THESIS

GENERATING DISCOURSE STRUCTURE
FOR THAI TEXT SUMMARIZATION

THANA SUKVAREE

GRADUATE SCHOOL, KASETSART UNIVERSITY
2008

THESIS

GENERATING DISCOURSE STRUCTURE FOR THAI TEXT SUMMARIZATION

THANA SUKVAREE

A Thesis Submitted in Partial Fulfillment of
The Requirements for the Degree of
Doctor of Engineering (Computer Engineering)
Graduate School, Kasetsart University
2008

THESIS APPROVAL

GRADUATE SCHOOL, KASETSART UNIVERSITY

Doctor of Engineering (Computer Engineering)

DEGREE

Computer Engineering | Computer Engineering

FIELD | **DEPARTMENT**

TITLE: Generating Discourse Structure for Thai Text Summarization

NAME: Mr. Thana Sukvaree

THIS THESIS HAS BEEN ACCEPTED BY

(Associate Professor Asanee Kawtrakul, D.Eng.) **THESIS ADVISOR**

(Associate Professor Yuen Poovarawan, M.Eng.) **COMMITTEE MEMBER**

(Associate Professor Punpiti Piamsa-nga, D.Sc.) **COMMITTEE MEMBER**

(Associate Professor Tasanalai Burapacheep, Ph.D) **COMMITTEE MEMBER**

(Associate Professor Pradondet Nilagupta, M.Eng.) **COMMITTEE MEMBER**

(Assistant Professor Kemathat Vibhatavanij, Ph.D.) **DEPARTMENT HEAD**

APPROVED BY THE GRADUATE SCHOOL ON _____

(Associate Professor Gunjana Theeragool, D.Agr.) **DEAN**

TABLE OF CONTENTS

Page

TABLE OF CONTENTS	i
LIST OF TABLES	ii
LIST OF FIGURES	iii
INTRODUCTION	1
OBJECTIVES	11
LITERATURE REVIEW	12
MATERIALS AND METHODS	51
Materials	51
Methods	58
RESULTS AND DISCUSSION	77
Evaluation Methods	77
Results and Discussion	78
CONCLUSION	82
LITERATURE CITED	84
APPENDICES	93
APPENDIX A Resources	94
APPENDIX B Outputs	108

LIST OF TABLES

Table		Page
1	An approach in summarization research	4
2	showing an existing number of linguistic information in each pattern	34
3	Rhetorical relations presented in our corpus (Thai agriculture domain)	38
4	The number of explicit discourse markers categories by rhetorical relations	38
5	The number of implicit discourse markers categorized by rhetorical relations	39
6	Rhetorical relations classified as paratactic relations and hypotactic relations	50
7	The example of cue phrases/words or discourse marker	54
8	Appearance of Discourse Markers occurs with Discourse relations, In corpus. Let given stand for the some rhetorical relations that EB = Elaboration relation, CSQ = Consequence relation, CR = Cause-Result relation, CD = Condition relation and JT = Joint relation	56
9	Accuracy of rhetorical relations in Thai agriculture corpus	70
10	Accuracy of COR/SUBR relations in Thai agriculture corpus	70
11	Comparing the ARF performance with the RFC	78
12	Comparing the ARF performance with the RFC in flashback Phenomenon	78
13	Result of COR&SUBR relations Identification, comparing between baseline system and our system	79
14	Result of Tree Spans evaluation	80
15	Result of Salience Extraction evaluation	81
16	The quality of Summary evaluation	81

LIST OF FIGURES

Figure		Page
1	Sample of RST tree from Text-2	21
2	Discourse Update, attach new node into Right Frontier Area	24
3	Discourse structure	28
4	The example of Discourse structure of document	29
5	Tree structure of document “เชื้อพันธุ์ข้าว”	31
6	An example of repeating terms at the being position of paragraph	35
7	An example of the repeating terms at the last sentence of paragraph	36
8	The possible APs for node 05 attach to PDT	46
9	The problems of multiple interpretations between [01], [02, 06] and [05], [06]	47
10	Paratactic relation and hypotactic relation in RST	49
11	Coordinating and Subordinating relations with nuclearity	50
12	The occurrence of Demonstrative anaphora and Zero anaphora	57
13	System architecture of Thai Text Structuring for Summarization	58
14	Example text for describing the discourse phenomenon (in Thai)	59
15	Example text for describing the discourse phenomenon (in English)	59
16	Local coherent tree construction algorithm	72
17	Global coherent tree construction algorithm	73
18	Local Coherence Tree Updating procedure following by Text-3	75
19	Snapshot, from local coherent tree to global coherent tree.	76

GENERATING DISCOURSE STRUCTURE FOR THAI TEXT SUMMARIZATION

INTRODUCTION

Introduction to Text Summarization

Nowadays, the Internet has become a source for all sorts of information. With more than 29.7 billion pages on 108 million websites (Netcraft's report, Feb-2007), people can get just about any information they desire by using web search engine. However, information obtained from traditional search engine is becoming more voluminous, fragmented into different formats, and duplicated in multiple physical locations. These make it extremely difficult to evaluate the usefulness and correctness of the returned information, and, thus, create an obstruction to promptly access the required information which is very important in the current Information Age. Text summarization research is one of the most effective methods that can be utilized to reduce the obstruction stated above.

Automatic text summarization (TS) is a process of generating a summary that contains important concepts and sentences of an original document. It is also a process that results in a decrease of the document length (Radev, 2000). Normally the text summarization process can be decomposed into three phases.

Firstly, the input document is analyzed to construct a representation that is suitable for further processing. In 1997, D. Marcu built a discourse parser program that constructs a tree-based representation of contents of the document based on Rhetorical Structure Theory (RST). Using RST, terminal nodes in a representation tree are corresponding to propositions in the document, and non-terminal nodes represents contiguous text spans where the text spans of its children are joined by some discourse relations. A linked list structure, such as cohesion chains (Morris and Hirst, 1991) or lexical chains (Barziley, 1997), has also been utilized to represent contents of the document. In this representation, a concept in the documents is

represented by a sequence of related words. This sequence is independent of the grammatical structure of the text and in effect it is a list of words that captures a portion of the cohesive structure of the document. The text summarization methods that utilize this representation usually start from a set of words in the title of the document. It then constructs lexical chains, consisting of words that have similar meaning or are related to the title, to represent the contents of the document. For example, in an agricultural paper about pests, some related terms for this topic would be insect, leafhopper, carrier, and pest. The text summarization method will consider the sentences that contains these terms as a potential sentence to contain important information than sentences that do not.

The second phase is responsible for calculating scores for sentences considered to be important by the first step. These scores will be used to select important sentence for generating the summary in the next step.

Finally, the last phase is responsible for synthesizing a summary output (Elhadad, 1992; Knight and Hatzivassiloglou, 1995; Mckeown and Radev, 1995; Hovy and Wanner, 1996; Kan and McKeown, 1999). The process of this step is depended on the summarization approach whether it is extractive summarization (in which only the key informative segments of the documents are identified) or abstractive summarization (in which a coherent high level text is generated to describe the important information of the document). The output of extractive summarization will consist entirely of material copied from the input document, while the output of abstractive summarization will contain at least some of material that are not presented in the input.

Moreover, text summarization techniques can also be classified into two approaches: *statistical–based approach* and *knowledge-based approach*.

Based on the hypothesis that a word frequently occurred in the document is an important word, traditional statistical-based approach (Luhn, 1958) uses term frequencies as a significant measurement. Consequently, document features are added into the scoring formula in order to calculate the important of each sentence. For

example, Edmundson (1968) used cue phrase, keyword, title, and location of the sentence to calculate weighting scores of sentences that contain these document features. After that, Kupiec (1995) utilized Bayesian classifier to learn the weighting function and introduced additional document features such as sentence length cutoff, and uppercase word. In 1997, C.Y. Lin & Hovy used Optimum Position Policy (OPP) to give a different score to each sentence based on its position in the paragraph. For Thai text summarization research, Canasai and Chuleerat (2003) represented a document at paragraph level by a document vector that contains information about term frequency and inversed document frequency as well as other document features.

The knowledge-based approach to text summarization usually utilizes Natural Language Processing (NLP) technique to identify important contents of the document. For example, word-level semantic relation such as lexical chains methods (Morris and Hirst, 1991; Barzilay, 1997; Silber, 2000; Brunn, 2002; Alonso, 2003) and discourse processing (Ono, 1994; Marcu, 1997 – 2003) can be utilized to locate the position of an important text in the document.

The performance study of these two different approaches revealed that statistical-based approaches have an average precision between 20% and 52%, while an average precision of knowledge-based approaches is between 46% and 88%. This variation was depended on the domain and type of the documents. The main reason why the precision of both approaches are significantly different is that text-level processing requires relations to describe the continuation of an important text together with an explanation of meaning to explain whether the text is main clause or sub clause. Thus, statistical-based approaches that do not consider these issues are usually have low performance than knowledge-based approaches that bring these issues into consideration. For knowledge-based approaches, the researches that utilize lexical cohesion techniques usually have less precision than researches that utilize discourse processing techniques since lexical-cohesion techniques do not consider the main clause and sub clause feature of the sentence.

Text summarization techniques can also be classified into two approaches based on summarization result as *extractive summarization* (Mani and Maybury, 2001), which only the key informative segments of the documents are identified and extracted to generate a summary, and *abstractive summarization* (Mckeown and Radev, 1995; Kan and Mckeown, 1999), which generates high level text summary to describe the important information of the original document. The tendency survey in this research revealed that most of the methods proposed in the literature trend to generate an extractive summarization rather than abstractive summarization.

The problems and complexities of text summarization are also depended on the document dimension which can be classified as *single-document* and *multiple-document*. In addition, we can also view the perspective of text summarization by author's purpose as: *indicative summarization* and *informative summarization*. The main purpose of an indication summary is to suggest the contents of the article without giving away detail on the content. It is usually used to convince the reader to retrieve the full information from the original source. The examples of such summary are book jackets, card catalog entries and movie trailers. On the other hand, the main purpose of an informative summary is to summarize (an often replace) the original document. Thus, it must contain all pertinent information necessary to convey the core information and omit assistant information. In conclusion, text summarization techniques can be classified in several dimensions as showed in Table 1.

Table 1 An approach in summarization research

Summarization criteria	**Dimension**	
Input	Single-document	Multiple-document
Purpose	Indicative Summarization	Informative Summarization
Output	Extractive Summarization	Abstractive Summarization

Source: Kan and McKeown (2003)

Motivation

Text Summarization has benefits in many dimensions. Not only it will contribute to natural language processing community, it also contribute to the progress of text understanding whose target is to develop a program that can simulate human thinking behavior in the sense that the program should be able to decide which parts of the document are important components that the author want to communicate with their readers and which parts are not. Beside the above reason, automatic text summarization also provides its users with several benefits including:

1. Text summarization can be used as a tool to reduce the time required for reading Thai document by reducing the size of the document so that it is harmonized with readers' time constraints.

2. Text summarization can be used to increase efficiency of other researches such as information retrieval by reducing the time to choose documents/information from search engine results which usually contain an excess amount of duplicated information, some of which are not matched with users' requirement.

3. Text summarization can be used to solve the limitation of information presentation on small communication devices such as PDA (Orkut, 2001; Buyukkokten, 2001) and mobile phone (Simon, 2001; Yang and Wang, 2001, 2003) by creating a summary that is fit and short enough to present on such devices.

4. Text summarization can be used to reduce a task of machine translation by reducing the size of the document. By translating a summarized document instead of an original document, the ruining time of machine translation is significantly reduced (Nguyen, 2004).

This research also still maintains all above benefits for Thai text summarization by using the way to generate the good representation of texts. Discourse structure as the tree-like was considered to construct it.

Crucial problems in Thai language for text summarization

In general, text summarization has three main steps: *pre-processing* for making the unstructured text suitable for further processing by using natural language processing techniques (e.g. POS tagging, sentence segmentation and anaphora resolution), salience extraction and summary document generation. The detail of each step will be explained in the subsequence section.

With all the above considerations, there are two main problems in Thai text summarization which are:

1. The basic problems that appear with a specific feature of language

1.1 Thai language does not have special delimiters to indicate word and sentence boundaries. These make Thai language more difficult to analyze than English language that has delimiters such as space, period, comma, and semicolon. Consequently, these create boundary determination problem. If the determined boundary is too small or too big, it will be more difficult to control the length of the summary result.

1.2 Thai language does not have linguistic information to indicate the different of timer. This makes the analysis of relations that are related to time and order more complicate than some languages (e.g. English) that had linguistic information to indicate the different of time. Thus, a discourse summarization process proposed in this research requires an information about text order and information of time for determining whether the considered text are main clause or sub clause.

1.3 Thai language has a flexible linguistic structure in the sense that the positions of subject and object in the sentence are interchangeable without altering the meaning of the sentence. This creates ambiguities both in structure and semantic level, and more complex linguistic rules are required to correctly analyze Thai sentence. Since text summarization process requires sentence-level linguistic

information for analyzing the focus of the sentence, Thai text summarization is more complicate than other language that has less complicate language structure.

1.4 Like other languages, Thai language has an ellipsis and reference problem. However, zero anaphora occurs very frequently in Thai language. In a corpus of 2850 sentences, nearly 30% of the sentence contains zero anaphora. This creates a lot of difficulty in noun phrases analysis, and requires linguistic information to solve the problem. This is a crucial problem that needs to be solved before we can identify the focus of the text.

2. Salience extracting problem

Study of previous researchers discovered that extractive text summarization techniques that utilize statistical-based approaches have several disadvantages. Firstly, the summarization result would not be coherent since the selection process does not consider the coherence between texts. Secondly, the summarization result may also contain duplicated information. In addition, the average precision of such approaches was less than 55%.

Therefore, this thesis selected to use a knowledge–based approach by utilizing Discourse Theory RST as a main concept for describing a relation between texts before designing which parts of the text are crucial to improve the coherence of the result. However, there are several problems in applying RST with text summarization:

1. Problems in identifying relation between text units which could be classified into two levels as:

1.1 Ambiguity of Discourse Marker (DM) that represent the relation between text such as [โรคนี้พบได้เกือบทุกระยะการเจริญเติบโต]$_A$ [แต่พบมากในระยะกะหล่ำปลีห่อหัว]$_B$. In text B, the word “แต่” is a discourse marker between two sentences, text A and B, that has two semantic relations: contrast and elaboration relation.

1.2 Some sentences may not contain a discourse marker. Since most discourse processing techniques utilize discourse markers as a main factor to identify semantic relation between texts, it will not be able to identify the relation when discourse markers are missing. For example, in sentence [บนแผลมีเส้นใยเชื้อรางอกออกมา]$_A$ [บริเวณปลายเส้นใยมีสีดำ]$_B$ [มองเห็นได้ชัดเจน]$_C$, text A, text B and text C do not have a word or phrase for conjunction. This problem makes it difficult to analyze the relation between texts.

2. Problem in matching position of constituent

Constituent is the relationship between lexicon units, which are parts of a larger unit. Constituency is usually shown by a tree diagram or by square brackets:
Ex. [บนแผลมีเส้นใยเชื้อรางอกออกมา] $_A$ [บริเวณปลายเส้นใยมีสีดำ] $_B$ [มองเห็นได้ชัดเจน] $_C$

Constituent acts as a chunk that can be moved together and it often occurs in Thai language. In each constituent can be moved position. For example, [บนแผลมีเส้นใยเชื้อรางอกออกมา]$_A$ [บริเวณปลายเส้นใยมีสีดำ]$_B$ [มองเห็นได้ชัดเจน]$_C$, Constituent C can be match the position with constituent B in A [B, C] or constituent A and B in [A, B] C.

Contributions

This research makes four contributions to the fields of text summarization and discourse processing for Thai language:

1. Provide the Thai discourse markers in the domain of agriculture. (see in Appendix A)

This resource provides knowledge about discourse relations. We investigate discourse markers in the agriculture domain. 126 discourse markers were identified and classified into two categories which are rhetorical relations and coordinating relations/subordinating relations.

2. Provide a corpus annotated by discourse tree-structure and its alignment. (see http://naist.cpe.ku.ac.th/~ts)

This resource contains 200 documents in agriculture domain that were annotated at multiple levels i.e., Elementary Discourse Unit (EDU), discourse relationship, anaphoric expression and discourse markers. It is useful for studying the phenomena of the occurrence of the discourse structure.

3. Provide the practical methodology for discourse structure construction.

This thesis proposes a methodology for discourse construction from text corpus by using DDF (Discourse Dependency Function) together with Right Frontier Constraint (Polanyi, 1988) to identify suitable attachment points by treating the context that has the highest relevance scores as a winner. The COR&SUBR relationship was computed based on the lexical and co-occurrence features.

Thesis Organization

The rest of the document is organized as follows:

Literature Review, this thesis presents the introduction to text summarization, and describes the necessity background knowledge in text summarization which consists of discourse processing, text summarization from computational perspectives, related theories, and related works. Furthermore, this section will be focus on the crucial problems of how to generate text representation for producing the quality result of TS.

Materials and Methods describes the materials used in this study and methodologies for structuring Thai texts. It shows how to construct Thai text coherent from Text Corpora by using Discourse Dependency Function together with Right Frontier Constraint (Polanyi, 1988) in Attachment Point Identification step. Moreover, two relations; coordinating relations (COR) and subordinating relation (SUBR) instead of the number of rhetorical relations are also proposed (see Appendix A).

Finally, the algorithm for constructing the discourse coherent tree based on linguistic knowledge and Information Retrieval (IR) techniques is described.

Result and Discussion, empirical evidence is provided to support the theoretical proposal in section of Literature Review and Materials and Methods. We present experiments with human judges and experiments with automatic procedures. Moreover, we also present a set of experiments of discourse tree structuring in our proposed model. We have measuring with five parts: Attachment Point Identification, COR&SUBR relations interpretation, Discourse tree structure, Saliencies and Coherence in summary.

Conclusion and Recommendation, this thesis is conclude by summarizing its contributions and sketching lines of research that are left for future work.

OBJECTIVES

The final goal of this thesis is to research and develop the representation of discourse that improves text summarization and can be obtained by discourse shallow parsing. This general goal can be divided in the following research objectives:

1. Study and develop the system for Thai text summarization.

2. Study an approach in order to generate a summary text which can be obtained with NLP knowledge resources: word cohesion, discourse markers, anaphoric expression, to construct the organization of texts which was described the pieces of text relationship with two relations: Coordinating relation (COR) and Subordinating relation (SUBR).

3. Develop methodology for disambiguating COR and SUBR

4. Develop the models for generating discourse structure tree to enhance the summary efficient.

LITERATURE REVIEW

Introduction to Text Summarization

In this information era, with the rapid growth of Internet and on-line information services, a huge quantity of information is available and accessible on-line. This explosion of information has resulted in a well-recognized information overload problem. There is no time to read everything and yet we have to make critical decisions based on whatever information is available. Automatic text summarization is recognized as one of the solutions needed to tackle the problem of overwhelming amounts of available information. However, research on automatic text summarization still requires further development. Recent studies focus on the identification of the most relevant information in the text, its transfer to the summary. There are two main approaches for procedure of the statistics-based approach and the knowledge-based approach. The statistics-based approach (Edmundson, 1969; Hovy, 1997) often gives incoherent results, which cause further misunderstandings by human. By contrast, the knowledge-based approach takes this problem into consideration and tries to extract the most salient parts from a structured text representation. Text representation is a tree containing discourse relations. The creation of text representation requires a strong knowledge base (Marcu, 1997; Cristea, 2005) that is often justified by the assumption that salience extraction cannot be achieved without a linguistically sophisticated and detailed representation.

About corpus in this research, they are in agriculture domain. Text examples have been translated from Thai into English. Thai texts are slightly different from English texts in some aspects at discourse level. For example, zero anaphora occurs frequently in Thai texts, and disappeared in English written texts. That has an effect on model for generating discourse structure tree, which is slightly more complicated than a model for English texts.

In the next section, the necessary background knowledge and the related works that are required for constructing the discourse structure from textual data is described. Finally, the problems in construction of coherence tree were illustrated as prior knowledge of reader that before walking through this research details in the next chapter. The background knowledge is classified in three sections; text summarization from a computational perspective, related theories and related work. Furthermore, problems in construction of the coherence tree are described at the end of this part.

Background Knowledge

This thesis concerns many factors that involve formulating the computational model of text summarization on the first. Present incoherent text is the cause of the problems, which are the summary result of TS. Discourses processing and discourse analysis in NLP are the solution to the present incoherent text problem.

Text summarization from a computational perspective

There have been many efforts devoted to analyze the problem of TS and systematize the process (Spärck-Jones, 1993, 1997, 1999; Hovy and Marcu, 1998; Mani and Maybury, 1999; Radev, 2000; Hahn and Mani, 2000; Hovy, 2001), here we will describe the aspects that most have generally considered essential for a good understanding of TS concern. From a computational perspective, “A summary is a reductive transformation of a source text into a summary text by extraction or generation” (Spärck-Jones, 2001). The problem of summarization has traditionally been decomposed into three phases:

analyzing the input text to obtain a representation,

transforming it into a summary representation,

synthesizing an appropriate output form to generate the summary text.

Many the early researches on TS have been devoted to the analysis of the source text, probably for two main reasons: first, a summary that fulfills an information need can be built by simple concatenation of literal fragments of the source text. Second, the generation of natural language expressions automatically requires huge NLP resources, which are far beyond the capabilities of most of the current research groups in the area even today. As it is, the analysis of the source text can currently be considered as the main factor influencing the quality of the resulting summary. Indeed, once the crucial aspects of texts have been identified and characterized, virtually all heuristics for selection of relevant information will perform well.

Besides the internal steps for building a summary, many contextual factors affect the process of summarization, mostly concerning the kind and number of documents to be summarized, the medium of communication, the expected format of the summary, the intended audience, etc. Effective summarizing requires an explicit and detailed analysis of context factors, since summaries are configured by the information need they have to fulfill. Spärck-Jones (1999a) distinguishes three main aspects of summaries: input, purpose and output, which we develop in what follows.

1. Input Aspects

The features of the text to be summarized crucially determine the way a summary can be obtained. The following aspects of input are relevant to the task of TS:

Document Structure Besides textual content, heterogeneous documental information can be found in a source document, for example, labels that mark headers, chapters, sections, lists etc. If it is well systematized and exploited, this information can be of use to analyze the document. For example, Kan (Kan, 2003) exploits the organization of medical articles in sections to build a tree-like representation of the source. Teufel and Moens (2002b) systematize the structural properties of scientific articles to assess the contribution of each textual segment to the article, in order to build a summary from that enriched perspective.

However, it can also be the case that the information it provides is not the target of the analysis. In this case, document structure has to be removed in order to isolate the textual component of the document.

