10.95	12.90	29.86	10.69	12.38
	PR	22.12	7.74	8.77	28.24	9.74	11.53	31.86	10.69	12.67	**31.94**	10.88	12.74	30.44	10.75	**12.48**
	UPB	47.44	25.85	30.13	55.64	27.01	31.31	54.09	26.08	29.77	49.90	24.62	27.45	45.18	22.74	24.65
PA	DN	6.06	3.19	4.02	20.42	7.61	9.83	26.31	**9.44**	12.21	28.28	9.74	12.84	28.71	9.74	12.84
	DR	5.05	1.84	2.38	19.12	6.81	8.62	24.74	8.36	11.08	27.62	9.33	12.45	29.51	10.33	13.63
	PN	5.05	1.84	2.38	19.68	6.91	9.16	25.32	8.37	11.50	**29.42**	9.92	13.84	31.84	11.24	**15.31**
	PR	5.05	1.84	2.38	19.53	6.92	9.16	25.17	8.35	11.42	29.06	9.71	13.57	31.16	10.67	14.78
	UPB	8.79	4.90	6.69	26.85	11.27	14.76	36.43	14.86	19.02	40.80	16.21	20.03	40.38	15.74	18.55

Table 2. Ranked methods by the average of ROUGE-1, ROUGE-2, and ROUGE-SU4. Here, compression rates used are between 0.1 and 0.5.

Rank	Method	ROUGE-1			ROUGE-2			ROUGE-SU4			AVG
		P	R	F	P	R	F	P	R	F	F
1	PR	0.2119	0.4222	**0.2644**	0.0744	0.1882	0.0991	0.0882	0.2010	**0.1158**	**0.1598**
2	PN	0.2087	0.4274	0.2624	0.0739	0.1938	**0.0993**	0.0876	0.2066	**0.1158**	0.1592
3	DR	0.2005	0.4252	0.2547	0.0684	0.1929	0.0939	0.0805	0.2017	0.1083	0.1523
4	DN	0.1913	0.4141	0.2440	0.0682	0.1960	0.0936	0.0806	0.2036	0.1078	0.1485

summarization application which is a part of All-News system[4] with the concepts from unit segmentation, unit graph formulation, and unit selection. For hardware and software implementation, the system requires a storage, at least a size of 1 TB, RAM of 8 GB., and an operating system of Ubuntu version 10.04. While Thai automatic summarization are compiled by C++ language and shell script. While PHP (Hypertext Preprocessor) is utilized for creating All-news website and collecting the news articles online in MySQL database.

Two main functions of Thai automatic summarization are developed; TextSummarizer() and CoreSummarizer().

1. TextSummarizer() focuses on the unit segmentation. This function firstly reads the set of related news from database within 30 days before the summaried day, segments Thai running text into a sequence of tractable unit (word/NE) and tags with POSs and NEs. Finally, the rule-based approach can be segmented Thai running text into a unit as CTEDU for summarization.

2. CoreSummarizer() constructs a graph of units (CTEDU as a output) and assigns the weight of node and link with the concept of unit graph formulation. Every week, this function summarizes multiple documents into a news summary which is a part of unit selection and applies the inclusion-based approach for Thai summarization.

[4]http://203.131.209.100/c/summarizedPageMain.php

8 Conclusion

This paper provided a definition of Thai Elementary Discourse Unit (TEDU) and then presented our three-stage method of Thai multi-document summarization. We investigated three different units; TEDU+COMP, CTEDU, and PARA. The results show that TEDU+COMP yields the best performance in terms of ROUGE-2 and ROUGE-SU4 while CTEDU is superior in terms of ROUGE-1. By ranking our proposed methods by the average ROUGE f-score, PR and PN have higher performance than the others. It is effective to select units based on their weights with consideration of redundancy removal and weight recalculation. Finally, the conceptualization of Thai news summarization is applied for web-based application.

Acknowledgment

This work was partially supported by Government Research Fund via Thammasat University, Thailand, the National Research University Project of Thailand Office of Higher Education Commission, as well as the National Electron-

ics and Computer Technology Center, Thailand under Project Number NT-B-22-KE-38-54-01.

References

[1] C. Jaruskulchai and C. Kruengkrai. A practical text summarizer by paragraph extraction for thai. In *Proc. of the sixth international workshop on Information retrieval with Asian languages - Volume 11*, AsianIR '03, pages 9–16, 2003.

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

[3] N. Suwanno, Y. Suzuki, and H. Yamazaki. Extracting thai compound nouns for paragraph extraction in thai text. In *Proc. of IEEE International Conference on Natural Language Processing and Knowledge Engineering*, IEEE NLP-KE '05, pages 657–662, October 30-November 1, 2005.

[4] N. Ketui and T. Theeramunkong. Inclusion-based and exclusion-based approaches in graph-based multiple news summarization. In *KICSS*, volume 6746 of *Lecture Notes in Computer Science*, pages 91–102. Springer Berlin / Heidelberg, 2010.

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

[6] O. Sornil and K. Gree-ut. An automatic text summarization approach using content-based and graph-based characteristics. In *Proc. of IEEE Conference on Cybernetics and Intelligent Systems*, pages 1–6, 2006.

[7] T. Sukvaree, A. Kawtrakul, and J. Caelen. Thai text coherence structuring with coordinating and subordinating relations for text summarization. In *Proc. of the 6th international and interdisciplinary conference on Modeling and using context*, CONTEXT'07, pages 453–466, 2007.

[8] L. Carlson, D.Marcu, and M. E. Okurowski. Building a discourse-tagged corpus in the framework of rhetorical structure theory. *Current Directions in Discourse and Dialogue*, 2003.

[9] N. Ketui, T. Theeramunkong, and C. Onsuwan. A rule-based method for thai elementary discourse unit segmentation (tedseg). In *Knowledge, Information and Creativity Support Systems*, pages 195–202. IEEE Computer Society, 2012.

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

[11] N. Kittiphattanabawon, T. Theeramunkong, and E. Nantajeewarawat. Exploration of document relation quality with consideration of term representation basis, term weighting and association measure. In *Intelligence and Security Informatics*, volume 6122 of *Lecture Notes in Computer Science*, pages 126–139. Springer Berlin / Heidelberg, 2010.

[12] N. Tongtep and T. Theeramunkong. Multi-stage annotation using pattern-based and statistical-based techniques for automatic thai annotated corpus construction. In *Proc. of the 9th Workshop on Asian Language Resources collocated with IJCNLP 2011*, pages 50–58, 2011.

[13] C-Y. Lin. Rouge: A package for automatic evaluation of summaries. In *Proc. of ACL Workshop on Text Summarization*, pages 74–81, 2004.