Domain. Domain-sensitive systems are only capable of obtaining summaries of texts that belong to a pre-determined domain, with varying degrees of portability. The restriction to a certain domain is usually compensated by the fact that specialized systems can apply knowledge-intensive techniques which are only feasible in controlled domains, as is the case of the multi-document summarizer SUMMONS (McKeown and Radev, 1995), specialized in summaries in the terrorism, domain applying complex Information Extraction techniques. In contrast, general purpose systems are not dependent on information about domains, which usually results in a more shallow approach to the analysis of the input documents. Nevertheless, some general purpose systems are prepared to exploit domain-specific information.

Specialization level A text may be broadly characterized as ordinary, specialized, or restricted, in relation to the presumed subject knowledge of the source text readers. This aspect can be considered the same as the domain aspect discussed above.

Restriction on the language The language of the input can be the general language or restricted to a sublanguage within a domain, purpose or audience. It may be necessary to preserve the sublanguage in the summary.

Scale Different summarizing strategies have to be adopted to handle different text lengths. Indeed, the analysis of the input text can be performed at different granularities, for example, in determining meaning units. In the case of news articles, sentences or even clauses are usually considered to be minimal meaning units, whereas for longer documents, like reports or books, paragraphs seem more adequate units of meaning. Also the techniques for segmenting the input text in these meaning units differ: for shorter texts, orthography and syntax, even discourse

boundaries; (Marcu, 1997a) indicate significant boundaries, for longer texts, topic segmentation (Kozima, 1993; Hearst, 1994) is more usual.

Media Although the main focus of summarization is textual summarization, summaries of non-textual documents, like videos, meeting records, images or tables have also been undertaken in recent years. The complexity of multimedia summarization has prevented the development of wide coverage systems, which means that most summarization systems that can handle multimedia information are limited to specific domains or textual genres (Maybury and Merlino, 1997). However, research efforts also consider the integration of information of different media (Benitez and Chang, 2002), which allow a wider coverage of multimedia summarization systems by exploiting different kinds of documentary information collaboratively, like metadata associated to video records (Wactlar, 2001).

Genre Some systems exploit typical genre-determined characteristics of texts, such as the pyramidal organization of newspaper articles, or the argumentative development of scientific articles. Some summarizers are independent of the type of document to be summarized, while others are specialized to some type of documents: healthcare reports (Elhadad and McKeown, 2001), medical articles (Kan, 2003), agency news (McKeown and Radev, 1995), broadcast fragments (Hauptmann and Witbrock, 1997), meeting recordings (Zechner, 2001), e-mails (Muresan *et al.*, 2001; Alonso *et al.*, 2003a), web pages (Radev *et al.*, 2001), etc.

Dimension The input to the summarization process can be a single document or multiple documents, either simple text or multimedia information such as imagery audio, or video (Sundaram, 2002). Language Systems can be language-independent, exploiting characteristics of documents that hold cross-linguistically (Radev *et al.*, 2003; Pardo *et al.*, 2003), or else their architecture can be determined by the features of a concrete language. This means that some adaptations must be carried out in the system to deal with different languages. As an additional improvement, some multi-document systems are able to deal simultaneously with documents in different languages (Chen, 2002; Chen *et al*., 2003).

2. Purpose Aspects

Situation TS systems can perform general summarization or else they can be embedded in larger systems, as an intermediate step for another NLP task, like Machine Translation, Information Retrieval or Question Answering. As the field evolves, more and more efforts are devoted to task-driven summarization, in detriment of a more general approach to TS. This is due to the fact that underspecification of the information needs poses a major problem for design and evaluation of the systems. Evaluation is a major problem in TS. Task-driven summarization presents the advantage that systems can be evaluated with respect to the improvement they introduce in the final task they are applied to.

Audience In case a user profile is accessible, summaries can be adapted to the needs of specific users, for example, the user's prior knowledge on a determined subject. Background summaries assume that the reader's prior knowledge is poor, and so extensive information is supplied, while just-the-news are those kind of summaries conveying only the newest information on an already known subject. Briefings are a particular case of the latter, since they collect representative information from a set of related documents.

Usage Summaries can be sensitive to determined uses: retrieving source text (Kan *et al.*, 2001), previewing a text (Leuski *et al.*, 2003), and refreshing the memory of an already read text.

3. Output Aspects

Content A summary may try to represent all relevant features of a source text or it may focus on some specific ones, which can be determined by queries, subjects, etc.

Generic Summaries are text-driven, while user-focused (or query-driven) ones rely on a specification of the user's information need, like a question or keywords. Related to the kind of content that is to be extracted, different

computational approaches are applied. The two basic approaches are top-down, using information extraction techniques, and bottom-up, more similar to information retrieval procedures. Top-down is used in query-driven summaries, when criteria of interest are encoded as a search specification, and this specification is used by the system to filter or analyze text portions. The strategies applied in this approach are similar to those of Question Answering. On the other hand, bottom-up is used in text-driven summaries, when generic importance metrics are encoded as strategies, which are then applied to a representation of the whole text.

Format. The output of a summarization system can be plain text, or else it can be formatted. Formatting can be targeted to many purposes: conforming to a pre-determined style (tags, organization in fields), improving readability (division in sections, highlighting), etc.

Style. A summary can be *informative*, if it covers the topics in the source text; *indicative*, if it provides a brief survey of the topics addressed in the original; *aggregative*, if it supplies information non present in the source text that completes some of its information or elicits some hidden information (Teufel and Moens, 2002b); or critical, if it provides an additional evaluation appreciation of the summarized text.

Production Process. The resulting summary text can be an *extract*, if it is composed by literal fragments of text, or an *abstract*, if it is generated. The type of summary output desired can be relatively polished, for example, if the input text is well-formed and connected, or else more fragmentary in nature (e.g., a list of keywords). There are intermediate options, mostly concerning the nature of the fragments that compose extracts, which can range from topic-like passages, paragraph or multi-paragraph long, to clauses or even phrases. In addition, some approaches perform editing operations in the summary, overcoming the incoherence and redundancy often found in extracts, but at the same time avoiding the high cost of a NL generation system. Jing and McKeown (Jing and McKeown, 2000) apply six re-writing strategies to improve the general quality of an extract-based summary by edition operations like deletion, completion or substitution of clausal constituents.

Surrogation. Summaries can stand in place of the source as a surrogate, or they can be linked to the source (Kan *et al.*, 2001; Leuski *et al.*, 2003), or even be presented in the context of the source (e.g., by highlighting source text, (Lehmam and Bouvet, 2001)).

Length. The targeted length of the summary crucially affects the informativeness of the final result. This length can be determined by a compression rate, that is to say, the ratio of the summary length with respect to the length of the original text. Traditionally, compression rates range from 1% to 30%, with 10% as a preferred rate for article summarization. In the case of multi-document summarization though, length cannot be determined as a ratio to the original text(s), so the summary always conforms to a pre-determined length. Summary length can also be determined by the physical context where the summary is to be displayed. For example, in the case of delivery of news of summaries to hand-helds (Boguraev *et al.*, 2001; Buyukkokten *et al.*, 2001; Corston-Oliver, 2001), the size of the screen imposes severe restrictions to the length of the summary. Headline generation is another application where the length of summaries is clearly determined (Witbrock and Mittal 1999). In very short summaries, coherence is usually sacrificed to informativeness, so lists of words are considered acceptable (Kraaij *et al.*, 2002; Zajic *et al.*, 2002).

From the perspective of TS, input aspect, purpose aspect and output aspect were considered as the influence factors for processing in TS. There are many researches involves in one or more factors to work on TS research such as salience identification, salience extraction and generating the discourse structure as the surrogate of the source text to maintain the coherence relation in summary result. This thesis also pays attention to solve the last one. In order to light up our work, we will describe the principle of the related theories; Rhetorical Structure Theory (RST), Right Frontier Constraint (RFC) and naïve bayes learning in the next section.

Related Theories

This section is described about theories of Rhetorical Structure Theory (RST), Rhetorical relations Right Frontier Constraint (RFC) , Naïve Bayes Classifier and Learning Techniques Vector Space Model (VSM).

1. Rhetorical Structure Theory (RST)

RST was developed during the 1980s by researchers in natural language generation, many of whom were then involved with projects at the Information Sciences Institute in Southern California. Since much of my research is informed by the theoretical approach taken within RST, it is first necessary to outline the theory, criticisms of it, and the modifications which we have made to adapt it to my purposes.

RST (Mann and Thompson, 1986, 1988) models the discourse structure of a text by means of a hierarchical tree diagram. The terminal nodes of RST tree are propositions encoded in text. (Although RST analysts usually take care to distinguish contiguous stretches of text, termed text spans, from the propositions expressed in the text, in the discussion below text spans are simply referred.) Non-terminal nodes represent contiguous text spans, whose daughter spans are joined by discourse relations. These discourse relations are divided into two kinds: symmetric and asymmetric. For the example of the RST tree, The example applied Text -1 with RST relations into RST tree (see in Figure 1) where text unit [05] is a cause of result in text unit [06] which denoted by CR[05, 06] where CR is Cause-Result relation, EB is Elaboration relation, JT is Joint relation and CSQ is consequence relation and SUM is summary relation.

Text-1:

[Blast disease can commonly spread to every parts of Thailand]$_{01}$ [this disease caused by a fungus called Pyricularia]$_{02}$ [which the conidia of this fungus can be blown by the wind] $_{03}$ [therefore blast disease distributes its through the wind] $_{04}$ [when the conidia of the fungus settle on various parts, rice that are highly moist]$_{05}$ [it will sprout in a fiber form, destroying the plant] $_{06}$

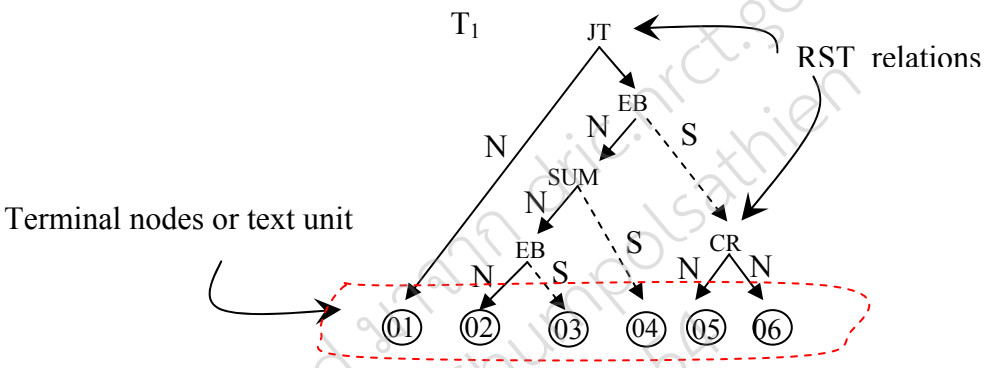


Figure 1 Sample of RST tree from Text-2

Figure 1, a symmetric relation involves two or more text spans, each of which is equally important in realizing the writer's goals. We denoted it with straight line such as the JT relation consist of two straight lines. By convention, each of these text spans is labeled a nucleus. Straight lines are used to represent the connection between the child nodes of a symmetric relation to their parent node. An asymmetric relation involves exactly two text spans. One text span, the nucleus (N), is more important in realizing the writer's goals. The other text span, the satellite (S), is in a dependency relation to the nucleus, modifying it in ways specified in the definition of the particular relation. In the same figure, illustrates one kind of asymmetric relation, the ELABORATION relation. An EB labeled is used to represent the connection between the satellite and the nucleus. The satellite denoted by dash line and nucleus denote by straight.

Although units as large as paragraphs, sections or chapters may be used as terminal nodes for a coarse-grained analysis, the terminal nodes of an RST tree are usually clauses with "independent functional integrity" (Mann and Thompson 1988: 248). Restrictive relative clauses, which by definition serve to modify a head noun and are therefore not directly in significant discourse relations to other clauses, do not qualify as minimal textual spans under this criterion. (We have also chosen to disregard non-restrictive relative clauses on the grounds that they also serve to modify a head noun.) Similarly, clausal subjects and complements do not qualify as terminal nodes.

2. Rhetorical relations

Although there is widespread acceptance by advocates of RST and advocates of other theories of discourse (among them, Ballard, 1971; Grimes, 1975; Halliday and Hasan, 1976; Longacre, 1976; Hobbs, 1979) that relations of the type proposed by RST are useful for describing the structure of discourse. Almost the researcher in RST analysis have a questions arise how many relations are there?

In answer to the question "How many relations are there?" Hovy (1990) identifies a total of approximately 350 relations which have been posited in the linguistics, philosophy, and artificial intelligence literature. Within RST, for example, Mann and Thompson (1986) proposed fifteen relations and twenty-four relation in 1988 which can be classified into subject matter (e.g., Elaboration, Circumstance, Solution hood, Cause, and Restatement) and presentational relations (Motivation, Background, Justify, Concession). The classification is based on the effect intended: in subject matter relations the text producer intends for the reader to recognize the relation; in presentational relations the intended effect is to increase some inclination on the part of the reader (positive regard, belief, or acceptance of the nucleus). However, they also classify the relations into two categories, by nuclearity; paratactic relation and hypotactic relations. Marcu (1997) propose seventy-six relations in fifteen categories, but he still classify based on effect intention of reader.

Hovy (1990) distinguishes a Parsimonious Position, advocated by Grosz and Sidner (1986) in their work on Centering and Focusing, which posits two very basic relations, Dominance and Satisfaction-Precedence. These two relations are claimed to be sufficient for describing speaker intentions in discourse. Indeed, Grosz and Sidner (1986) claim that it is futile to try to identify a larger finite set of relations, since closer inspection always reveals increasingly subtle semantic nuances.

In summary, the number of RST relations depends on the criteria of RST analyst to determine for their purpose. Therefore, it's no matter to find out an answer to the question "how many relations are there?"

3. Right Frontier Constraint (RFC)

The Right Frontier Constraint, originally proposed by Polanyi (1988) is mentioned in discourse processing with two settings: as an attachment constraint in an incremental discourse development of the tree structure (Cristea and Webber, 1997), and as a referential constraint defining the regions of the discourse model taken to be in focus, therefore introducing discourse entities which are recoverable by referential expressions contained in the last mentioned discourse unit (Polanyi, 1988; Webber, 1991).

The right frontier (RF) of a tree is the sequence of nodes that starts in the root of the tree and continues with all nodes placed in the right extreme of the tree at any level, so-called "right frontier area". The terminal frontier (TF) is the sequence of nodes, counted left-to-right, on the lowest levels of the tree (therefore which do not have any daughters).

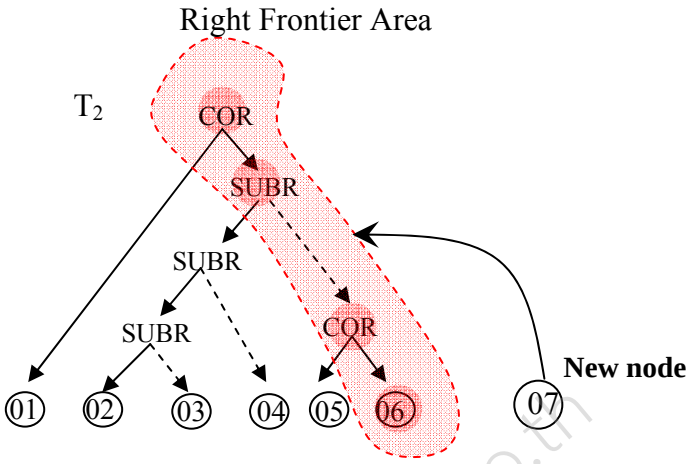


Figure 2 Discourse Update, attach new node into Right Frontier Area

Figure 2, discourse updating a new node 07 to discourse tree T_2, RFC has mention to attach a new node 07 at a some point in right frontier area. As the same way, to identify an antecedent of anaphor in the new node, RFC has recommended to locate its antecedent in the right frontier area.

However, the RFC are not complete in some phenomenon of discourse updating such as discourse interruption and flashback, the new node maybe suitable attaches at a node that posit out of the right frontier area. This thesis pay attention on this problem, and propose the solution in term Adaptive Right Frontier (ARF).

4. Naïve Bayes Classifier, Learning Techniques

The availability of reliable learning systems is one of strategic importance, as there are many tasks that cannot be solved by classical programming techniques since no mathematical model of the problem is available. Then, it is necessary to apply the learning techniques, e.g. Naïve Bayes Classifier, Support Vector Machine, ID3, and etc., for determining the model or rules of solving the tasks. Additionally, this work applied Naïve Bayes Classifier for determining the discourse cues (or discourse markers), word co-occurrence and semantic noun phrases for determining

the discourse relations; coordinating relation and subordinating relations, as described by the following.

The Naïve Bayes (NB) classifier or the NB learner is the highly practical Bayesian learning method based on probabilities together with observed data. According to (Mitchell T.M., 1997), the NB classifier applies to learning tasks where each instance x is described by a conjunction of attribute values and where the target function $f(\mathbf{x})$ can take on any class value, v, from some class finite set V. A set of training examples of the target function is provided, and a new instance is presented, described by the tuple of attribute values$<a_1, a_2,..a_n>$. The learner is asked to predict the target value, or classification, for this new instance.

The Bayesian approach to classify the new instance is to assign the most probable target value, v_{MAP} called $v_{maximum\ a\ posterior}$, given the attribute values $<a_1, a_2,..a_n>$ that describe the instance (see equation (1) which can be derived to equation(2)).

$$v_{MAP} = \arg\max_{v_j \in V} P(v_j \mid a_1, a_2 ... a_n) \quad (1)$$

$$v_{MAP} = \arg\max_{v_j \in V} P(a_1, a_2 ... a_n \mid v_j) P(v_j) \quad (2)$$

The NB classifier is based on the simplifying assumption that the attributes values are conditionally independent given the target value. In other words, the assumption is that given the target value of the instance, the probability of observing the conjunction $a_1, a_2,..a_n$ is just the product of the probabilities for the individual attributes. Then, the NB classifier, v_{NB} , is shown as in the following.

$$v_{NB} = \arg\max_{v_j \in V} P(v_j) \prod_i P(a_i \mid v_j) \quad (3)$$

In our research, we applied NB classifier from equation (3) for learning to classify the annotated the pair of elementary discourse unit as the coordinating relation (COR) or subordinating relation (SUBR) by having V as the finite set of discourse relation classes as {*COR, SUBR*} (see discourse relation interpretation in the MATERIAL AND METHOD section). The attributes from a_1, a_2,.. to a_n are linguistic features which consist of a discourse marker feature, word co-occurrence feature and semantic noun phrase feature.

5. Vector Space Model (VSM)

Many current researches in automatic summarization, including this research use the Vector Space Model (Salton and McGill, 1983) to measure, or at least approximate the measurement similarity of semantic content. In the original IR model, a set of documents is conceptualized as a two-dimensional co-occurrence matrix, where the columns represent the documents and the rows represent the unique terms (usually words or short phrases) occurring in the documents. Sometimes every term appearing in the source document will be represented by a row, though it is more common to exclude a stop list of prepositions, function words, and other lexemes with negligible semantic content. The value in a particular cell may be a simple binary 1 or 0 (indicating the presence or absence of the term in the document) or a natural number indicating the frequency with which the term occurs in the document. Typically, each cell value is adjusted with an information-theoretic transformation. Such transformations, widely used in IR (e.g., Spärck-Jones, 1972), weight terms so that they more properly reflect their importance within the document. For example, one popular measure known as TF.IDF (term frequency–inverse document frequency) uses the following formula:

$$w_{ij} = tf_{ij} \log_2 (N/n_i)$$

Here w_{ij} is the weight of term i in document j, tf_{ij} is the frequency of term i in document j, N is the total number of documents, and n_i is the number of documents in which i occurs. After the weighting, pairs of documents can be compared by their

column vectors, using some mathematical measure of vector similarity. Perhaps the most popular measure is the cosine coefficient,

$$\cos (A, B) = \sum_i A_i B_i / |A_i| |B_i|$$

In our work, we use the vector space model to compare the semantic similarity of discourse entity (such as noun phrase) within a single document. In this case, the "documents" of the term–document co-occurrence matrix are actually words or phrases.

Discourse processing

This part focuses on prior knowledge in discourse processing. Present incoherent text is the cause of the problems, which are the summary result of TS. Discourses processing and discourse analysis in NLP are the solution to the present incoherent text problem.

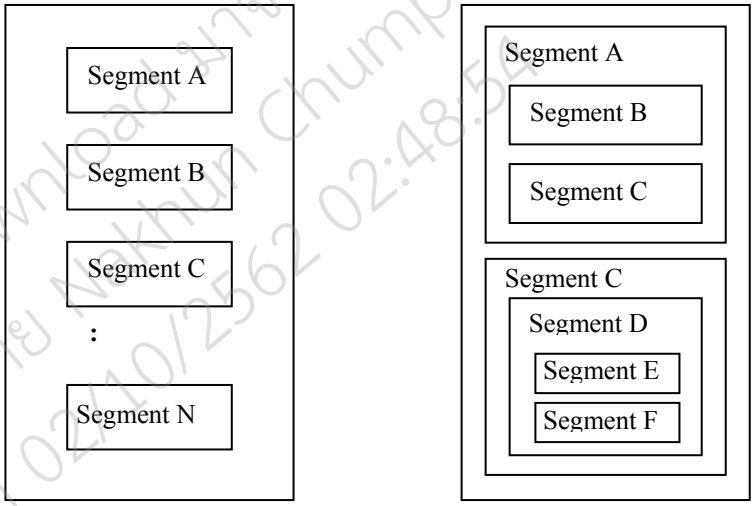
1. Discourse analyzing for extracting an important text

In this way, the knowledge based TS approach in summarization research uses the linguistic knowledge obtained from an analyzing of linguistic behavior as a tool for human communication. This research, the important texts are given to linguistic knowledge usage for document processing at the discourse level. Grosz (1986) proposes a discourse theory by giving the importance in bringing the discourse structure to describe the relation of text. Mann and Thomson (1988) has proposed RST theory by using the rhetorical relation as a tool for describing semantic relations between texts. Both theories would have a purpose to describe a relation of text in meaning. This research will use concepts from both theories to analyze the discourse relation and structure of Thai language in order to help the processing step of text in Thai language during the important text extraction step.

2. Discourse structures

Discourse structure consist of paragraph or embedded discourses (Longacre, 1979 and Somsong, 1984) . Each paragraph could be containing embedded discourse. Each embedded paragraph consists of sentence that harmonized with the Attention State Theory (AST) by B. Grosz. Each segment would describe similar interests (intention). In the same way, the relation of the structure between segments were explained by embedded discourse structures.

The discourse structure can be classified into two structures: linear structure or segments as coordinating relation, and hierarchical structure or segments as subordinating relation (Hobbs, 1990) as shown in Figure 3. The example of text in these structures is shown in Figure 4.



(A) Linear Structural document (B) Hierarchical structure document

Figure 3 Discourse structure

Linear structure: Hobbs described the discourse structure seen in Figure 3(A) as linear structure. This is because each segment {A, B, C …} shares similar structures in text presentation. That is if A, B, C … was independent from each other texts in each segment were organized in the text relation called parallelism. When the

linear structure was applied to an extractive important text with RST, it should select an important text of each segment that had linear structure relation to bring about RST schema in multi–nucleus type that might lead to the conclusion of discourse relation in joint, list, summation relation etc (the example as show in figure 4).

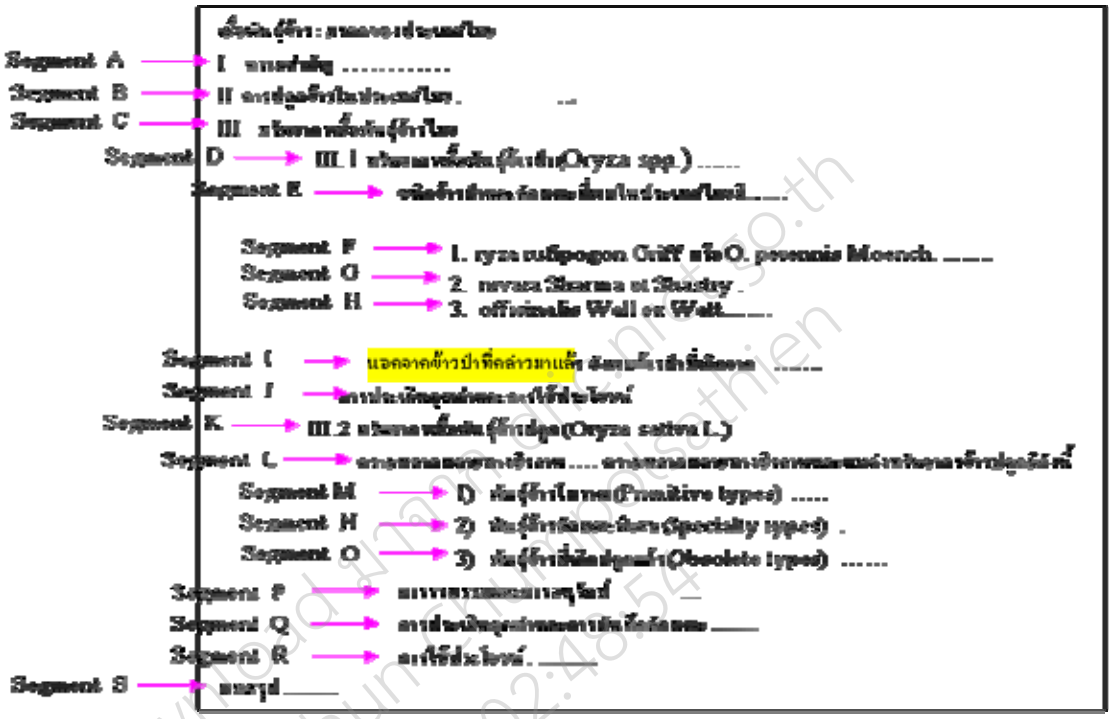


Figure 4 The example of Discourse structure of document

Segments [F, G, H] describe a variety of wild rice discovered in Thailand together with under the list of relation, shown in item in numeric ordering. While the segments [E, K] describe a coordinating relation under the joint relation of the resource of Thai rice variety under segment C. Moreover, it can be said that if the full document has a discourse linear structure, it will be able to produce an important text of full document from the combination of salient each segment (local salient). For example the document describes the rice cultivation that consists of an item of paddling, Fertilizer method, irrigation method, and pest control by each item is independent to each other but it is coordinating to describe rice cultivation. So the summarization is able to use salient of each item and combines to explain as an importance of full document.

A hierarchical structure: Hobbs described the discourse structure in this type as subordinating relation which contained a segment being dominant to the other segment in embedded segment feature or a recursive. For example as shown in Figure 4, a segment D is an embedded segment inside a segment C, a segment E is an embedded segment inside a segment D, and segments [F, G, H] are embedded inside a segment E. This kind of discourse structure will have a dominant segment to give main information. In addition, embedded segment gives additional information to the main segment that might be lead to the conclusion of discourse relation in discourse structure in type of Elaborate, Explanation and Background etc. For example as segment E describes the detail of the resource of Thai rice variety with giving the information about the wild rice variety and the feature that is discovered in Thailand to segment D. With the relation in type Elaboration-Additional. While the segment [F, G, H] show the relation with segment E as the example of wild rice variety and feature discovered in Thailand with the relation type as Elaboration-Example etc.

From the document “เชื้อพันธุ์ข้าว: มรดกของประเทศไทย” in Figure 4 shows in other type as tree structure in Figure 5 that describes to segment D and K has a relation depending on a segment C and a segment [E, I, J] have the relations depending on a segment D. Segment [L, P, Q, R] has a relation depending on a segment K. In the same way, segment [F, G, H] have relations depending on a segment E and segment [M, N, O] have relations depending on segment L where each segment contains a condition relation as an embedded segment to a dominant segment that can be figure in string as [A, B, C [D [E [F, G, H], I, J], K [L [M, N, O], P, Q, R], S].

A concept of discourse structure applied with extraction of an important text can explain the data that is in embedded segment. The additional details to the dominant segment mean that discourse structure can be brought to describe with RST. The text existed in embedded segment will work as satellite by a dominant segment working as nucleus.

For example as segments [M, N, O] shown an example of biological diversity of segment L is refer to the relation between segments [M, N, O] and segment L with Elaboration-Example type. Segment L works as a main data segment existing as nucleus. And segments [M, N, O] are addition example data to segment L with existing as satellite.

Moreover, under the relation in hierarchical structure, it can be identified the relation type with existing a cue phrase (or discourse marker) in paragraph level or text segmented, such as segment I contained a cue phrase "นอกจาก...ที่กล่าวมาแล้ว / moreover… as said" working as conjunction back to the segment E (Flashback). Therefore, segment I gives addition information to segment E by elaborate relation.

A document: "เชื้อพันธุ์ข้าว: มรดกของประเทศไทย"

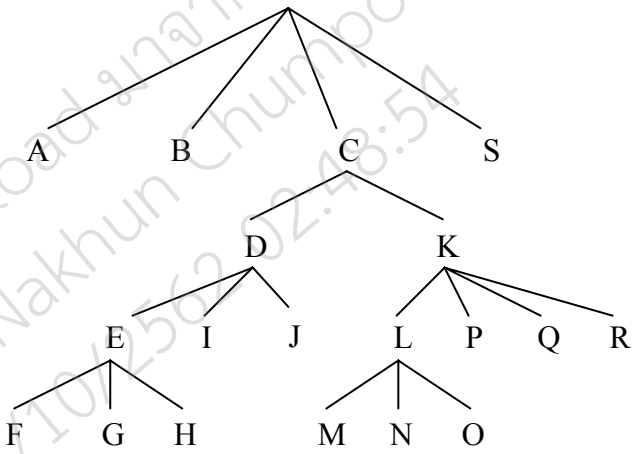


Figure 5 Tree structure of document "เชื้อพันธุ์ข้าว"

Generally, the co-explanation of sentence, subordinate sentence or paragraph might be use a word or phrase to work as a conjunction or preposition. To help to be a continuous text as "และ/and", "หรือ/or", "แต/but", "แม้ว่า/ever", "เพราะว่า/ because", "นอกจากนี้/moreover", "ถ้า..จะ/if …will…" etc. But in the other function of this word or phrase give a semantic relation together which calls a word or phrase in both functions as discourse marker. For example, S8 and S9 in Text-2, **[S**8 เพลี้ยกระโดดสีน้ำตาล

สามารถทำลายได้ทุกระยะการเจริญเติบโตของข้าว] [S9 นอกจากนี้ยังเป็นพาหะนำเชื้อวิสา]. It contains a discourse maker as "นอกจากนี้" to bind S8 to S9 describing an addition data more than the explanation in S8. This discourse maker describes the relation in elaboration-addition type and a pair of text [S20 ถ้าสภาพแวดล้อมเหมาะสม] [S21.ปริมาณเพลี้ยกระโดดสีน้ำตาล จะเพิ่มขึ้นตามอายุข้าว] shown in a pair of text will exist discourse marker as "ถ้า ...จะ..." describing text in condition relation between S20 and S21 where S20 shows the condition and S21 will be happen if it is a real condition etc.

Text 2:

S1: ทั้งตัวอ่อนและตัวเต็มวัยจะดูดกินน้ำเลี้ยงบริเวณโคนต้นข้าวเหนือระดับน้ำ
S2: ในขณะเดียวกันจะคอยขับถ่ายมูลน้ำหวานออกมา
S3: เป็นสาเหตุให้เกิดโรคราดำ
S4: เมื่อมีเพลี้ยกระโดดสีน้ำตาลจำนวนมากดูดกินน้ำเลี้ยงต้นข้าว
S5: จะทำให้ต้นข้าวแสดงอาการใบเหลืองแห้ง คล้ายถูกน้ำร้อนลวก
S6: ซึ่งเรียกว่า "อาการไหม้เป็นหย่อม"
S7: ถ้ารุนแรงมาก ต้นข้าวจะแห้งตาย
S8: เพลี้ยสามารถทำลายได้ทุกระยะการเจริญเติบโตของข้าว
S9: นอกจากนี้ยังเป็นพาหะนำเชื้อวิสา
S10: ซึ่งทำให้เกิดโรคใบหงิกหรือโรคจู๋ มาสู่ต้นข้าวอีกด้วย
S11: โรคนี้เกิดกับต้นข้าวได้ทุกระยะการเจริญเติบโต
S12: ต้นข้าวอายุตั้งแต่ 15-45 วัน
S13: ถ้าได้รับเชื้อโรคจู๋
S14: จะแสดงอาการรุนแรงมาก
S15: ส่วนต้นข้าวอายุเกิน 60 วันไปแล้ว
S16: ได้รับเชื้ออาการจะไม่รุนแรง
S17: ต้นข้าวที่ได้รับเชื้อแล้ว
S18: จะมีอาการต้นเตี้ยแคระแกรน
S19: และไม่ออกรวงหรือออกรวงน้อย
S20: ถ้าสภาพแวดล้อมเหมาะสม
S21: ปริมาณเพลี้ยกระโดดสีน้ำตาลจะเพิ่มขึ้นตามอายุข้าว

S22: จากระยะกล้าถึงระยะออกรวง
S23: ซึ่งในระยะตั้งท้องและออกรวงมักจะพบประชากรเพลี้ยกระโดดสีน้ำตาลสูงที่สุด
S24: และอาการใบไหม้มักจะเกิดในระยะนี้

3. Analyzing discourse structure in paragraph.

In each item, the author can describe data under the item such that it may consist of data more than one paragraph. In each paragraph will be a relation, which is rhetorical relation paragraph, would be any relation depending on the author who needs to use for the ending paragraph or expanding text different from the previous paragraph. Consequently, the type of relations in paragraph will determine the type of relation on tree structure as a mono-nucleus or multi nucleus. The condition of nuclerity on tree structure is an important factor to assign score level leading to an important text in summary. Therefore, the assignment of relation types of each paragraph is very important.

From the study of Thai linguistic behavior of an academic document in an agriculture domain for 286 paragraphs discovered each paragraph of document existed linguistic information used in giving a relation species for 66.08%. Also it could be divided into two patterns as Discourse Cues and repeating terms for the beginning of new paragraph. In Table 2 being show a number of linguistic information existing in each pattern.

Table 2 showing an existing number of linguistic information in each pattern

Linguistic information	Discourse Cues				
	numberings	**bullets**	**DM at beginning of the paragraph**	**DM at ending of the paragraph**	**Repeating word**
Existing number (%)	57.67	18.52	2.65	6.35	14.81

From Table 2 can explain as the following:

1. Use discourse Cues as numberings, bullets and discourse marker.

1.1 Numbering gives a types of relation as list relation, multi-nucleus, being a pattern as {(1., 1), (1), (1.1), (1.1.1)}. For example as existing in Figure 4, segment F, G, H and M, N, O.

1.2 Bullets gives species of relation as list relation, multi-nucleus, being a pattern as {– , · } and exist in the beginning of each paragraph being a status same as to use a bullets as numberings type.

1.3 Discourse marker at the beginning of a paragraph. Discourse marker was exists in the relation type of Elaboration, Summary and consequence relation only. The example of discourse maker of each species with the relation as follows.

Elaboration relation = {นอกจากนี้/more over, นอกจากนั้น/more over}

Summary relation = {สรุปว่า/conclusion}

Consequence relation = {ต่อมา/after that}

1.4 Discourse marker at the end of a paragraph. Discourse marker that was exists in the relation type of list relation only. Discourse mark of each species of the relation is {as, as follow, was}.

2. Use of repeating terms.

The repeating terms/words at the first sentence of each paragraph: at the first sentence of each paragraph is the repeating word existing in the first sentence of each paragraph. From the study of repeating word would show that the author needs to divide a data of each page to be the same item with numbering and bullets. Thus we find out the repeating words in this type to be assigned the species of relation as List Relation as multi–nucleus. For example shown in figure 6.

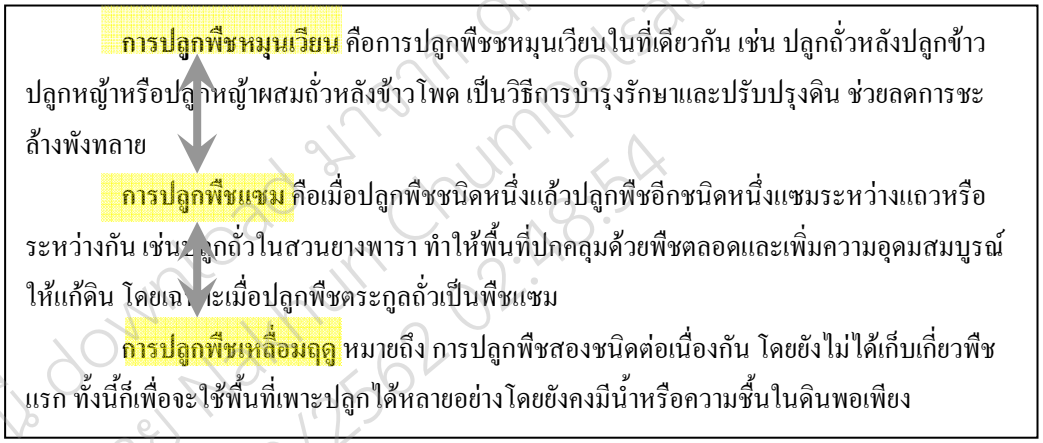


Figure 6 An example of repeating terms at the being position of paragraph.

The repeating word at the position of the last sentence of the previous paragraph is repeating terms existing in the last sentence of the previous paragraph with word which existed in the position of the first sentence of the present paragraph. From the study of repeating terms would show that the author needs to add the following paragraph to enlarge or to describe the detail of the previous paragraph so it find out the repeating terms in this type relation as Elaboration relation being a status mono-nucleus. For example shown in Figure 7.

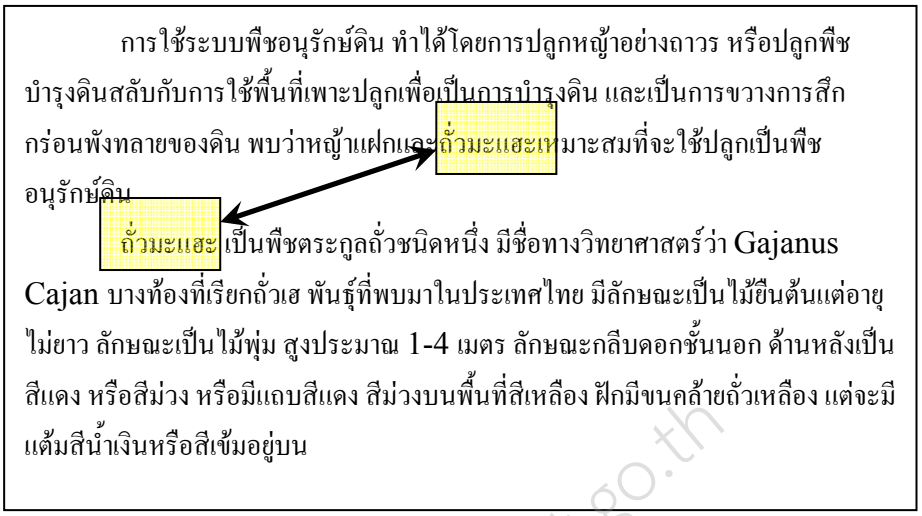


Figure 7 An example of the repeating terms at the last sentence of paragraph.

4. Analyzing of discourse relation with Rhetorical Relation.

Many researchers, Asher (1993), Grimes, (1975), Mann and Thomson (1988), Mani and Bloedorn (1977), were interested in analyzing of discourse relation. Generally, the relation in existing text level is divided in 15 groups and 78 relations i.e. mono-nucleus 50 relations and multi-nucleus 25 relations. All relation is classified (Mann and Thomson, 1988) as subject matter and presentational relation. Subject matter relation gives the more detail of question for reader understanding. Moreover, an interesting text of the thing find a relation as the relation in type of Elaboration, Circumstance, Cause-Result, Condition, Contrast, Summary and Sequence. While the relation in presentational relation gives an importance result to readers , it effected to them being believe in part of nucleus as Motivation, Antithesis, Background, Enablement, Evidence, Justify and concession.

Although this number of existing of each relation depending on the genre and domain; for instance, the analyzing Thai discourse relation with an agricultural domain for 2,950 sentences. We could classify the main relation that was found into 5 relations as Table 3. This databank containing a ratio of existing species of the relation by descending as Elaborate, Condition, Cause-result, Joint and Consequence, respectively. According to this databank has a feature as an academic document

which stressed to describe the fact that it happened. Although the author is not necessary to write text for convincing with the reader, that did not exist a relation in type of Presentational in this databank analyzing. The example of Rhetorical relation shown as in the following:

Example 1

Elaboration EDU 1ปริมาณเพลี้ยเพิ่มขึ้นตามอายุข้าว จากระยะกล้าถึงระยะออกรวง

EDU2***ซึ่ง***ในระยะตั้งท้องมักจะพบประชากรเพลี้ยกระโดดสีน้ำตาลสูงที่สุด

Example 2

Cause-Result EDU 1 ถ้าเพลี้ยทำลายช่อดอกไหม้

EDU2 จะ***ทำให้***ดอกร่วง

Example 3

Condition EDU 1 ***เมื่อ***เมล็ดสืบพันธุ์ของเชื้อราตกลงบนส่วนต่างๆ ของต้นข้าว

EDU2 ***ก็จะ***งอกเป็นเส้นใยเข้าทำลายต้นข้าว

Example 4

Joint EDU 1 เชื้อราสามารถทำให้คอรวงข้าวเน่าเป็นสีน้ำตาลแก่

EDU2 ***และ***ยัง***ทำให้***เมล็ดลีบ

Example 5

Consequence EDU 1 ต้นข้าวจะมีอาการต้นเตี้ยแคระแกรน

EDU2 ***หลังจากนั้น***ไม่ออกรวงหรือออกรวงน้อย

Table 3 Rhetorical relations presented in our corpus (Thai agriculture domain)

Document	**Rhetorical relation**					
An academic document (1)	Elaborate (27.77 %)	Condition (26.19 %)	Cause-Result (19.84 %)	Joint (11.90 %)	Consequence (8.73 %)	Other (5.57%)

Remark: (1) data was obtained from webpage of department of agriculture for 2,950 sentences from 200 files in academic document and classified with a discourse marker as a tool to describe the relation in text level as in Table4. Consequently, existing of discourse marker (Explicit discourse marker: EDM) is 42.86% and in case of no discourse marker at 57.14 % that was in the nearly ratio.

Table 4 The number of explicit discourse markers categories by rhetorical relations.

Relation types	**Discourse markers**				
Elaboration	“โดย” (40.19%)	“แต่” (31.82%)	“ซึ่ง” (18.19%)	“ในกรณี” (4.55%)	“นอกจากนี้” (4.55%)
Condition	“เมื่อ...จะ....” (80%)	“ถ้า...จะ...” (12%)			
Cause-Result	“ทำให้” (66.66%)	“เพราะว่า” (13.33%)	“เนื่องจาก” (13.33%)	“เป็นสาเหตุ” (6.66%)	
Joint	“และ” (100%)				
Consequence	“ต่อมา” (86.67%)	“หลังจากนั้น” (13.33%)			

Table 5 The number of implicit discourse markers categorized by rhetorical relations.

The number of occurrences of relations	Implicit Discourse Markers		
Joint (43.59%)	“และ” (100%)		
Elaborate (41.03%)	“โดย” (56.25%)	“ซึ่ง” (31.25%)	“นอกจากนี้” (12.50%)
Cause-Result (7.69%)	“เมื่อ” (33%)	“เมื่อ...แล้ว” (66.67%)	
Consequence (7.69%)	“หลังจากนั้น” (7.69%)		

From Table 5 shown the percentage of relation, in case of text disappear discourse marker or implicit discourse marker.

The studying and analyzing discourse marker in rhetorical structure is important because we has used discourse marker as a tool to describe the relation in text level and to indicate the rhetorical relation of text as A - [เมื่อมีเพลี้ยกระโดดสีน้ำตาลจำนวนมากดูดกินน้ำเลี้ยงต้นข้าว] B - [จะทำให้ต้นข้าวแสดงอาการใบเหลืองแห้ง คล้ายถูกน้ำร้อนลวก] in text A relative to text B with rhetorical relation called as causal relation having text A describing a cause such that effect to text B as a result action of text A (result) .We found out discourse marker (DM) that combined between text A and text B to{“เมื่อ”… “จะทำให้”....} rhetorical relation in RST Theory as a cause-result called as discourse marker. (“เมื่อ”… “จะทำให้”....) Discourse marker clearly existed as explicit discourse marker: EDM. However, something that was discovered EDM being an ambiguity problem in meaning as semantic ambiguity, meaning that EDM could be given a rhetorical relation more than one meaning in the different context. For example A_1-[ขอบเขตของแผลมีขอบเขตไม่แน่นอน] B_1-[***แต่***ค่อนข้างเป็นรูปสี่เหลี่ยม] and A_2-[โรคนี้เป็นได้ทุกส่วนของต้น] B_2–[**แต่**มักเกิดมากที่ปลายกิ่งและใบ] with this example which was discovered that DM “แต่” this had two meaning

as in text A_1 - B_1 that "แต่" gave rhetorical relation to show a conflict between text A_1 and text B_1 while "แต่" in text A_2-B_2 that gave the rhetorical relation as Elaboration General-Specific as text B_2 giving an addition detail to A_2 in more specific. Both relations had a result to different in text extraction. According to the relation in type of contrast relation was a status of a nuclerity in type of multi-nucleus while an Elaboration relation gave a status of nuclerity in type of mono-nucleus.

In case of combining of text did not exist of discourse marker; it was phenomena that existed in the same feature of the ellipsis and anaphora problem in word or phase level. For example as A-[โรคนี้เกิดได้ในทุกระยะการเจริญเติบโต] B-[โรคนี้เกิดกับต้นข้าวได้ทุกระยะการเจริญเติบโต] C-[ถ้าได้รับเชื้อโรคจู๋] D-[จะแสดงอาการรุนแรงมาก]. Text A and text B were non existing DM creating 2 problems as text B might be co-describe with text A or text C. This problem was called as a structural ambiguity 2. Text B might be co-described with text A or text C with any rhetorical relation. That was directly impact to procedure for building RST. Tree structure was difference and difficult problem to give a real rhetorical relation to the pair of text, which was non-existing DM. By two types of problem was to impact the step of decision making in selection of salient in extraction procedure that used a depth of node score to be a parameter for decision in salient selection to get in summary for the problem of offering the real rhetorical relation to the pair of text. The selection of RR relation would be impact to a status of text node which was multi–nucleus or mono–nucleus so that bringing DM as a tool for discourse structure and rhetorical relation for RST created an important problem as semantic ambiguity and structural ambiguity as said in previous paragraph.

Related Works

There have been various approaches to exploit the discursive organization of text to improve the relevance and quality of final summaries. Many approaches exploit discursive properties in an unprincipled way, for example, by removing all subordinated clauses, by including the sentence immediately preceding a sentence introduced by a discourse marker, etc. These approaches do not contribute to significant progress in the area because they do not increase our understanding of the

problem and do not provide an insightful representation of source texts, even if they may produce improvements in particular summaries.

There have been also some approaches basing summaries on a representation of the discursive aspect of texts. Some of these approaches exploit deep understanding of texts; others are based on shallow evidence. We will specially focus on the latter, because, as we have said, shallow evidence seems more adequate to address the task of text summarization, and it is also closer to our own approach. The most popular theory of text organization underlying summarization approaches has been the RST.

Ono *et al*. (1994b) are the first known researchers to have applied RST to analyze a text as a hierarchical tree, where the relevance of discourse units is relative to their proximity to the root of the tree. They propose that a summary of the text can be obtained exploiting one of the properties derived from the nuclearity principle of rhetorical relations RST: the fact that, in an asymmetric relation between two discourse units, where one, the nucleus, is more important than the other, the satellite, the least important unit can be removed, preserving the main aim of the text. However, their proposal has not been pursued further on.

Corston-Oliver (1998) applied an RST approach to represent the structure of text, and applied this representation to summarization. The rhetorical analysis of texts is based on the deep analysis provided by Microsoft language processing tools, which is said to reach the level of propositional analysis and even allows establishing inferential relations between clauses.

Polanyi *et al.* (2004) produced summaries applying a set of heuristics to a representation of discourse based on the Linguistic Discourse Model (Polanyi, 1988; Polanyi, 1996). Just as Corston-Oliver (1998), this analysis is based on a structure of discourse that builds directly upon sentential syntax and semantics, and considers different kinds of discourse structuring devices, such as basic hierarchical and linear structuring, genre-determined schemata and interactional frameworks. It relies on the language processing tools of Xerox PARC.

Another interesting approach to summarization based on deep analyses of the discursive structure of text is that of Hahn (1990). Text is represented as a hierarchy of thematic units obtained by analyzing it with a knowledge base of the domain. Then, summaries can be obtained by retrieving parts of this structure. This approach is clearly domain-dependent, since it strongly relies on the existence of a knowledge base to obtain the hierarchical representation of the thematic units of the text. A similar approach was proposed by Kan (2003), with the main difference that Kan also provides a methodology to induce the knowledge base from the texts themselves.

Some other approaches rely on shallow cues to obtain a discursive representation of text; they are closer to what we consider useful to introduce progress in the field of text summarization and are also closer to our own approach, so we will discuss them deeper.

Marcu (1997b) is the best known application of the RST-based approach to summarization, because he provides a thorough description of the procedures he applies and also of the resources he exploits. He implements an RST-based discourse parser for English that makes no use of world knowledge, but instead is fully based on shallow textual evidence, namely, discourse markers and word form co-occurrence. The aim of this parser is to obtain a discourse structure that conforms to the well-formedness requirements of RST (Mann and Thompson 1988).

Marcu's parser is totally text-based, it does not depend on domain-dependent sources of knowledge, but exploits the general properties of discourse markers. This guarantees the robustness of the system, but, in contrast to other shallow approaches, partial analyses are not allowed in case a complete one can not be provided.

The architecture of the parser leaves room for incorporating heterogeneous information on discourse structure; for example, word co-occurrence is used to identify cohesion-based relations between discourse units. However, the fact that the approach is relation-based makes it difficult to incorporate various discourse information, because any new information must be expressed in terms of relations and the new relations must be comparable to the previous ones. This makes it difficult,

for example, to incorporate such useful information for discourse processing as information structure, topics, etc. Also, the constraints imposed by RST itself have been criticized for lack of descriptive adequacy that is a one cause of problems when RST was applied to obtain representation of text.

Soricut and Marcu (2003) follow the philosophy of Marcu (1997b) but apply a machine learning approach to build the discourse parser. A corpus with syntactic and RST based rhetorical annotations (Carlson, Marcu and Okurowski, 2001) serves as the basis that of the learning process. The performance of the machine learning parser is better than for the manually built one, probably because rules are based on the objective weighting of a number of examples, instead of the subjective impression of the analyst, and also because the machine learning parser is able to exploit much more information on the examples, more concretely, the lexico-syntactic structure of the training examples.

Schilder (2002) implements an underspecified version of SDRT (Lascarides and Asher, 1993) to obtain a representation of discourse that he argues can be useful to summarize texts. A two-step strategy combining deep and shallow approaches is applied. First, rich structures are obtained with a hand-crafted, grammar-based analysis, but only for those parts of text where discursive clues are found. Then, less informative structures, but covering the whole text, are derived with a more robust strategy based on word forms.

More concretely, discourse makes provide information to apply the rules of an underspecified version of SDRT (Lascarides and Asher, 1993) that determines immediate dominance, dominance, precedence and equivalence relations between minimal discourse constituents, obtaining rhetorical schemata of local scope.

Higher-level discourse structure should be treated with more complex knowledge (intentions, beliefs, plans, genres, etc.). Since neither this knowledge nor the ability to deal with it are available, Schilder obtains higher order relations between discourse units via a topicality measure, an adaptation of the tf * idf metric. Then, the

relations between these schemata are constrained by a topicality measure based on the tf*idf index to determine the relative relevance of each schema.

In essence, Schilder's approach exploits the same kind of information as Marcu (1997b), in short: a rich structure is derived from the available discourse markers and word-based measures account for the relations between those units with no discourse markers. However, Schilder's and Marcu's parsers crucially differ in their ability to integrate heterogeneous discourse knowledge. The stepwise analysis of Schilder (2002) allows the progressive enrichment of the resulting structure, while keeping relative independence between the information to be taken into account.

Finally, an interesting approach to discourse analysis is mentioned, although it has not yet been applied to summarization. Its interest lays in the fact that it establishes a very principled connection between shallow evidence and an insightful theoretical framework. DLTAG (Forbes, *et al.*, 2003) incorporates the syntactic and semantic properties of discourse markers into sentential analysis to go beyond sentential level and reach what they call low level discourse structure and discourse semantics.

The function of discourse markers in structure derivation is to establish relations between textual entities of various sizes (ranging from clauses to the whole previous text) and kinds (from syntax-based units to abstract objects (Asher, 1993). Some discourse markers, like "although" or "because", take their arguments structurally, and some others anaphorically, like "however" or "in that case". For the second kind, it is problematic to determine the referential argument with precision.

Within DLTAG, no specific machinery for the analysis of discourse is developed, but rather the existing mechanisms at clausal level are adapted to the discursive level. This supposes an important economy of development that allows high portability of the system. However, as in the case of the two parsers presented so far, the coverage of the system is critically limited by the amount of discourse markers that has been accounted for.

In summary, many approaches have exploited an analysis of discourse in TS systems, to improve the quality of the resulting systems. Some of these approaches are based on shallow textual clues. Discourse markers are among the most used of these clues, because they are highly informative on the relations between discourse units. Notwithstanding, other kinds of information, like topicality measures or lexico-syntactic structures, improve the accuracy of the analysis.

However, the previous works had not presented the practical solution to solve the lack of coherence in summary result that caused by extracting the salience units from representation of a source text as the RST tree. Because, some the source text has occurs discourse topic discontinuity while its content is progress. And then, the immediately context refer to the previous discourse topic, so-called the event that "flashback". Even though the number of flashback in texts are depends on writing style of a writer. But the flashback can also found in the general texts. So, this thesis focuses on flashback behavior as the root cause of incoherent texts in summary.

Problems in construction the coherence tree

To construct the coherence tree, there are two problems that must be solved. The first problem is to locate the attachment point of incoming node. The second problem is to interpret the existing relations in the text. Furthermore, there is an integration of two previous steps to build up a coherent tree.

1. Attachment Point Identification problem.

This problem is from the point of view of matching, a problem of incoming node (INC) considered being a part of previous discourse tree (PDT) as in Text-Example 1. A possible Attachment Point (AP) that connects the incoming node (INC) with PDT tree consists of AP_1, AP_2, AP_3, AP_4, AP_5, AP_6, and AP_7 as shown in Figure 8. Different attachment points will result in different discourse tree structures. This wills affect the extraction of the discourse tree which will cause incorrect text summarization.

Text-Example 1:

[ถ้าใบข้าวถูกทำลาย]$_{01}$ [ใบจะมีแผลโรคเป็นจำนวนมาก]$_{02}$ [และมีกลิ่นเหม็น] $_{03}$ [นอกจากนี้, ยังทำให้คอ รวงเน่า] $_{04}$ [แต่อาการจะไม่รุนแรง]$_{05}$

[If the rice leaf is destroyed]$_{01}$ [leaf will have many the scar]$_{02}$ [and the scars have bad smell] $_{03}$ [Furthermore, can cause the rice kernels to rot] $_{04}$ [but may be not severe]$_{05}$

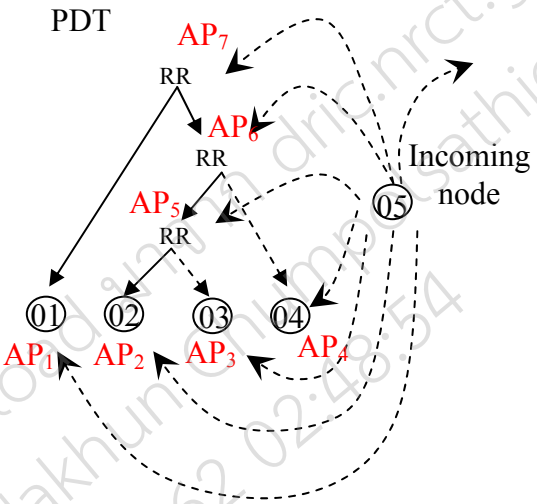


Figure 8 The possible APs for node 05 attach to PDT

When considering time complexity of all possible positions, we see that it is in $O(n^2)$. However, this problem can be solved by using the Right Frontier algorithm (Polanyi, 1988) together with linguistic information, which decreases time complexity to linear order that we will discuss later.

2. Discourse relation interpretation problem.

The interpretation of the discourse relation can be in various forms (Marcu, 1997; Mann and Thompson, 1988; Moore and Pollack, 1992; Asher and Lascarides). For example, there are two methods for interpreting the discourse in

Text-Example 2: [01] is the joint of [02, 06] which denoted by JT[01,[02,06]], or the other way is [01] is a elaboration of [02, 06] which denoted by EB[01,[02,06]], see Figure 9. Also, T_3 and T_4 are the same structure but the top label relations of each tree are different in nuclearity which JT (Joint) is multi-nucleus and EB (Elaboration) is mono-nucleus.

Text-Example 2:

[โรคไหม้ระบาดทั่วไปในทุกภาคของประเทศไทย]$_{01}$ [***เกิดจาก***เชื้อราชื่อ ไพริคู-ลาเรีย]$_{02}$ [***ซึ่ง***เมล็ดสืบพันธุ์ของ ไพริคู-ลาเรีย แพร่กระจายไปได้โดยปลิวไปกับลม] $_{03}$ [***ฉะนั้น*** โรคไหม้จึงแพร่กระจายไปโดยลม] $_{04}$ [***เมื่อ*** เมล็ดสืบพันธุ์ของเชื้อราตกลงบนส่วนต่างๆ ของ ต้นข้าวที่มีความชื้นสูง]$_{05}$ [***ก็จะ***งอกเป็นเส้นใยเข้าทำลายต้นข้าว]$_{06}$

[Blast disease can commonly spread to every parts of Thailand]$_{01}$ [this disease caused by a fungus called Pyricularia]$_{02}$ [which the conidia of this fungus can be blown by the wind] $_{03}$ [therefore blast disease distributes its through the wind] $_{04}$ [when the conidia of the fungus settle on various parts, rice that are highly moist]$_{05}$ [it will sprout in a fiber form, destroying the plant] $_{06}$

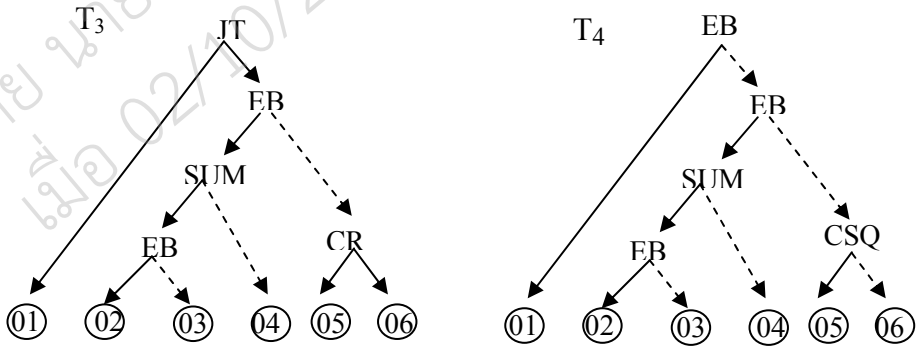


Figure 9 The problems of multiple interpretations between [01], [02, 06] and [05], [06]

If the salience units are extracted from the different tree, the summary were different, i.e. [01, 02] for T_3 and [01] for T_4. This thesis proposes a solution to those problems by the reduction of relations using a transformation from an N-dimensional space of traditional rhetorical relations into a 2-dimensional space of COR and SUBR.

In summary, the problems of ambiguity in attachment points in RST tree and ambiguity in RST relation interpretation will create different tree structure and different nuclearity interpretation, respectively. According, these problems will affect the text summarization.

Our Proposed Solution

Prior work on discourse processing, there are many researchers (Dijk, 1972; Longacre, 1983; Grosz and Sidner, 1986; Mann and Thompson, 1988; Polanyi, 1988, 2004; Asher, 1993, 2002, 2005; Moser and Moore, 1996; Marcu, 1997) paid attention to discourse analysis and text theories that are tree-like representation of discourse structure. And, Grimes, 1975; Halliday and Hasan, 1976; Grosz, 1986; Mann and Thompson, 1988, they had argued to support that some textual unit plays a more important role in the text than others. Paratactic relation and hypotactic relation are two distinction relations that play on the important role, defined by Grimes (1973). Mann and Thompson (1988) had introduced the RST by classifying their rhetorical relations (RR) into two relations, namely a paratactic (multi-nucleus) relation and a hypotactic (mono-nucleus) relation (Marcu, 1997). A paratactic relation is defined to be a relationship that exists between two discourse units (du) which have the same value of interest. A hypotactic relation can be defined as a relationship that exists between discourse units where the value of interest is inequivalent. These two types of relation play a role as nuclearity functions for salience extraction in text summarization. By a Nucleus (N), we mean a discourse unit which is more important or has more value of interest. A Satellite (S) is a discourse unit that contains less value of interest. The example is shown in Text-Example 3 and figure 10.

Text-Example: 3

[เพลี้ยกระโดดสีน้ำตาลจะดูดกินน้ำเลี้ยง บริเวณโคนต้นข้าว]$_{01}$ [จะคอยขับถ่ายมูลน้ำหวาน ออกมา]$_{02}$ [**จะทำให้**ต้นข้าวแสดงอาการใบเหลืองแห้ง คล้ายถูกน้ำร้อนลวกเ]$_{03}$ [**ซึ่งเรียกว่า** "อาการไหม้เป็นหย่อม"]$_{04}$

[The Brown hopper likes to live in the bottom area of the rice plant]$_{01}$ [and infest the sap in that area]$_{02}$ [rice plant shows symptoms of dried leaf as if it has been boiled or burnt]$_{03}$ [which is called "symptom of inconsistent burning"]$_{04}$

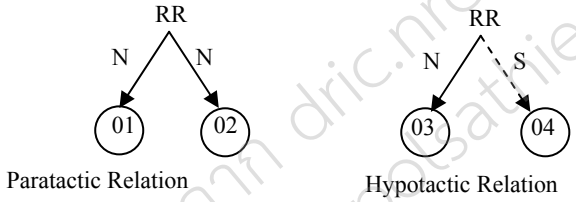


Figure 10 Paratactic relation and hypotactic relation in RST

Because RST proposes too many possible relation such as Elaboration, Explanation, Cause-result, Conditional, Contrast, Sequence, Consequence, Joint, List, Background etc., it is difficult to specify the relations to generate a rhetorical structure. The root of this problem stems from the vagueness of the definition of RST relations (Fukumoto, 1994; Marcu, 2001). Previous text summarization researches (Ono, 1994; Marcu, 1997, Alonso, 2003; Thanh, 2004; Cristea, 2005) used RST theory to obtain surrogate of source text, but they no need to know how many relations for processing discourses therefore text summarization require know where are the salience located in text. Thus, we propose to reduce the number of relations to only two, namely subordinating and coordinating relations. These two relations have been well studied in (Polanyi, 1988, Asher, 2005). A relation R (α, β) between α and β is called subordinating if β adds something to what is said in α so that the information expressed by β is in a sense more granular than the information expressed by α. If R (α, β) is not subordinating relation then it is the coordinating relation.

Let RR is the main set for this discourse rhetorical relation and if we work in accordance with the RST theory, then we would have the following set:

***RR**= {Elaboration, Cause-Result, Condition, List, Joint, Contrast, Explanation, Evidence ...}*

When considering the way in which we decide over the subsets of RR to be either the Coordinating relation (COR) or Subordinating relation (SUBR), we can make the following subsets (see Table 6).

Table 6 Rhetorical relations classified as paratactic relations and hypotactic relations

Types	Rhetorical Relations
Paratactic (COR)	List, Joint, Cause-Result, Problem-Solution, Topic-Shift, Contrast, Question-Answer, …
Hypotactic (SUBR)	Elaboration, Background, Justify, Evidence, Condition, Explanation, Consequence, …

In addition to these two relations, we also take nuclearity into account in order to consider the salience discourse units as show in figure 11 into COR and SUBR only.

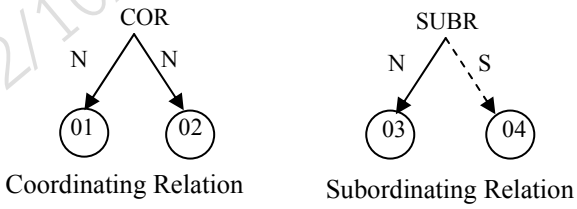


Figure 11 Coordinating and Subordinating relations with nuclearity

In general, the discourse unit in the tree structure with two discourse relations, following the RST theory, called COR&SUBR-tree. The leaves of the tree (the tree end-nodes) comprise of grammatical structures that are elementary discourse units and the internal parts of the tree structure contain relations that exist between the elementary discourse unit and the complex discourse segment.

MATERIALS AND METHODS

Materials

1. Computers

The experiments are run as a computer program. The algorithms are implemented by using PHP programming language. The computer has qualifications as follows:

1.1 Hardwares:

1.1.1 PC with Pentium processor 2.2 GHz.

1) RAM 2 GB.

2) Hard Disk 250 GB.

1.2 Software:

1.2.1 PHP 2.0

1.2.2 RST Tools version.3.41 (Mick O' Donnell)

1.2.3 MATLAB 7.0

1.2.4 WampServer

1.2.5 MS Office 2003

2. Data

This research proposes the methodologies for structuring the surrogate of text summarization, on Thai narrative document corpus in an agricultural domain.

2.1 Choosing texts for the corpus

Texts that used as a corpus in this thesis are in Thai and were chosen from agricultural domain, specifying on plant's disease that published recently in Thai Department of Agricultural Extension website. Most of the texts concern in the narrative the story that describes only one topic as plant's disease symptom. Average document length is 13.2 sentences. Several general design decisions provided a context for the choice of new article. Texts should:

2.1.1 expose a clear factual structure. In this case, the reader selects structured factual data without the consideration of the author's intentions and hence the reasoning within those facts is ignored.

2.1.2 be short enough to allow annotators to quickly gain a good understanding of their overall structure of a document and their goals. Yet, they should be long enough to expose the kinds of relations that possibly hold between larger spans of texts.

2.1.3 be stylistically well-formed written, in terms of a match between rhetorical signals and the intended factual structure,

2.1.4 belong to the same genre of text. While contrastive studies of rhetorical signals or even single instances of relations are an interesting application of a corpus, we need enough texts (per genre) to evaluate machine-learning approaches,

2.1.5 be written in a language that complements efforts in other languages. The corpus is in Thai to my knowledge; this is the first discourse corpus collection effort in this language.

All of these criteria are fulfilled by the short Thai language scientific article, which were written by experts. The 200 document files of plant's disease symptoms were prepared for our experiments. They have the EDU's (elementary discourse unit) average length of 17.34 words with a total number of 3320 EDUs. In

this research purposed, we emphasize to maintain the quality of result of text summarization; readable and comprehend, which our research pays attention to the coherence of textual unit in summary document.

2.2 Linguistics phenomena

The previous works (Halliday and Hasan, 1976, Polanyi, 1988; Mann and Thompson, 1988; Hobbs, 1993; Hovy, 1993, Marcu, 1997; Cristea, 2003) argues that discourse coherence helpful construct the hierarchical structure. The two linguistic phenomena, cue phrases/words or discourse marker (Marcu, 1997) and anaphoric expression, are the evidences that can be found in the articles of discourse structure analysis. This thesis also uses both of cue phrases/words and anaphoric expression as the devices for measuring the degree of coherence between the textual units.

2.2.1 Discourse Markers (DM)

Discourse markers are words or phrases that function to signal how the current utterance relates to prior discourse, also contributing to the meaning of the message. They are best realized by being used at the beginning of clauses. In view of that, a preliminary list of discourse markers can be specified, in terms of their relations, as follows:

The table 7 shows the example of cue phrases/words or discourse marker that used in this thesis.

Table 7 shows the example of cue phrases/words or discourse marker

No.	Discourse Markers	COR & SUBR relations	Rhetorical relations	No.	Discourse Markers	COR & SUBR relations	Rhetorical relations
1	ก็	SUBR	EB	41	ถ้าว่า	COR	CD
2	ก็ต่อเมื่อ	COR	CD	42	ถ้าหาก	COR	CD
3	กระทั่ง	COR	CSQ	43	ถ้าหากว่า	COR	CD
4	ก็แล้วแต่	COR	CD	44	ถึง	COR	CD
5	ก่อน	SUBR	SEQ	45	ถึงกระนั้น	COR	CD
6	กับ	COR	JT	46	ถึงแม้	COR	CD
7	ขณะ	SUBR	SEQ	47	ถึงแม้ว่า	COR	CD
8	ขณะเดียวกัน	COR	JT	48	ถึงอย่างไรก็ตาม	COR	CT
9	ขณะที่	SUBR	SEQ	49	ทั้ง	SUBR	EB
10	ขณะนั้น	SUBR	SEQ	50	ทั้งนี้	SUBR	SUM
11	ครั้น	COR	CR	51	ที่จริง	SUBR	EB
12	ครั้น	SUBR	SEQ	52	ที่แท้	SUBR	EB
13	จน	SUBR	SEQ	53	ทีนี้	SUBR	EB
14	จนกระทั่ง	SUBR	SEQ	54	ทีหลัง	SUBR	SEQ
15	จนกว่า	COR	CD	55	เท่าที่	COR	COMP
16	จนถึง	SUBR	SEQ	56	แทนที่	COR	CD
17	จวบจน	SUBR	SEQ	57	นอกจาก	SUBR	CD/EB
18	จากนั้น	COR	CSQ	58	นอกจากนั้น	SUBR	EB
19	จึง	COR	CR	59	นอกจากนี้	SUBR	EB
20	ฉะนั้น	COR	CR/EB	60	นอกเหนือ	SUBR	EB
21	เช่น	SUBR	EB	61	นอกเหนือจาก	SUBR	EB
22	เช่นเดียว	SUBR	EB	62	นับว่า	SUBR	SUM
23	ดัง	SUBR	EB	63	เนื่องด้วย	COR	CR
24	ดังนั้น	SUBR	SUM	64	เนื่องมาจาก	COR	CR
25	ดุจ	COR	COMP	65	ในขณะเดียวกัน	SUBR	SEQ
26	โดย	SUBR	EB	66	ในทำนองเดียวกัน	SUBR	EB
27	โดยเฉพาะ	SUBR	EB	67	ในเมื่อ	COR	CD
28	โดยเฉพาะอย่างยิ่ง	SUBR	EB	68	ในไม่ช้า	SUBR	SEQ
29	โดยทั่วไป	SUBR	EB	69	บางที	COR	CD
30	โดยที่	SUBR	EB	70	เป็นเพราะว่า	COR	CR
31	ตราบชั่ว	COR	CD	71	เป็นเหตุให้	COR	CR
32	ตราบใด	COR	CD	72	เผื่อ	COR	CD
33	ตราบเท่า	COR	CD	73	เผื่อว่า	COR	CD
34	ตลอดจน	SUBR	EB	74	พอ	COR	CD
35	ต่อมา	COR	CSQ	75	เพราะ	COR	CR
36	ตั้งแต่	SUBR	SEQ	76	เพราะฉะนั้น	SUBR	SUM
37	ตามที่	SUBR	EB	77	เพราะว่า	COR	CR
38	แต่	COR/SUBR	CT/EB	78	เพราะเหตุที่	COR	CR
39	แต่เดิมมา	SUBR	BG	79	เพราะเหตุว่า	COR	CR
40	ถ้า	COR	CD	80	เพียงแต่	COR	CD/COMP

Table 7 (Coun't)

No.	Discourse Markers	COR & SUBR relations	Rhetorical relations	No.	Discourse Markers	COR & SUBR relations	Rhetorical relations
81	เพียงแต่ว่า	COR	CD/COMP	106	สำหรับ	SUBR	EB
82	เพื่อ	COR	CR	107	หรือ	COR	DISJ
83	เพื่อให้	COR	CR	108	หรือไม่ก็	COR	DISJ
84	มิเช่นนั้น	COR	CD/CT	109	หรือไม่เช่นนั้น	COR	DISJ
85	เมื่อ	COR	CD	110	หลังจาก	COR	CSQ
86	เมื่อใด	COR	CD	111	หลังจากที่	COR	CSQ
87	แม้	COR	CD	112	หาก	COR	CT
88	แม้กระทั่ง	COR	CT	113	หากว่า	COR	CT
89	แม้แต่	COR	CT	114	เหมือน	COR	COMP
90	แม้ว่า	COR	CT	115	ให้		
91	ไม่ว่า	COR	CT	116	อนึ่ง	COR	LIST
92	ยิ่ง	SUBR	EB	117	อย่างไรก็ตาม	COR	CT
93	รวมทั้ง	SUBR	EB	118	อะไรก็ตาม	SUBR	EB
94	ระหว่าง	COR	JT	119	อันที่จริง	SUBR	EB
95	ราวกับ	COR	COMP	120	อันเนื่องมาจาก	COR	CR
96	ล้วนแล้วแต่	SUBR	SUM	121	ถ้า...แล้ว...	COR	CD
97	แล้ว	COR	CD	122	ถ้า...จึง...	COR	CD
98	แล้วจึง	COR	CR	123	เมื่อ...ก็...	COR	CD
99	และแล้ว	COR	SEQ	124	หาก...ก็...	COR	CD
100	และ	COR	JT	125	แต่ถ้า	COR	CD
101	ว่า	COR	CD	126	เมื่อ...จะ..	COR	CD
102	เว้นแต่ว่า	COR	CD				
103	เว้นเสียแต่	COR	CD				
104	ส่วน	SUBR	EB				
105	ส่วนใหญ่	SUBR	EB				

Remark: COR = coordinating relation, SUBR = subordinating relation, BG = background relation, EB = elaboration relation, CR = cause-result relation, CD = condition relation, CT = contrast relation, DISJ = disjoint relation, JT = joint relation, SEQ = sequence relation, CSQ = consequence relation, COMP = comparison relation,

In order to make explicitly appearance of the linguistic devices in the corpus, the discourse markers were analyzed into two discourse relations categories; rhetorical relations from the traditional RST and the COR-SUBR relation that is originated from our model.

Table 8 Appearance of Discourse Markers occurs with Discourse relations, in corpus. Let given stand for the some rhetorical relations that EB = Elaboration relation, CSQ= Consequence relation, CR=Cause-Result relation, CD=Condition relation and JT=Joint relation.

Relations	Rhetorical Relations	Explicit Discourse Markers (52.09%)						Implicit DM (30.69%)
		Strong Type (36.74%)	Weak type (10.23%)				Couple DM (10.13%)	
			R1	R2	R3	TOTAL		
Subordinations (SUBR)	EB	84.81	CT (4.35)	JT (4.35)	CD (2.17)	39.13	22.22	56.06
	CSQ	2.53	CR (21.74)	JT (8.69)	-			
Coordinations (COR)	CR	5.06	CD (8.69)	CSQ (21.74)	-	50	77.77	43.94
	CD	2.53	CR (8.69)	EB (2.17)	-			
	JT	5.06	EB (4.35)	CSQ (4.35)	-			
	Others	N/A	N/A	N/A	N/A	10.87	0.01	1.52

In table 8, discourse markers were categorized into two groups; the explicit discourse markers and the implicit discourse markers. The explicit discourse markers are able to infer some coherence relations among the text argument of its marker. Therefore, the number of explicit discourse marker will be the crucial information for designing the system. If the number of discourse markers appearance is low, it is hard to infer those coherence relations. Then, the anaphoric expression which includes anaphor entity and its antecedence is used to infer the coherence relations between sentences instead.

In the table 8, the percentage of the appearance of implicit discourse marker in the last column is 56.06 %, that nearly the half of occurring number in corpus. This means that only one of discourse marker is not enough to be evidence for computing the coherence of two textual units occurred in everywhere of the document. Therefore discourse markers have not presented in every discourse constituents.

2.2.1 Anaphoric Expression (AE)

Anaphora is used as a reference to “point back” to some entities called referent or antecedent. By the observation, the number of anaphoric expression is high as 68 % in this corpus. However, this thesis interests in only demonstrative anaphora and zero anaphora that are outstanding occurred with 66.67 %, 25.59 %, respectively.

Demonstrative anaphora in this thesis included the demonstrative pronoun that occurred after head noun such as “เชื้อรานี้ (**this** Fungus)” in the following sentence.

[โรคไหม้ระบาดทั่วไปในทุกภาคของประเทศไทย]$_{01}$ [***(?โรคไหม้)***เกิดจากเชื้อราชื่อ ไพริคูลาเรีย]$_{02}$ [ซึ่งเมล็ดสืบพันธุ์ (conidia) ของเชื้อรานี้***(?ไพริคูลาเรีย)*** แพร่กระจายไปได้โดยปลิวไปกับลม]$_{03}$

Figure 12 The occurrence of Demonstrative anaphora and Zero anaphora

Zero anaphora is the use of a gap, in a phrase or clause that has an anaphoric function similar to a pro-form. It is often described as “referring back” to an expression that supplies the information necessary for interpreting the gap as in the sentence “Øเกิดจากเชื้อราชื่อ ไพริคูลาเรีย” from figure …….. above that illustrates zero anaphora referring to the word ***(?โรคไหม้)***.

However, the both of discourse markers and anaphoric expression are still not to present in every constituent of texts. This causes the problem of coherence determination. The content relevance which is the relationship measurement between two discourse units will be proposed to solve this problem as describe in the next section.

Methods

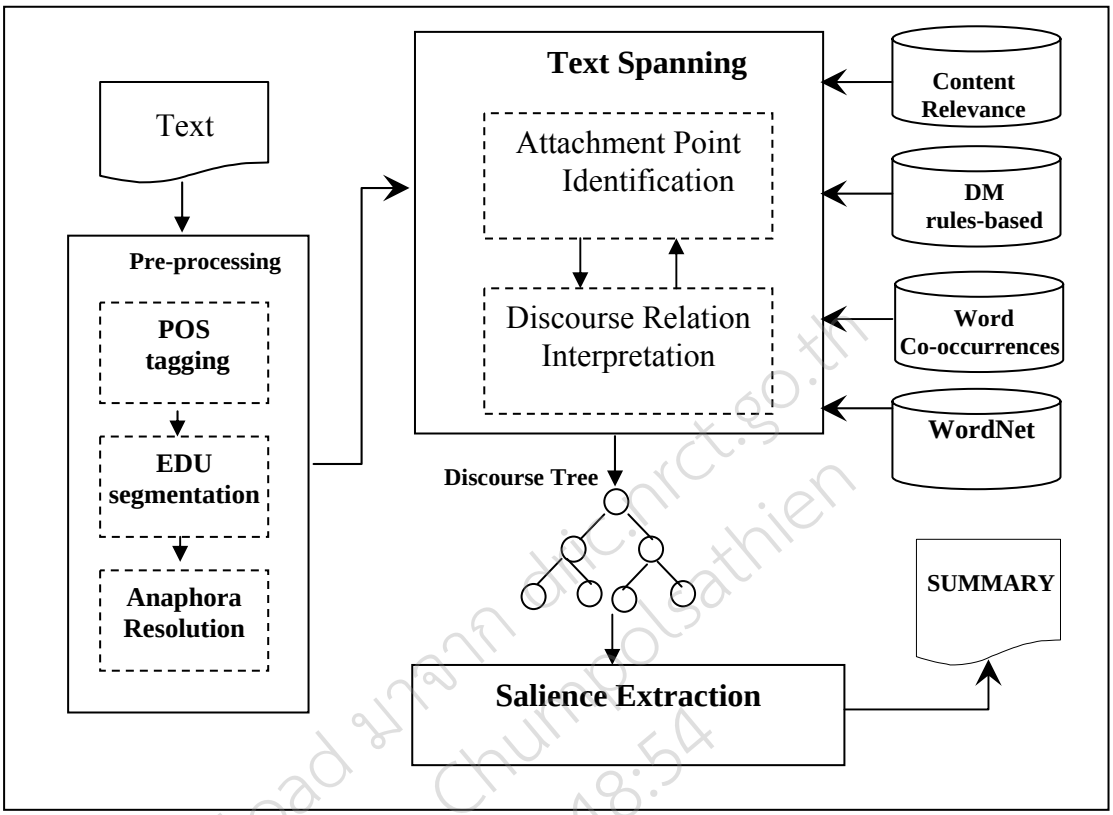


Figure 13 System architecture of Thai Text Structuring for Summarization.

Figure 13 shows the overview of Thai Text Structuring for the Summarization process and its non-trivial resource. The system consists of 3 parts; Pre-processing, Text spanning and Salience Extraction. They work together as the sequential pipe line processing. The pre-processing part composes of three sequential modules developed by the NAiST laboratory at Kasetsart University as the in-house software. At first, a free text document (or raw text) will be the input of the pre-processing. Then the POS tagging program will tag the part of speech of words and then send to the Element Discourse Unit Segmentation system (EDU) in order to segment string of words to be sub-element. The final of this state is the manually tag anaphora. In order to ease the description how the system works as in the flow of figure 12. The text in figure 13 will be used to be the example of working step and discourse phenomenon. Figure 14 was translated from Thai to English, and it was aligned with subscript number at the end of each textual unit.

[โรคไหม้ระบาดทั่วไปในทุกภาคของประเทศไทย]$_{01}$ [(?โรคไหม้)เกิดจากเชื้อราชื่อ ไพริคูลาเรีย]$_{02}$ [ซึ่งเมล็ดสืบพันธุ์ (conidia) ของเชื้อรานี้(?ไพริคูลาเรีย) แพร่กระจายไปได้โดยปลิวไปกับลม]$_{03}$ [ฉะนั้น โรคไหม้จึงแพร่กระจายไปโดยลม]$_{04}$ [เมื่อเมล็ดสืบพันธุ์ของเชื้อรา(?ไพริคูลาเรีย)ตกลงบนส่วนต่าง ๆ ของ ต้นข้าวที่มีความชื้นสูง]$_{05}$ [มัน(?เมล็ดสืบพันธุ์ของเชื้อราไพริคูลาเรีย)ก็จะงอกเป็นเส้นใยเข้าทำลาย ต้นข้าว]$_{06}$ [ปกติโรคนี้(?โรคไหม้)จะทำให้ใบของต้นกล้าเกิดเป็นแผล รูปกลมหรือคล้ายรูปตาของคน เป็นสีเทา]$_{07}$ [และบางครั้ง (?ใบของต้นกล้า)จะมีขอบของแผลเป็นสีน้ำตาลด้วย]$_{08}$ [เมื่อใบข้าวถูกเชื้อโรค(?ไพริคูลาเรีย)เข้าทำลายอย่างรุนแรง]$_{09}$ [แต่ละใบก็จะมีแผลเป็นจำนวน มาก]$_{10}$ [แล้วทำให้ใบข้าวแห้งตาย]$_{11}$ [ถ้าใบข้าวจำนวนมาก แห้งตายไปเพราะโรค(?โรคไหม้)]$_{12}$ [ในที่สุดก็จะทำให้ต้นกล้าแห้ง ตายไปด้วย]$_{13}$ [นอกจากนี้ เชื้อรา(?ไพริคูลา

Figure 14 Example texts for describing the discourse phenomenon (in Thai)

[This disease **(@blast)** can commonly spread to every parts of Thailand]$_{01}$ [**(@ blast)** caused by fungus called Pyricularia]$_{02}$ [which the conidia of this fungus **(@Pyricularia)** can be blown by the wind]$_{03}$ [therefore blast disease distribute its through the wind]$_{04}$ [when the conidia of the fungus settle on various parts, rice that are highly moist]$_{05}$ [it **(@conidia)** will sprout in a fiber form, destroying the plant]$_{06}$ [normally this disease will cause the leaf of the seedling to be scarred; circular or eye-like shape, grayish in color]$_{07}$ [and sometimes **(@leaf of the adult)** will have the a brown epidermis of the scar as well]$_{08}$ [when leaf of the rice is severely infested with viruses] $_{09}$ [each leaf will contain many scars]$_{10}$ [and will cause leak to death]$_{11}$ [if many of the rice leaf dry to death due to the disease]$_{12}$ [will result in the seedling to dry to death]$_{13}$ [furthermore the fungus can cause the tiller, rice kernels to rot, having dark brown color] $_{14}$ [**(@fungu)** can cause blighted seed]$_{15}$ [therefore this fungus can cause a disease infection in the rice plant during the seedling to the production of kennels]$_{16}$

Figure 15 Example texts for describing the discourse phenomenon (in English)

From Figure 14, a string "โรคไหม้ระบาดทั่วไปในทุกภาคของประเทศไทยเกิดจากเชื้อราชื่อไพริคูลาเรีย ซึ่งเมล็ดสืบพันธุ์ของเชื้อรานี้แพร่กระจายไปได้โดยปลิวไปกับลม" will be tagged by POS tagger (Sudprasert, 2003). The result of this step is "โรคไหม้/ncn ระบาด/vi ทั่วไป/adv ใน/prep ทุก/adj ภาค/ncn ของ/prep ประเทศ/ncn ไทย/npn เกิด/vi จาก/prep เชื้อรา/ncn ชื่อ/ncn ไพริคูลาเรีย/npn ซึ่ง/prel เมล็ด/ncn สืบพันธุ์/vi ของ/prep เชื้อรา/ncn นี้/det แพร่/vt กระจาย/vi ไป/vpost ได้/vpost โดย/prep ปลิว/vi ไป/vpost กับ/prep ลม/ncn". (The part of speech abbreviation description is in Appendix A.) After that, the output from the previous step will be the input of EDU segmentation module. The EDU segmentation (Chalernsuk, 2005) will be segmented into the form of simple sentence or the clause-like (Marcu, 1997), for example, [โรคไหม้ระบาดทั่วไปในทุกภาคของ/

ประเทศไทย]$_{01}$ [เกิดจากเชื้อราชื่อไพริคูลาเรีย]$_{02}$ [ซึ่งเมล็ดสืบพันธุ์ของเชื้อรานี้แพร่กระจายไปกับลม]$_{03}$. Then, these EDUs will be captured the anaphoric expression manually as in this example [***{โรคไหม้}***$_{antecidence}$ระบาดทั่วไปในทุกภาคของประเทศไทย]$_{01}$ [***{โรคไหม้}***$_{anaphor}$เกิดจากเชื้อราชื่อ ***{ไพริคูลาเรีย}***$_{antecidence}$]$_{02}$ [ซึ่งเมล็ดสืบพันธุ์ของ***{เชื้อราไพริคูลาเรีย}***$_{anaphor}$แพร่กระจายไปกับลม]$_{03}$. Hence, these texts will be the input text of the main module, Text spanning, to generate the representative of the source text.

1. Constructing the coherent COR&SUBR tree.

As mentioned in previous section, we propose the following model and algorithm for constructing discourse tree representation within system architecture.

1.1 Identifying the attachment point of an appropriate incoming node to PDT

This problem is an attachment point problem. The traditional Right Frontier Constraint: RFC (Polanyi, 1988) only concerns with the anaphoric pronoun. Consequently, the traditional RFC cannot solve our problems, an incoming EDU (elementary discourse unit (Marcu, 1997) becomes a new topic of the content or a new segment, and when the attachment point does not locate in the right frontier area in PDT (Sassen, 2005). Moreover, it could happen that the incoming discourse unit have no any anaphoric pronoun. Therefore, we propose the algorithm of using Adaptive Right Frontier which combines the discourse dependency function (DDF) with the RFC in order to cover these three phenomena.

The DDF is considered having 3 linguistic parameters: Content Relevance (CV), Discourse Marker (DM) and Anaphoric Expression (AE). We use a linear combination model to formulate this function. Let x be a node in PDT, and y be the incoming discourse unit. We define:

$$DDF(x, y) = \alpha\, CV(x, y) + \beta\, DM(x, y) + \gamma\, AE(x, y) \quad (1)$$

where α, β and γ are weighting constants. To determine whether the incoming discourse unit is a new segment or not, the node in PDT that has maximum

value will be tested against a predefined threshold, δ. If max $_{x \in PDT}$ DDF (x, y) ≥ δ, the incoming node will be a node in PDT; otherwise, it will becomes a new segment.

1.2 How to compute the parameter of the coherence dependency function: CV, DM and AE

1) CV (Content Relevance):

Content Relevance is a measurement of the relationship between two discourse units; its value is computed form discourse context. Conceptually, a CV value can be used to determine whether the two discourse units are the same topic or not. The high CV value means two discourse units describing the same topic whereas the low CV value means two discourse units describing the different topics. For example, in the text i, j and k *[Blast disease can commonly spread to every parts of Thailand.]$_i$ [It caused by fungus called Pyricularia.]$_j$ [Each leaf will contain many scars]$_k$* Thus, the value of CV is one component to decide whether the new coming discourse unit should create a new segment or not. To compute this, we define

$$CV(x, y) = \text{sim}\ (f_x\ ,\ f_y)$$

where f_x and f_y are the foci of the discourse unit x in PDT and the incoming discourse unit y respectively. According to the (B.J. Grosz 1995), we use the NP (noun phrase) that precedes the main verb which has the highest potential of the discourse entities as a focus of the discourse unit. The similarity between the foci f_x and f_y is then computed by using the cosine of the angle between the word vectors in the vector space model (Salton, 1975). For a fixed collection of corpus, a m-dimensional vector is generated for each word, where m is the number of unique noun word in the corpus. The weight associated with each noun word is calculated based on the number of occurrence of word w_i and w_j in the same k-consecutive EDUs. The reason of this is that the consecutive EDUs usually describe the same topic, and, in this paper, k is set to 3 which is the average length of EDU span that has the same discourse topic.

Text-4:

[The Brown hopper causes Blast disease,**]$_{01}$ [**especially, this disease always occurs in summer.]$_{02}$ [spay the insecticide only in the morning,**]$_{03}$ [**don't spray the insecticide in the evening.]$_{04}$

To compute the word similarity in example 2, we firstly create a set of important words of the corpus which are “Brown hopper”, “Blast disease”, “summer”, “Insecticide”, “morning”, and “evening”. Then, the 6-dimentional vector is generated for each word w_i. The value of j-th dimension is generated by counting the number of time w_i appear together with w_j in the same k-consecutive EDUs; for example, the number of time the word ‘summer’ appears in the same 3-consecutive EDUs with the word ‘morning’ is 2 (S1:S2:S3 and S2:S3:S4) and the number of time the word ‘Brown hopper’ appears in the same 3-consecutive EDUs with the word ‘evening’ is 0. After obtaining the word vector for all words, we can calculate the similarity between word w_i and w_j by using the inner product of the word vector w_i and w_j, which can be computed from the equation:

$$sim\ (w_i, w_j) = \frac{\sum_{k=1}^{m} \boldsymbol{w_i[k]} \times \boldsymbol{w_j[k]}}{\sqrt{\sum_{k=1}^{m} \boldsymbol{w_i[k]}} \quad \sqrt{\sum_{k=1}^{m} \boldsymbol{w^2_j[k]}}}$$

2) DM (Discourse Marker):

DM is the connective device for discourses processing; identify text coherence and semantic relation between EDUs. In this section, we use DM in the first role as the traffic policeman to point out the attachment point of INC. If DM has the signal for the left hand side attach to PDT, such as *“(The Brown hopper likes to live in the bottom area of the rice plant)$_{EDU\ 1}$ (and infest the sap in that area)$_{EDU\ 2}$”*, then the value of DM, is assigned as +1, hereafter called DM_{left} . On the other hand, if the DM has the signal to create a new segment such as *“(In conclusion, agriculture commodity and food standards prepared by ACFS can be applied by all*

stockholders)$_{EDU\ 1}$ (which can benefit all parties concerned in the industry and the economy as a whole)$_{EDU\ 2}$", then its value is -1, hereafter called DM_{right} . When the DM can not be used to guide the direction of the attachment point, the value will be set to zero. To formulate this, we analysis the corpus and create a set of discourse marker that can be used to guide the attachment point direction namely DM_{left} and DM_{right} . The function used to determine the value of the output is defined as in Equation 2.

$$DM(x, y) = \begin{cases} +1 & \text{; if discourse unit } y \text{ contain a discourse marker in } DM_{left} \\ -1 & \text{; if discourse unit } y \text{ contain a discourse marker in } DM_{right} \\ 0 & \text{; otherwise} \end{cases} \quad (2)$$

where DM_{left} = {*ซึ่ง*: which, *โดย*: by means of, *และ*: and, *หรือ*: or ...

DM_{right} = {*ในที่สุด*: finally, *สรุปได้ว่า*: conclusion *โดยปกติ*: normally ...}

3) AE (Anaphoric Expression):

AE value could be used for describing the strength of the relation between discourse units using anaphoric expression. The intuition of assumption is that the unit should be related if they describe the same thing. Specifically, the unit that contains the antecedent of the anaphor of the incoming EDU should be more related with the incoming EDU than the EDU that does not contain any antecedent. To compute this, we use anaphora resolution described in (Kongwa, 2005) to identify antecedent of y. and the value of AE is compute from the number of antecedents of anaphor in y that exists in x.

AE(x, y) = the number of antecedent of anaphor in *y* that exists in *x* (3)

1.3 Identifying the attachment point of an incoming discourse unit

Finally, we propose the Adaptive RFC for computing the coherence value to decide the attachment point of the incoming discourse unit in (4)

$$\text{Adaptive RFC}(x, y) = \text{RFC}(x, y) + \text{DDF}(x, y) \quad (4)$$

where the RFC is the ranking value of the rightmost node of PDT. The RFC value decreases when it is as long as the distance from the bottom of the rightmost of the PDT. The attachment point of an incoming discourse unit is the unit that has maximum adaptive RFC value.

We tested 200 document files about plant diseases in agricultural domain which has 126 discourse markers (COR/SUBR). We used the threshold δ at 0.057 and the coefficient value of (2) with (0.88, 0.47, and 0.55). Maximum Likelihood Estimation (MLE) was used to compute the coefficient numbers of DDF and we adjust a suitable threshold with trial and error method.

1.4 How to adjust the properly coefficient α, β and γ

The three coefficients α, β and γ of the DDF in (1) are playing role as weighting factor for parameter CV, DM and AE, are respectively. Each of the coefficient number has a value in between 0 to 1. We can use the proportion of these numbers to describe the significant to compute the coherence value among them such as if the either value of coefficients α, β and γ is 1 then its meaning that each parameters have been given the equal significant to compute the coherence value by discourse dependency function or DDF. In the same way, if we given [0.25, 0.50, 0.75] are the coefficients α, β and γ respectively. It can be described that the parameter DM has twice significant computing the coherence value to the parameter CV and the parameter AE has triple significant computing the coherence value to the parameter CV and has twice significant computing the coherence value to the parameter DM respectively. In case that what are the suitable value of coefficients α,

β and γ in this work. In order to find out the answer, we formulate the normal form of these coefficients into coefficient tuple < α, β, γ > and T.

$$T = < \alpha, \beta, \gamma >$$

where t_h □ T and α_{hi}, β_{hj} and γ_{hj} is the instance of tuple < α, β, γ >, including a condition that the instance must be the monotonic increasing number with increasing factor 0.01 such as 0, 0.01, 0.02, 0.03, …, 0.99, 1.0 and h is the integer number in [1, 10^7]. In addition to i, j and k are also the integer number in [1, 10^2]. We used trial method to indicate the suitable instance of tuple t_h which given the good DDF, was designed into five steps. The first step, to generate all tuple T (10^7 tuples) by combination the three coefficient α_{hi}, β_{hj} and γ_h. The second step, to apply the instance t_h from the first step to existing text documents, the maximum DDF at this time was so-called the reference $DDF_{ref,h}$. The reference $DDF_{ref,h}$ are corresponding to α_{hi}CV, β_hDM and γ_hAE. In the third step, compute the differentiate value of existing DDF and reference $DDF_{ref,h}$, this result was consider be unsigned number. After that repeat the step 1 to 3 until coverage text document. Step-4, to compute the average differential value of DDF in step-3. In the final step, choose the reference $DDF_{ref,h}$ which has given the most nearby the average of differential value of DDF.

2. Discourse relations disambiguation by reducing transformations approach

This approach is to use coordinating relation and subordinating relation which have common nuclearity as a connected relation to generate discourse tree. This gives a transformation from a problem in an n-dimensional space of relations in RST to the 2-dimensional space of Coordinating and Subordinating relations which have common nuclearity.

Let RR^n be the set of rhetorical relations having n relations r_k , k = 1,2, …,n, where r_k is in R. That is

$RR^n = \{ r_k \in R, k = 1,2, \ldots,n\}$. Let $RR^2 = \{r_c = COR, r_s = SUBR\}$

RR^2 is set of 2 dimension space of 2 semantic relations; COR/SUBR relation. Based on RR^2, we could simplify the resolution of the ambiguity problem in n-dimension to 2 dimensions. We define COR/SUBR properties with nuclearity function of RST.

- **Definition 1:** Coordinating relation means the relation between discourse topics having the same important interrelated events or objects.

- **Definition 2:** Subordinating relation means the relation between entity and interrelated proceeding in the form that one proceeding event or object depends on the other. This definition can be explained as follows;

2.1 Analysis the advantage of COR/SUBR relations comparison with N-relations RST

The result of reducing the N-relations in RST to 2-relations (COR & SUBR relations) have effect two sides of the discourse relation interpretation process. For the first side is the search space, we can estimate the search space complexity of N-relations in RST is the logarithmic order in the best case while the COR&SUBR can produce 1 times constant complexity for search space complexity. Certainly, the COR&SUBR relations is better than N-relations RST in search space complexity. And the second side, the relation ambiguity, we consider the relation ambiguity of N-relations RST by assuming that the particular relation R_k have possibly ambiguous with remain the relations is n-1. And the other relations N-1 also can be computed the relation ambiguity with the same way. In consequence, the whole of relation ambiguity will occur into order n^2. While the COR&SUBR relations has relation ambiguity complexity as 1 constant order. In summary that the performance of search space and relation interpretation of COR&SUBR relation is better than the N-relations RST in the same measure.

2.2 COR & SUBR recognition.

We use Naïve Bayes classifier to classify the COR/SUBR relations and Thai discourse cues defined in (Wattanamethanont et al., 2005), key phrases, and focus continuity will be used as the features in this learning process.

Discourse Marker feature (DM feature), this research classifies discourse markers into two groups that corresponding with RST's nuclearity (Marcu, 1997) as DM_{COR}, DM_{SUBR} . For example,

DM_{COR} = {"และ:and" , "หรือ:or", "แต่:but", "ถ้า-แล้ว:if-then", …},

DM_{SUBR} = {"ซึ่ง:which", "โดย:by", "ดังนั้น:therefore", "นอกจากนี้:in addition to", …}

DM_{COR} is the discourse markers set of coordinating relations which corresponding with multi-nucleus of nuclearity of Marcu. We illustrate COR's discourse marker "และ:and" with text A and text B.

[ที่ใบข้าว จะแสดงอาการใบแห้ง/At the leaf, will show the dried-leaf symptom.]$_A$ [และจะปรากฏแถบสีน้ำตาลที่บริเวณขอบใบ/and the leaf area will appear brown strip at the leaf edge]$_B$

Text A and text B have coordinating relation (COR) because discourse marker "และ/and" plays the role equivalent importance of two texts. In the same way, the others discourse marker in the COR set have function as the same discourse marker "และ/and".

DM_{SUBR} is the discourse markers set of subordinating relation which corresponding with mono-nucleus of nuclearity of Marcu(1997). We show the example of DM_{SUBR} "ซึ่ง/which" with text C and text D.

[โรคใบไหม้เกิดจากเชื้อราชื่อไพริคูราเรีย/Blast caused by fungus called Pyricularia]$_C$ [ซึ่งเมล็ดสืบพันธ์ของเชื้อรานี้สามารถปลิวไปตามลม/which the conidia of this fungus can be blown by the wind]$_D$.

In this thesis, we give the role unequivalent importance of two texts to discourse marker in DM_{SUBR}. Therefore, text C is more important than text D because text D provides information to elaborate text C. But the main of context [C, D] is still place on text C.

Even though, discourse marker is the high potential linguistic device to identify discourse relations, but discourse markers are not always appear that normally, DM is not present in constituent of discourses. Furthermore, some DM has occurs relation ambiguity. So, the other features will need to close off the weak point of DM usage for discourse relation interpretation process. Then this research is using two features; key phrase (KP) feature and focus continuity (FC) feature works together with DM feature for Naïve Bayes learning.

Key phrase feature (KP feature), key phrase is the phrase which can infer the semantic relation but KP is not be a conjunctive device of discourse units pair such as "เกิดจาก/caused by", "เป็นสาเหตุ/be the cause", "ทำให้เกิด/produce an effect", "เรียกว่า/to be called","ชอบที่จะ/be pleased", " มากกว่า/more than", "น้อยกว่า/less than". The first three KP examples can infer the cause-result relation in rhetorical relations which correspond to COR relation in this research. While the next two KP example can infer elaboration relation in rhetorical relation which correspond to SUBR relation.

Focus continuity feature (FC feature), we has hypothesis that if the foci of narrative discourses that are progress in the same topic to continue, then the succeeding discourses express the preceding discourse. The evidence of this hypothesis caused by our observation on corpus and general writing text, the author lays the most important ideas go first and then, the more text authors added by incremental discourse building consist mostly of expansion of the right branches. That is consistent to research in psycholinguistic (Segal, 1991). In the other words, the

succeeding discourse unit has a SUBR relation to the preceding discourse if this hypothesis is exist. For the example*, [Blast disease can commonly spread to every parts of Thailand]$_A$ [can be found at northern in January to March]$_B$ [and can be found at southern in March to July]$_C$* . Within discourse A, B and C, these are the same foci as Blast disease, so the relation SUBR will be assigned to discourse A, B and discourse A, C.

These features computed by supervised learning technique whose annotated data are separated into two sets of COR and SUBR. Each set has 1000 EDU pairs in agricultural domain, and the testing data of each set are 300 EDU pairs. Naive Bayes is applied to calculate the weight of each individual feature.

In order to prove our hypothesis that COR/SUBR relation can help to improve the accuracy of discourse relation interpretation, is a one of two important processes for constructing discourse tree. We had experimented to compare the performance of traditional rhetorical relations in our corpus that consist of five relations {EB, CSQ, CR, CD, JT} and COR/SUBR relations by using Naïve Bayes technique.

Table 9 illustrates the performance of rhetorical relation recognition in Thai agricultural corpus that returns the average accuracy at 80.08%. While we applied the same corpus and the same features for learning the COR/SUBR relations, have given the average accuracy at 87.46% (see table 10). In consequence, the performance of discourse relations COR/SUBR can outperform the rhetorical relations {CD, CR, CSQ, EB and JT} with nearly 7%.

Table 9 Accuracy of rhetorical relations in Thai agriculture corpus.

Accuracy of rhetorical relation recognition (%)			
Rhetorical relations	**DM**	**DM+KP**	**DM+KP+FC**
CD	63.7	74.55	77.01
CR	71.4	79.68	80.23
CSQ	76.47	82.51	76.65
EB	82.3	81.7	86.9
JT	74	74.06	79.63
Average accuracy	73.57	78.5	**80.08**

Table 10 Accuracy of COR/SUBR relations in Thai agriculture corpus.

Accuracy of COR/SUBR relation recognition (%)			
COR/SUBR relation	**DM**	**DM+KP**	**DM+KP+FC**
COR	79.66	80.01	84.95
SUBR	83.52	85.31	89.96
Average accuracy	81.59	82.66	**87.46**

Furthermore, the KP feature and FC feature also have high potential to work together with DM feature in order to increase an accuracy of discourse relation recognition process. In the other word, KP feature and FC feature can disambiguate discourse relations which were inferred by DM.

3. Coherent Tree Construction.

As mention about coherent tree construction in section 4.1 and section 4.2, our algorithm using left-to-right style for the reading input text and we also using the bottom-up style for merging the local trees to the global tree. The merging tree process will be run as iterative process until remain a unique coherent tree, according to the RST theory. We design the algorithm into two phrases that consist of the local tree construction and the global tree construction.

3.1 Local Coherent Tree Construction

See figure 15, our purpose is to build up the coherent tree by reading the each EDUs from left to right. The input of this process that is the EDUs sequence that are correspond to the sequence of sentences or clauses in the input text. And then, the function Get_Edu() will read an EDU from the EDU sequence by left-to right style, the output called the incoming node(inc) variable, at line 4. Next, to check that if inc variable is the first EDU, then it is assigned to the previous discourse tree (PDT), at line 05 otherwise it will be to inc. After that, to check all nodes in the right frontier area by function RFC_Area(), at line 7. The DDF function in section 4.1 will be compute and store those coherence vales with Coherent_value_table variable, at line 8. Consequently the ARF process called for computing the maximum coherence value between inc and the nodes in PDT, at line 09-10, and then, at line 13, COR&SUBR interpretation process will assign the relation name (COR or SUBR) to AP_node and possibly nodes in function variable Coherent_value_table , at line 11. The output in this step may be produced by the new segment that become to the PDT of the next iteration or the INC node added into a node in the PDT. However, before the new PDT occurs for the next iterative, the current PDT will be store into the tree space

```
[01] Local_Tree_Construction(text_segmented){
[02]    PDT =inc= null;
[03]    while (text_segmented<>null)
[04]       inc = Get_Edu(text_segmented);
[05]       if(PDT == null) {PDT=inc; break;}
[06]       else {
[07]          RFC_nodes = RFC_Area(PDT);
[08]          Coherent_value_table = DDF(PDT, inc);
[09]          if(Max_Coherent(Coherent_value_table)) >threshold
[10]               RFC_Promote(Coherent_value_table, RFC_nodes);
[11]               AP_node = Check_Max_Coherent(Coherent_value_table);
[12]          if (AP_node.coherent)>threshold {
[13]            Relation_Interpretation(AP_node, inc);
[14]            Update_Local_Tree(PDT, AP_node, inc);
[15]            }
[16]          else{ /* inc become to the new segment */
[17]                    Add_Tree_Space(PDT);
[18]                    PDT = inc;
[19]               }
[20]    } /* while */
[21] }/* Build_Up_Local_Tree() */
```

Figure 16 Local coherent tree construction algorithm

Summary, at the finally of the local coherent tree construction may have the several individual coherent trees that be stored in the storage space with namely Tree_Space in line 17 of figure 16.

3.2 Global Coherent Tree Construction

The global coherent tree construction, start after that the local coherent tree construction phrase finished, the global tree construction in figure 12 is continued. The linguistic rules-based which was obtained by the expert will be applied to merge each local coherent trees in tree space. Next, the remaining of local tree in tree space will be merged by repeating the ARF process in section 4.1 and the COR&SUBR interpretation process in section 4.2, that are the same as process in the local tree construction. Finally, the unique global tree was generated as if the representative of input text. To fine-gain the algorithms, we will describe its detail with step-by-step in the next.

```
[01] Global_Tree_Construction (Tree_Space) {
[02]    left = right = Global_Tree =null;
[03]    if(Tree_Space<>null)
[04]         Merging_Local_Tree_by_Rules(Tree_Space);
[05]     else exit();
[06]    while(Tree_Space<>null ) {
[07]      left = Get_Tree(Tree_Space);
[08]      if(Tree_Space<>null)
[09]           right= Get_Tree(Tree_Space);
[10]      else {  Global_Tree = Pop(Tree_Space); /* remain only one tree */
[11]              return Global_Tree;}
[12]      new = Merging_Tree(left, right)
[13]      Push(Tree_Space, new)
[14]    }/* while */
[15]      return Global_Tree;
[16]    } /* Global_Tree_Construction */
```

Figure 17 Global coherent tree construction algorithm

In order to refine the global coherent tree algorithm in figure 16, we describe its detail as the form line-by-line. The tree space from the local coherent tree phrase become to the input of this phrase. At line 02, to initiate address of argument of function Merging_Tree that will be referenced at line 07 and line 09. The function Merging_Local_Tree_by_Rules, at line 04, was called to merge the local coherent tree in tree space by using heuristic rules which was prepared by expert. If some local coherent trees can match with rules, then the new local coherent tree are producing by replacing the local coherent trees that was applied with a rule. So, the number of local coherent tree will be decreased, and convergent to the unique tree. To verify the existing member of tree space again, at line 06, and then go out while-loop at line 15, the otherwise to prepare left argument and right argument of function Merging_Tree at line 07 and line 09 respectively. At line 08, to check that are remain the local coherent tree in tree space or not. If tree space has not the member, then the tree which was pointed by left pointer is become to the global_tree, so call function Pop that popping the last member in tree space to the result of this phrase with address of variable global_tree. Otherwise, function Mering_Tree will be run and its output is a new local coherent tree that will be push into tree space, at line 13, by replacing the left argument tree and right argument tree of Merging_Tree. After that, to repeat at

line 06 and will be run this loop until tree space is empty or remaining only one local coherent tree depend on that one condition is reach. Exploring Coherent Tree algorithm by the Existing Data

3.3 Snapshot of Global Coherent Tree Construction

In this section, we illustrate the existing text document form our corpus that was applied with the tree construction algorithm in this research. During the explanation, we have to using Text-5 and snapshots of the output, in figure 17, are take turns through the whole text.

Text-5:

[Blast disease can commonly spread to every parts of Thailand] $_{01}$ [this disease caused by fungus called Pyricularia] $_{02}$ [which the conidia of this fungus can be blown by the wind] $_{03}$ [therefore blast disease distribute its through the wind] $_{04}$ [when the conidia of the fungus settle on various parts, rice that are highly moist] $_{05}$ [it will sprout in a fiber form, destroying the plant] $_{06}$

The output of each iteration of the local tree construction process that show with an incremental tree such as the output of interation-1 is the tree that are composes of text span with edu 01, iteration-2, the text span are composed of edu 01, 02 and the iteration-03 which as following by iteration-02 that produced the text span which are composed of edu 01,02 and 03.

Note that, this tree will span from left to right that are correspond to the sequence of text reading, in figure 17, texts 01, 02, 03, …, 06. So, the process of local tree construction will be finish when the last edu from the input text was already processed in the iteration-6. Note that again, at the end of local coherent tree, iteration-06, in the tree space have two PDT { PDT_{01}, PDT_{02}}. The PDT_{01} was produced into the tree space after that the edu-05 was decided to be the new segment, but the PDT02 caused by the end of edu to finish of its process.

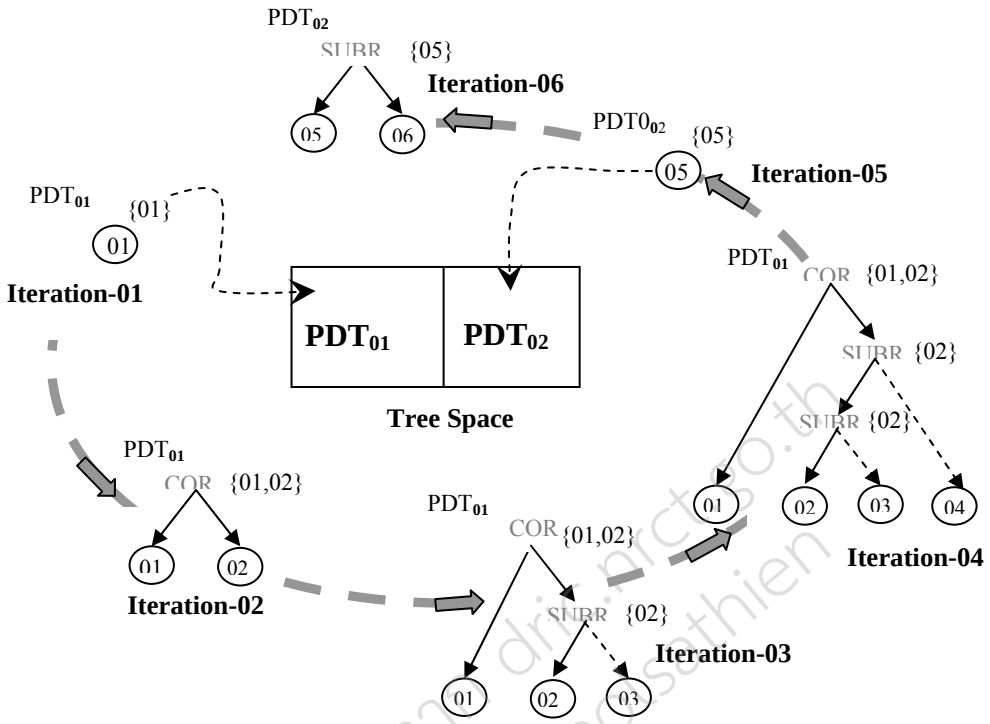


Figure 18 Local Coherence Tree Updating procedure following by Text-3

See figure 18, illustrates how the representative of local PDT can be derived. The local PDT is the output of algorithm in Figure 17. From section 1, the nuclearity property has two statuses {N: nucleus denotes with line arrow and S: satellite denotes with dash line arrow}. We can use nucleus property of discourse relation to decide the representative of individual discourse unit pair. The PDT_j , the discourse unit pair [02, 03] has relationship with subordinating relation; 02 has a nucleus status and 03 has a satellite status, thus we select 02 as the representative of [02, 03], denoted with {02} as parent of [02, 03].

In the same method, we consider two discourse units [02, 03] and [04] as subordinating relation; also the {02} was selected to represent the two discourse units [02, 03] and [04]. Next, [01] and [02, 04] is consecutive two discourse units which are coordinating relation; [01] and [02, 04] has a nuclearity property as a nucleus. Therefore, the representative of [01] and [02, 04] is {01, 02}. In the same way, PDT_j , {05} will be the representative of discourse unit pair [05,06]. Finally, we compute the

discourse relationship between PDT_i and PDT_j with the method in section 2.2 and the $PDT_{i,j}$ will be generated as a result of this step.

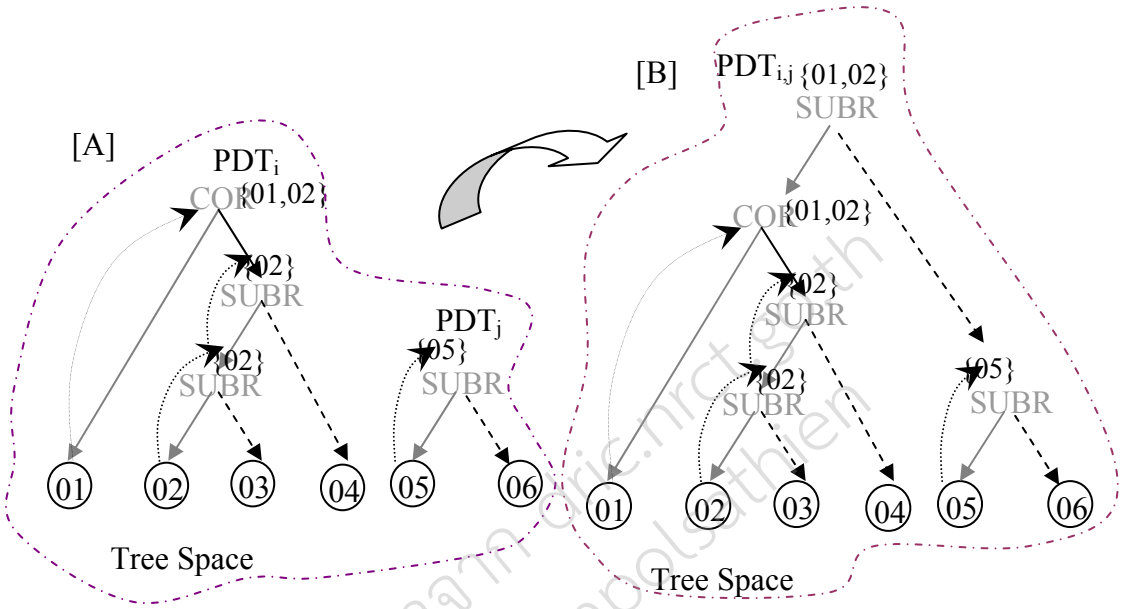


Figure 19 Snapshot, from local coherent tree to global coherent tree.

The experiment was maintained under the condition that the input document has one topic which is a plain text. Therefore, this experiment does not cover the case of a long text. And its result will be reported in the next section.

RESULTS AND DISCUSSION

Evaluation Methods

In order to evaluate the performance of the systems, the human prepared a gold standard align with the results which were produced by our system. We use standard well-know measures with three parameters; precision(PR), recall(RC) and f-measurement(F), are the indicator to capture the accuracy of system's result.

$$precision = \frac{\text{the number of extracted correct results}}{\text{the number of total extracted results}}$$

$$recall = \frac{\text{the number of extracted correct results}}{\text{the number of total correct results}}$$

The F-measurement (F) takes balance of two parameters precision and recall with the formulas:

$$F = \frac{2 \cdot precision \cdot recall}{(precision + recall)}$$

Our experiments were done by testing 200 document files of plant's disease symptoms in agricultural domain. The average EDU length is 17.34 words with the total number of 3320 EDUs. The measurement consists of five parts. In part-1, we consider the accuracy of the position of an attachment point of incoming EDU. In part-2, we measure the average of accuracy of relation identification between COR and SUBR relations. In part-3, we measure accuracy of spanning the discourse structure. In part-4, we measure accuracy result form salience extraction. Finally, we measure the quality of summary with two factors; readable and comprehensible. In each part, we align our system's result with human's result that produced by 3 persons, two of which are linguist and another is not. The win voted 2/3 is the criteria to setup the gold standard for this thesis.

Results and Discussion

The experimental are divided into 5 parts according to our propose models; Attachment Point Identification, COR&SUBR interpretation, the whole of Discourse structure, Salience Extraction and the Quality of summary result.

1. Attachment Point Identification evaluation

In this case, we measure the performance of ARF by comparing with RFC (Polanyi, 1988), and also align with the gold standard. For testing in special case, we test the ARF can closing off the RFC's weak point when discourse updating has occurred discourse phenomena, called flashback. Flashback is an event in a narrative presented out of sequence from an earlier time, and return to the previous discourse topic. This measuring only consider the local coherence tree construction level because it is not fairly to measure the accuracy of ARF in global coherent tree construction level therefore the RFC not claim to in this level.

Table 11 Comparing the ARF performance with the RFC.

API testing	PR(%)	RC(%)	F(%)
RFC	78.38	66.85	72.16
ARF	84.85	77.51	81.01

To test ARF with the flashback, we count the number of occurrences 23 times in corpus. So, we tested the performance of ARF on 23 occurrence of flashback. See at table 12.

Table 12 Comparing the ARF performance with the RFC in flashback phenomenon.

API testing	PR(%)	RC(%)	F-measure(%)
RFC	21.58	34.28	26.49
ARF	77.79	75.33	76.54

Our system can show the explicitly outperform of RFC in the both of general case and the flashback phenomenon. Especially in the flashback phenomenon, the accuracy of ARF is explicitly over than the RFC in all measuring parameter. Thus is the evidence to advocate that the only RFC is not enough for discourse updating and ARF can close off this point. In addition to, this experiment still support that three parameters (CV, DM and AE) in DDF model have the potential to capture the coherence relations between discourse units.

2. COR&SUBR Interpretation evaluation.

For this section, we report the performance of COR&SUBR interpretation from the testing phrases. We prepare 1000 EDU pairs in each relation types for recognition and 300 EDUs pairs for testing the rules that given by Bayesian technique. To show the explicitly system performance, we setup the simple baseline by heuristic rule that the every internal nodes should be the subordinating relations. Because the subordinating relation mostly occurs more than 79% in corpus.

Table 13 Result of COR&SUBR relations Identification, comparing between baseline system and our system.

Result	**COR relation**			**SUBR relation**		
	PR(%)	**RC(%)**	**F(%)**	**PR(%)**	**RC(%)**	**F(%)**
Baseline	53.27	47.19	50.05	63.05	57.41	60.10
System	79.31	76.38	77.82	83.19	79.26	81.18

As the result of COR&SUBR interpretation, there is not different result between COR relation identification and SUBR relation identification. But, it has significant to comparing the result of our system and baseline system. In conclusion, our system outperforms the baseline system which has not the knowledge. This is an evidence that the features of learning; cue words/phrases feature, key phrase feature and focus continuity feature, have the potential to identify COR&SUBR relations. To explore the weight of each feature has effect to identify COR&SUBR relation.

3. COR&SUBR Structure evaluation.

This thesis use the evaluation method of syntactic parser to measure the accuracy of tree span, we compare every tree spans, in the other words know as the subtree, in COR&SUBR structure was generated by our system and gold standard. And we also setup baseline system by using the heuristic rules that are consist of two prior knowledge in attachment point identification step and relation identification step to construct the COR&SUBR structure. The procedure of measurement divided into two procedures. The first, we count every tree spans if textual unit of tree span as the same as textual unit of tree span in gold standard then count it into the candidate set of tree, otherwise is not account. The second, we consider the candidate tree from the previous step that has the same labels (COR or SUBR) as the gold standard then count it be 1, otherwise is not account. We experimented with 200 COR&SUBR tree files.

Table 14 Result of Tree Spans evaluation.

Testing	**Tree spans**		
	PR(%)	**RC(%)**	**F(%)**
Baseline	12.03	21.58	15.45
System	69.33	62.51	65.74

This result shows that the COR&SUBR structures have average accuracy over baseline system nearly 50%. But the accuracy of system has slightly decline when consider with the two previous measurements. We track back the cause of error of tree spans that relate to the error from two previous steps, called propagation errors.

4. Salience Extraction evaluation.

We measure the accuracy in this part by aligning with the gold standard, document by document. Three parameters measure; PR, RC and F-measure would be measured seem as the local measuring. And then the average of PR, RC and F-measure were calculated from the 200 summary files in corpus. See table 14.

Table 15 Result of Salience Extraction evaluation.

Testing	Salience Extraction evaluation		
	PR(%)	**RC(%)**	**F(%)**
System	64.05	53.89	58.53

The accuracy of result in this section seen that more decline than the COR&SUBR structure measurement. However, the number of accuracy of salience extraction not surprise because it nearly the average of the real world summarizer. The propagation error from the previous process is still be a cause of midrange accuracy.

5. Quality of Summarizer's result evaluation.

We mention the quality of summary from summarizer into 2 factors; readable and comprehensible. We still using the win-voted 2/3 is the criteria to decide score for measuring readable factor and comprehensible factor such as, 3 was assigned to strongly agree, 2 was assign for moderate agree and 1was assigned for not agree. We measure the readable and comprehensible on the summary documents which have the local accuracy F-measure more than 80% , there are the number of 15 summary files.

Table 16 The quality of Summary evaluation.

Testing	Criteria Measurement (%)		
	Not Agree	**Moderate**	**Agree**
Readable	10	11	79
Comprehensible	7	13	80

The result on this section can show that the summary document which has coherence between the pieces of textual units can help the user of summarizer is the both of readable and comprehensible text at nearly 80%.

CONCLUSION

This thesis directly addresses the problem of how to construct the suitable discourse structure for text summarization. Therefore the mostly result of currently text summarization has be the text fragments or less coherent text or incoherent text, occurs in the summary. We had ever argue in the former parts that incoherent text in summary causes from the salience extraction process to select the salience units on the text representation which each text units selected are the lack of coherence together. However, the discourse structure construction for text summarization has not the properly algorithms. Especially, discourse structure construction n which referent to rhetorical structure (Mann and Thompson, 1988) would be paid attention to cue words and phrases employ with the set of heuristic rules. The possible results of rhetorical trees were formed by this method are approaching to combinatorial explosion, when text will be increased.

In order to avoid the problem of combinatorial explosion, we apply the greedy in our algorithm. The process of discourse update are forwarding as the pipe line sequence form left to right reading which most likely the human reading a text. We select a best coherence relation in each text span at the right position by using ARF function. At this step, it guarantees that combinatorial explosion has not occurs.

Furthermore, we had found that the number of discourse relations are not necessary identify the salience for text summarization. This thesis also propose to using the only two discourse relations; coordinating relation, subordination relation, within the concept of reducing N-relations in traditional RST to 2-relations in COR-SUBR relation. Obviously, the search space complexity of discourse relation interpretation can be reduced from polynomial order degree to constant order 1. And, it can show that this has effect to improve the accuracy of discourse relations interpretation.

For sank of focusing in depth on issues of quality of summary into two parameters; readable and comprehensible, we evaluates these with by human perspective. Its accuracy result shows that more over 73%.

The results can meet to the thesis's promise, given that the experiment is preliminary, but the vital limitation of our approach is still depending on linguistic phenomenon as discourse markers and anaphoric expressions. Based on our error analysis the performance of the system can be improved and the methodologies can be extended to other sets of lexical semantic and applied to other domain. Moreover, further works to complete the research are performing more tests on large corpora and evaluating by using in the real applications.

LITERATURE CITED

Alemany, L. and M. Fuentes. 2003. Integrating cohesion and coherence for automatic summarization. In **Proceedings of the 11th Meeting of the European Chapter of the Association for Computational Linguistics**, Budapest, Hungary, April 12-17 2003.

Asher, N. 1993. **Reference to abstract objects in discourse**. Kluwer, Dordrecht.

Ballard, D., R. Conrad and R. Longacre. 1971. The deep and surface grammar of interclausal relations. **Foundations of Language** 4: 70-118.

Barzilay, R. and M. Elhadad. 1997. Using lexical chains for text summarization. **ACL Workshop on Intelligent Scalable Text Summarization**, July 1997.

Benitez, A. and S. Chang. 2002. Multimedia knowledge integration, Summarization and evaluation. In **Proceedings of the 2002 International Workshop On Multimedia Data Mining** in conjunction with the International Conference on Knowledge Discovery and Data Mining (MDM/KDD-2002). Edmonton, Alberta.

Chris, D.P. 1981. **The automatic generation of literary abstracts: an approach based on the identification of self-indicating phrases**. In Robert Norman Oddy, S. E. Robertson, C. J. van Rijsbergen, and P. W. Williams, editors

Edmundson, H.P. 1969. New methods in automatic extraction. **Journal of the ACM** 16(2): 264- 285.

Fox, B. 1987. **Discourse Structure and Anaphora.** Cambridge Studies in Linguistics 48. Cambridge: Cambridge University Press.

Fukumoto, J. and J. Tsujii. 1994. Breaking down rhetorical relations for the purpose of analyzing discourse structures. **COLING 94: The 15th International Conference on Computational Linguistics,** August 5-9, 1994, Kyoto, Japan. Proceedings, vol. 2:1177-1183.

Gregory, H. Silber and McCoy, K. 2000. Efficient Text Summarization Using Lexical Chains. In **Proceedings of the ACM Conference on Intelligent User Interfaces** (IUI'2000), January 9-12.

Grimes, E. 1975. The Thread of Discourse. **Jangua Linguarum**, Series Minor, (207)

Grosz, B. and S. Candice. 1986. Attentions, intentions, and the structure of discourse. **Computational Linguistics** 12 (3): 175-204.

_____. 1995. Centering A Framework for Modeling the Local Coherence of Discourse. **Computational Linguistics** 21 (2): 203-226.

Haiman, J. and S.A. Thompson. 1988. **Clause Combining in Grammar and Discourse**. John Benjamins: Amsterdam and Philadelphia.

Hahn, U. 1990. Topic Parsing: Accounting for Text Macro Structures in Full-Text Analysis. In **Information Processing & Management**. 26 (1):135-170.

_____ and I. Mani. 2000. The challenges of automatic summarization. **IEEE Computer** 33 (11): 29-36.

Halliday, M.A.K. 1985. **An Introduction to Functional Grammar**. Edward Arnold Press, Baltimore.

_____ and R. Hassan. 1976. **Cohesion in English**. Longman, London.

Hearst, M.A. 1994. Multi-paragraph segmentation of expository text. In 32nd **Annual Meeting of Association for Computational Linguistics**.

Hobbs, J.R. 1979. coherence and coreference. **Cognitive Science** 3: 67-90.

Hovy, E. 1988. Planning coherent multisentential text. **ACL 1988**,163-169.

_____. 1990. Parsimonious and Profligate Approaches to the Question of Discourse Structure Relations. In **Proceedings of the 5th International Workshop on Natural Language Generation**, pages 128, Dawson, PA.

_____ and D. Marcu. 1998. Automated Text Summarization. **COLING-ACL**. Tutorial.

_____ and C.Y. Lin. 1999. Automated Text Summarization in SUMMARIST. *In* Mani and Maybury, eds., **Advances in Automatic Text Summarization**.

_____. 2001. **Handbook of Computational Linguistics.** Summarization. Oxford University Press.

Jaruskulchai, C. and C. Kruengkrai. 2003. A pratical text summarizer by paragraph extraction for thai, **Proceedings of the Sixth International Workshop on Information Retrieval with Asian Languages** (IRAL 2003), Sapporo, Japan.

Jensen, K. and J.L. Binot. 1987. Disambiguating prepositional phrase attachments by using on-line dictionary definitions. **Computational Linguistics** 13: 251-260.

Kan, M.Y. and K. McKeown. 2001. Domain-specific informative and indicative summarization for information retrieval. **The Document Understanding Conference** (DUC).

Kan, M.Y. 2003. **Automatic text summarization as applied to information retrieval: Using indicative and informative summaries**. Ph.D. thesis, Columbia University.

Kintsch, W. and T.A. Van Dijk. 1983. **Strategies of Discourse comprehension**. Academic Press.

Knott, A. and R. Dale. 1995. Using linguistic phenomena to motivate a set of coherence relations. **Discourse Processes** 18: 35-62.

Kurohashi, S. and M. Nagao. 1994. Automatic detection of discourse structure by checking surface information in sentences. **COLING 94**: The 15th International Conference on Computational Linguistics, August 5-9, 1994, Kyoto, Japan. Proceedings, 2: 1123-1127.

Kozima, H. 1993. Text segmentation based on similarity between words. In **Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics,** 286-288.

Kupiec, J.P. and F. Chen. 1995. A trainable document summarizer. In **Proceedings of the 18th International ACM Conference on Research and Development in Information Retrieval** (SIGIR-95), Seattle, WA, USA, 68–73. Information Retrieval Research, 172-191. Butterworth, London.

Luhn, H.P. 1958. The automatic creation of literature abstracts. **IBM Journal**, 4: 159-165.

Litman, D. and R. Passonneau. 1995. Combining multiple knowledge sources for discourse segmentation. In Proceedings of the 33rd Meeting, 26-30 June, Massachusetts Institute of Technology, Cambridge Massacheutts, USA. **Association for Computational Linguistics**. 108-115.

Longacre, R. 1976. **An Anatomy of Speech Notions**. Ghent: The Peter de Ridder Press.

Maier, E. and E. Hovy. 1991. A metafunctionally motivated taxonomy for discourse structure relations.

Mani, I. and E. Bloedorn. 1997. Multi-Document Summarization by Graph Search and Matching. In **Proceedings of the 14th National Conference on Artificial Intelligence**, 622-628. Providence, Rhode Island.

Mann, W.C. and S.A. Thompson. 1986. Relational Propositions in Discourse. **Discourse Processes** 9: 57-90. Also available as Information Sciences Institute Research Report 83-115, 4676 Admiralty Way, Marina del Rey, CA 90292-6695.

_____. and S.A. Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. 8: 243-281.

Marcu, D. 1996. Building Up Rhetorical Structure Trees. In **Proceedings of the Thirteenth National Conference on Artificial Intelligence**, vol. 2. 1069-1074, Portland, Oregon, August 1996.

_____. 1997a. The rhetorical parsing of natural language texts. In **Proceedings of the Thirty-fifth Annual Meeting of the Association for Computational Linguistics.** 96-103.

_____. 1997b. From discourse structures to text summaries. In **Proceedings of the ACL '97/EACL '97** Workshop on Intelligent Scalable Text Summarization, Madrid, Spain, July 11, 1997. 82-88.

Marcus, M. P., B. Santorini and M.A. Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. **Computational Linguistics** 19: 313-330.

Matthiessen, C. and S. Thompson. 1988. **The structure of discourse and subordination**. In Haiman and Thompson (eds.). 1988: 275-329

Maybury, M. and A. Merlino. 1997. Multimedia summaries of broadcast news. In **International Conference on Intelligent Information Systems.**

_____ and I. Mani. 2001. Automatic summarization. **ACL/EACL'01**. Tutorial.

McKeown, K. 1985. **Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language Text**. Cambridge University Press, Cambridge.

_____ and D. Radev. 1995. Generating summaries of multiple news articles. In **ACM Conference on Research and Development in Information Retrieval SIGIR'95**. Seattle, WA.

_____ *et al.* 1999. Towards multidocument summarization by reformulation: Progress and prospects. In **AAAI 99**.

_____. 2002. The Columbia multi-document summarizer for DUC 2002. In **Workshop on Text Summarization**. Philadelphia.

Moore, J.D. and M.E. Pollack. 1992. A problem for RST: The need for multi-level discourse analysis. **Computational Linguistics** 18: 537-544.

Morris, J. and G. Hirst. 1991. Integrating cohesion al Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text. **Computational Linguistics** 17 (1): 21-43.

Ono, K., K. Sumita and S. Miike. 1994. Abstract generation based on rhetorical structure extraction. **COLING 94**: The 15th International Conference on Computational Linguistics, August 5-9, 1994, Kyoto, Japan. Proceedings, vol. 1: 344-348

Polanyi, L. 1988. A formal model of the structure of discourse. **Journal of Pragmatics** 12: 601-638.

Radev, D. 2000. Text Summarization. **ACM SIGIR**. Tutorial.

Reitter, D. 2003a. **Rhetorical analysis with rich-feature support vector models**. Master's thesis, University of Potsdam

Sanders, T. 1992. **Discourse Structure and Coherence: Aspects of a Cognitive Theory of Discourse Representation**. Lundegem, Nevelland.

_____, W. Spooren and L. Noordman. 1992. Toward a taxonomy of coherence relations. **Discourse Processes** 15: 1-35.

_____. 1993. Coherence relations in a cognitive theory of discourse representation. **Cognitive Linguistics** 4: 93-133.

_____ and W. Spooren. 2001. **Text representation as an interface between language and its users**. *In* T. Sanders, J. Schilperoord and W. Spooren, eds. Text representation: Linguistic and psycholinguistic aspects. Benjamins, Amsterdam.

Schilder, F. 2002. Robust discourse parsing via discourse markers, topicality and position. *In* **Natural Language Engineering**, vol. 8(2&3). Special issue on robust methods in analysis of natural language data.

Sidner, C.L. 1983. Focusing and discourse. **Discourse Processes** 6: 107-130.

Sparck-Jones, K. 1993. What Might Be In A Summary. **Information Retrieval** 93: 9-26.

Sparck-Jones, K. 1997. Summarising: Where are we now? where should we go? In Inderjeet Mani and Mark T. Maybury, eds., Proceedings of the Workshop on Intelligent Scalable Text Summarization at the 35th **Meeting of the Association for Computational Linguistics, and the 8th Conference of the European Chapter of the Association for Computational Linguistics**. Madrid, Spain.

_____. 1999a. Automatic summarising: factors and directions. Inderjeet Mani and Mark Maybury, eds., **Advances in Automatic Text Summarization**. MIT Press.

_____. 1999b. Automatic Summarizing: Factors and Directions. In In-derjeet Mani and Mark T. Maybury, eds**., Advances in Automatic Text Summarization** 1-13. The MIT Press.

_____. 2001a. Factorial Summary Evaluation. In **Proceedings of the 1st Document Understanding Conference**. New Orleans, LA.

_____. 2001b. Factorial summary evaluation. In **Workshop on Text Summarization in conjunction with the ACM SIGIR** Conference 2001. New Orleans, Louisiana.

_____. 2004. Language and information processing: numbers that count. **ESSLLI**.

Sumita, K., K. Ono, T. Chino, T. Ukita, and S. Amano. 1992. A discourse structure analyzer for Japanese text. In **Proceedings of the International Conference of Fifth Generation Computer Systems** 1992: 1133-1140.

Thompson, S. 1983. Grammar and Discourse: The English Detached Participial Clause. *In* Klein-Andreu, Flora (ed.), **Discourse perspectives on syntax**. Academic Press, New York.

Teufel, S. and M. Moens. 1998. Sentence extraction and rhetorical classification for flexible abstracts. In **AAAI Spring Symposium on Intelligent Text Summarisation** 1998: 16 - 25.

_____. 2002a. Summarising Scientific Articles - Experiments with Relevance and Rhetorical Status. In **Computational Linguistics** 28: 4.

_____. 2002b. Summarizing scientific articles - experiments with relevance and rhetorical status. In **Computational Linguistics** vol. 28(4). Special Issue on Automatic Summarization.

Vander, L.K. 1993. **Speaking of Actions: Choosing Rhetorical Status and Grammatical Form in Instructional Text Generation**. Ph.D. dissertation. University of Boulder, Colorado. Published as Technical Report, University of Boulder, Colorado.

Zechner, K. 1996. Fast generation of abstracts from general domain text corpora by extracting relevant sentences. In **COLING 96**: The 16th International Conference on Computational Linguistics, August 5-9, 1996, Copenhagen, Denmark. Proceedings, vol. 2:986-989.

APPENDICES

Appendix A

Resources

Part of speech:

	Kasetsart Part Of Speech	Examples
	NOUN	**NOUN**
1	npn (Proper noun)	น้ำดอกไม้ ตาขาวปะเหลียน
2	nnum (Cardinal number)	พัน หมื่น แสน ล้าน etc.
3	norm (Ordinal Number Marker)	ที่
4	nlab (label noun)	1 2 ก ข
5	ncn (Common Noun)	ช้าง,ม้า
6	nct (Collective Noun)	ฝูง,พวก,
7	ntit (title Noun)	นาย,นาง,นางสาว
	PRONOUN	**PRONOUN**
8	pper (personal Pronoun)	เขา,คุณ,ท่าน
9	pdem (Demonstrative Pronoun)	นี้,นั้น,นั่น
10	pind (indefinite Pronoun)	ใครๆ,ผู้ใด, ต่าง, บ้าง
11	ppos (Possessive Pronoun)	ของคุณ,ของเรา
12	prfx (Reflexive Pronoun)	เอง
13	prec (Reciprocal Pronoun)	กัน
14	prel (Relative pronoun)	ที่ ซึ่ง อัน
15	pint (Interrogative Pronoun)	ทำไม,อะไร

	Kasetsart Part Of Speech	**Examples**
	VERB	**VERB**
16	vi (intransitive Verb)	เดิน,นั่ง, ซึม,กดดัน,กระจาย
17	vt (Transitive Verb)	กรุณา,กลัว,กวนใจ
18	vcau (Causative Verb)	ให้,ทำให้
19	vcs (Complementary State Verb)	เป็น อยู่ คือ กล่าวคือ
20	vex (Existential Verb)	มี
21	prev (Pre-Verb)	จะ,ยัง,คง
22	vpost (Post-verb)	ไป มา ขึ้น ลง
23	honm (honorific marker)	พระ,ทรง,พระราช
	DETERMINER	**DETERMINER**
24	det (determiner)	นี้,นั้น
25	indet (Indefinite determiner)	ใด,อื่น, อย่างไร
	ADJECTIVE	**ADJECTIVE**
26	adj (Adjective)	ขยัน, กำยำ, กิตติมศักดิ์
	ADVERB	**ADVERB**
27	adv (adverb)	กลางคัน,กว่า, แรก สุดท้าย ก่อน หลัง
28	advm1 (Adverb Marker1)	อย่าง
29	advm2 (Adverb Marker2)	เป็น
30	advm3 (Adverb Marker3)	โดย
31	advm4 (Adverb Marker4)	สัก
32	advm5 (Adverb Marker4)	ตาม

	Kasetsart Part Of Speech	Examples
	CLASSIFIER	**CLASSIFIER**
33	cl (Classifier)	เชือก,เซนติเมตร, ทาง,ประเทศ,ชิ้น etc.
	CONJUNCTION	**CONJUNCTION**
34	conj (Conjunction)	และ ในที่นี้
35	conjc (Co-Conjunction)	ทั้ง...และ ไม่... ก็
36	conjncl (Noun Clause Conjunction)	ว่า ให้ ได้แก่
	PREPOSITION	**PREPOSITION**
37	prep (Preposition)	กับ โดย เมื่อ
38	prepc (Co-Preposition)	ระหว่าง....กับ
	INTERJECTION	**INTERJECTION**
39	int (interjection)	เอ๊ะ อ๋อ อุ๊ย ว้าย
	PREFIX	**PREFIX**
40	pref1 (Prefix1)	การ,ความ
41	pref2 (Prefix2)	ผู้,นัก
42	pref3 (Prefix3)	ชาว
	PARTICLE	**PARTICLE**
43	aff (Affirmative)	ค่ะ, ครับ, จ้า
44	part (particle)	นัก, นั่นเอง
	NEGATIVE	**NEGATIVE**
45	neg (Negative)	ไม่, มิ, ไร้
46	negc (Co-negative)	หา.....ไม่

	Kasetsart Part Of Speech	Examples
	PUNCTUATION	**PUNCTUATION**
47	punc (punctuation)	. - , ‘
	IDIOM	**IDIOM**
48	idm (IDIOM)	รักวัวให้ผูก รักลูกให้ตี
	PASSIVE VOICE MARKER	**PASSIVE VOICE MARKER**
49	psm (Passive Voice Marker)	ถูก โดน
	SYMBOL	**SYMBOL**
50	sym (Symbol)	ฯลฯ, ฯ

Non terminal Symbols = 8 symbols

S = Sentence
NP = Noun Phrase
VP = Verb Phrase
PP = Preposition Phrase
ADJP = Adjective Phrase
ADVP = Adverb Phrase
RELC = Relative Clause
NCL = Noun Clause

The example of Thai narrative text in corpus; plant's disease domain.

Text-A: เพลี้ยกระโดดสีน้ำตาลแมลงศัตรูพืชข้าว

Topic: ข้าว

Topic: แมลงศัตรูพืช เพลี้ยกระโดดสีน้ำตาล

[Thai]

S1: ทั้งตัวอ่อนและตัวเต็มวัยของเพลี้ยกระโดดสีน้ำตาลจะดูดกินน้ำเลี้ยง บริเวณโคนต้นข้าวเหนือระดับน้ำ

S2: ***ในขณะเดียวกัน***จะคอยขับถ่ายมูลน้ำหวาน ออกมา

S3: ***เป็นสาเหตุ***ให้เกิดโรคราดำ

S4: ***เมื่อ***มีเพลี้ยกระโดดสีน้ำตาลจำนวนมากดูดกินน้ำเลี้ยงต้นข้าว

S5: ***จะทำให้***ต้นข้าวแสดงอาการใบเหลืองแห้ง คล้ายถูกน้ำร้อนลวก

S6: ***ซึ่งเรียกว่า*** "อาการไหม้เป็นหย่อม"

S7. ***ถ้า***รุนแรงมาก

S8: ต้นข้าว***จะ***แห้งตาย

S9: เพลี้ยกระโดดสีน้ำตาลสามารถทำลายได้ทุกระยะการเจริญเติบโตของข้าว

S10: ***นอกจากนี้***ยังเป็นพาหะนำเชื้อวิสา

S11: ***ซึ่ง***ทำให้เกิดโรคใบหงิกหรือโรคจู๋ มาสู่ต้นข้าวอีกด้วย

S12: โรคนี้เกิดกับต้นข้าวได้ทุกระยะการเจริญเติบโต

S13: ต้นข้าวอายุตั้งแต่ 15-45 วัน ***ถ้า***ได้รับเชื้อโรคจู๋

S14 ***จะ***แสดงอาการรุนแรงมาก

S15: ส่วนต้นข้าวอายุเกิน 60 วันไปแล้ว ได้รับเชื้อ

S16: อาการจะไม่รุนแรง

S17: ต้นข้าวที่ได้รับเชื้อ***แล้ว***

S18: ***จะ***มีอาการต้นเตี้ยแคระแกรน และไม่ออกรวงหรือออกรวงน้อย

S19: ***ถ้า***สภาพแวดล้อมเหมาะสม

S20: ปริมาณเพลี้ยเพิ่มขึ้นตามอายุข้าว จากระยะกล้าถึงระยะออกรวง

S21: ***ซึ่ง***ในระยะตั้งท้องและออกรวงมักจะพบประชากรเพลี้ยกระโดดสีน้ำตาลสูงที่สุด

S22: ***และ***อาการใบไหม้มักจะเกิดในระยะนี้

[In English]:

S1: The nymph and the adult will suck the sap at tiller of the rice plant, above the water level

S2: in the meanwhile, will secrete waste as honey dew

S3: is the cause of fungus

S4: when there is the brown leafhopper in large numbers sucking the sap from rice

S5: will cause the rice plant to show dry and pronounced yellowing of leaf as if boiled

S6: which is called hopper burn

S7: if severe

S8: the rice plant will wilt to death

S9: brown leafhopper can destroy any stages of growth of rice

S10: in addition it is also a vector for viruses

S11: which will cause rice ragged stunt disease to spread in the rice plant as well

S12: this disease can occur at any stages of growth

S13: rice plants with ages between 15 - 45 days if infected with stunt disease

S14: will show extreme symptoms

S15: for rice plants again more than 60 days, infected,

S16: may be less severe

S17: The infected rice plant

S18: will have stunt symptoms and will either not produce any rice kernels or else low production amount

S19: if environmental conditions are suitable

S20: the amount of brown plant hopper will increase in accordance to the age of the rice, from the seedling period to production of kernels

S21: Which during the metamorphosis and production of rice kernels usually find the population of brown plant hopper at its highest

S22: and the symptom of drying leaf usually occurs during this time.

Text-A annotated:

<para id=001>
<topic id=001>ข้าว(Rice) <\topic >
<topic id=002 name=แมลงศัตรูพืช parent=001> เพลี้ยกระโดดสีน้ำตาล" <\topic>
<text>
<edu id=001>ทั้ง<focus type=0 >< CORef ref= เพลี้ยกระโดดสีน้ำตาล dist=-1 type=1 > ตัวอ่อนและตัวเต็มวัยของเพลี้ยกระโดดสีน้ำตาล </CORef></focus>จะดูดกินน้ำเลี้ยงบริเวณโคนต้นข้าวเหนือระดับน้ำ </edu>
<edu id=002>ในขณะเดียวกัน<focus type=0> <CORef ref=ตัวอ่อนและตัวเต็มวัยของเพลี้ยกระโดดสีน้ำตาล dist=-1 type=0> </CORef> </focus> จะคอยขับถ่ายมูลน้ำหวาน (honey dew) ออกมา </edu>
<edu id=003><focus> <CORef ref=มูลน้ำหวาน dist=-1 type=0> </CORef> </focus> เป็นสาเหตุให้เกิดโรคราดำ </edu>
<rel name=sub; dm=; co-dm=; dm-pos=; dm-type=; kp=เป็นสาเหตุ; relev=002,003; ns=n,s>
<rel name=co; dm=ในขณะเดียวกัน; co-dm=; dm-pos=; dm-type=s; kp=; relev=001,[002,003]; ns=n,n>
<edu id=004>เมื่อมี<focus>เพลี้ยกระโดดสีน้ำตาล</focus>จำนวนมากดูดกินน้ำเลี้ยงต้นข้าว </edu>
<edu id=005>จะทำให้<focus>ต้นข้าว</focus>แสดงอาการใบเหลืองแห้ง คล้ายถูกน้ำร้อนลวก </edu>
<edu id=006>ซึ่งเรียกว่า "<focus>อาการไหม้เป็นหย่อม</focus>" (Hopper burn) </edu>
<rel name=sub; dm=ซึ่ง; co-dm=; dm-pos=2; dm-type=s; kp=เรียกว่า; relev=005,006; ns=n,s>
<rel name=co; dm=เมื่อ; co-dm=จะ; dm-pos=3; kp=ทำให้; relev=004,[005,006]; ns=n,n>
<edu id=007>ถ้า<focus> <CORef ref=อาการไหม้เป็นหย่อม dist=-1 type=0></CORef></focus> รุนแรงมาก </edu>
<edu id=008>ต้นข้าวจะแห้งตาย </edu>
<rel name=co; dm=ถ้า; co-dm=จะ; dm-pos=3; dm-type=s; kp=; relev=007,008; ns=n,n>
<rel name=sub; dm=; co-dm=; dm-pos=; dm-type=; kp=; relev=[004,006],[007,008]; ns=n,s>
<rel name=co; dm=; co-dm=; dm-pos=; dm-type=; kp=; relev=[001,003],[004,008]; ns=n,n>

<edu id=009><focus>เพลี้ยกระโดดสีน้ำตาล</focus>สามารถทำลายได้ทุกระยะการเจริญเติบโตของข้าว </edu>

<edu id=010>นอกจากนี้ <focus> <CORef ref=เพลี้ยกระโดดสีน้ำตาล dist=-1 type=0></CORef></focus> ยังเป็นพาหะนำเชื้อวิสา </edu>

<edu id=011>ซึ่ง<focus> <CORef ref=เชื้อวิสา dist=-1 type=0></CORef></focus>ทำให้เกิดโรคใบหงิกหรือโรคจู๋ (Rice raggedstunt) มาสู่ต้นข้าวอีกด้วย </edu>

<edu id=012><focus>โรคนี้ <CORef ref=โรคจู๋ dist=-1 type=1></CORef></focus> เกิดกับต้นข้าวได้ทุกระยะการเจริญเติบโต </edu>

<edu id=013>ต้น<focus>ข้าวอายุตั้งแต่ 15-45 วัน</focus> ถ้าได้รับเชื้อโรคจู๋ </edu>

<edu id=014><focus> <CORef ref=ต้นข้าว dist=-1 type=1></CORef></focus> จะแสดงอาการรุนแรงมาก </edu>

<rel name=co; dm=ถ้า; co-dm=จะ; dm-pos=3; dm-type=s; kp=; relev=013,014; ns=n,n>

<edu id=015>ส่วน<focus>ต้นข้าวอายุเกิน 60 วัน</focus>ไปแล้วได้รับเชื้อ(ref=เชื้อโรคจู๋ dist=-2) </edu>

<edu id=016><focus>อาการ</focus>จะไม่รุนแรง </edu>

<rel name=co; dm=ส่วน; co-dm=จะ; dm-pos=3; dm-type=w; kp=; relev=015,016; ns=n,n>

<rel name=co; dm=; co-dm=; dm-pos=; dm-type=; kp=; relev=[013,014],[015,016]; ns=n,n>

<edu id=017><focus>ต้นข้าว</focus>ที่ได้รับเชื้อ <CORef ref=เชื้อโรคจู๋ dist=-4 type=1></CORef>แล้ว</edu>

<edu id=018><focus> <CORef ref=ต้นข้าว dist=-1 type=0></CORef></focus> จะมีอาการต้นเตี้ยแคระแกรน และไม่ออกรวงหรือออกรวงน้อย </edu>

<rel name=co; dm=แล้ว; co-dm=จะ; dm-pos=4; dm-type=s; kp=; relev=017,018; ns=n,n>

<rel name=sub; dm=; co-dm=; dm-pos=; dm-type=; kp=; relev=[013,016],[017,018]; ns=n,s>

<rel name=sub; dm=; co-dm=; dm-pos=; dm-type=; kp=; relev=012, [013,018]; ns=n,s>

<rel name=sub; dm=; co-dm=; dm-pos=; dm-type=; kp=; relev=011, [012,018]; ns=n,s>

<rel name=sub; dm=; co-dm=; dm-pos=; dm-type=; kp=; relev=010, [011,018]; ns=n,s>

<rel name=sub; dm=; co-dm=; dm-pos=; dm-type=; kp=; relev=009, [010,018]; ns=n,s>

<rel name=co; dm=; co-dm=; dm-pos=; dm-type=; kp=; relev=[001,008],[009,018]; ns=n,n>

<edu id=019>ถ้า<focus>สภาพแวดล้อม</focus>เหมาะสม </edu>

<edu id=020>ปริมาณ<focus>เพลี้ยกระโดดสีน้ำตาล</focus>จะเพิ่มขึ้นตามอายุข้าวจากระยะกล้าถึงระยะออกรวง </edu>

<edu id=021>ซึ่งใน<focus>ระยะตั้งท้องและออกรวง</focus>มักจะพบประชากรเพลี้ยกระโดดสีน้ำตาลสูงที่สุด

<edu id=022>และ<focus>อาการใบไหม้</focus>มักจะเกิดในระยะนี้</edu>

<rel name=co; dm=และ; co-dm=; dm-pos=2; dm-type=s;   kp=; relev=021,022; ns=n,n>

<rel name=sub; dm=ซึ่ง; co-dm=; dm-pos=2; dm-type=s; kp=; relev=020, [021,022]; ns=n,s>

<rel name=co; dm=ถ้า; co-dm=จะ; dm-pos=3; kp=; dm-type=s; relev=019, [020,022]; ns=n,n>

<rel name=co; dm=; co-dm=; dm-pos=; dm-type=; kp=; relev=[001,018] [019,022]; ns=n,n>

<\text>

<\para>

Text-B: อาการโรคไหม้ในข้าว

[In Thai]:

S1: โรคไหม้ระบาดทั่วไปในทุกภาคของประเทศไทย

S2: ***เกิดจาก***เชื้อราชื่อ ไพริคู-ลาเรีย

S3: ***ซึ่ง***เมล็ดสืบพันธุ์ของ ไพริคู-ลาเรีย แพร่กระจายไปได้โดยปลิวไปกับลม

S4: ***ฉะนั้น*** โรคไหม้จึงแพร่กระจายไปโดยลม

S5: ***เมื่อ***เมล็ดสืบพันธุ์ของเชื้อราตกลงบนส่วนต่าง ๆ ของ ต้นข้าวที่มีความชื้นสูง

S6: ***ก็จะ***งอกเป็นเส้นใยเข้าทำลายต้นข้าว

S7: ปกติโรคไหม้จะทำให้ใบของต้นกล้าเกิดเป็นแผลรูปกลมหรือคล้ายรูปตาคน

S8: ***และบางครั้ง*** จะมีขอบของแผลเป็นสีน้ำตาลด้วย

S9: ***เมื่อ***ใบข้าวถูกเชื้อโรคเข้าทำลายอย่างรุนแรง

S10: แต่ละใบ***ก็จะ***มีแผลโรคเป็นจำนวน มาก

S11: ***แล้ว***แผลโรคทำให้ใบข้าวแห้งตาย

S12: ***ถ้า***ใบข้าวจำนวนมาก แห้งตายไปเพราะเชื้อรา

S13: ***ในที่สุด***ก็จะทำให้ต้นกล้าแห้ง ตายไปด้วย

S14: ***นอกจากนี้*** เชื้อรายังสามารถทำให้คอรวงข้าวเน่าเป็นสีน้ำตาลแก่

S15: ***และยังทำให้***เมล็ดลีบ

S16: อาการจะไม่รุนแรง

S17: ต้นข้าวที่ได้รับเชื้อแล้ว

S18: จะมีอาการต้นเตี้ยแคระแกรน และไม่ออกรวงหรือออกรวงน้อย

S19: ถ้าสภาพแวดล้อมเหมาะสม

S20: ปริมาณเพลี้ยกระโดดสีน้ำตาลจะเพิ่มขึ้นตามอายุข้าว จากระยะกล้าถึงระยะออกรวง

[In English]:

S1: Blast disease can commonly spread to every parts of Thailand

S2: **(@ blast)** caused by fungus called Pyricularia

S3: which the conidia of this fungus can be blown by the wind

S4: therefore blast disease distribute its through the wind

S5: when the conidia of the fungus settle on various parts, rice that are highly moist

S6: **(@conidia)** will sprout in a fiber form, destroying the plant

S7: normally this disease will cause the leaf of the seedling to be scarred; circular or eye-like shape, grayish in color

S8: and sometimes **(@leaf of the adult)** will have the a brown epidermis of the scar as well

S9: when leaf of the rice is severely infested with viruses

S11: and will cause leak to death

S12: if many of the rice leaf dry to death due to the disease

S13: will result in the seedling to dry to death

S14: furthermore the fungus can cause the tiller, rice kernels to rot, having dark brown color

S15: and **(@fungus)** can cause blighted seed

S16: may be less severe

S17: The infected rice plant

S18: will has stunt symp toms and will either not produce any rice kernels or else low production amount

S19: if environmental conditions are suitable

S20: the amount of brown plant hopper will increase in accordance to the age of the rice, from the seedling period to production of kernels.

Knowledge Sources for Computing the Attachment Point Identification Content releVance (CV) of discourse entities in Agricultural corpus

Discourseentity	moistness	farmer	fungus	soil	sprout	rice plant	leaf of rice	nitrogen fertilizer	scar	Pyricularia	conidia	Blast disease
moistness	1.0	0.1272	0.2677	0.0599	0.3312	0.152	0.1684	0.0599	0.1961	0.2878	0.3953	0.2209
farmer	0.1272	1.0	0.0529	0.6482	0.312	0.3948	0.1874	0.6482	0.0279	0.2593	0.1619	0.3341
fungus	0.2677	0.0529	1.0	0.015	0.418	0.2585	0.2743	0.015	0.682	0.3216	0.3924	0.3406
soil	0.0599	0.6482	0.015	1.0	0.1499	0.177	0.0707	0.6667	0.0024	0.0667	0.0687	0.0788
sprout	0.3312	0.312	0.418	0.1499	1.0	0.4563	0.2519	0.15	0.4554	0.5085	0.4177	0.4816
rice plant	0.152	0.3948	0.2585	0.177	0.4563	1.0	0.3383	0.177	0.3144	0.299	0.2724	0.32
leaf of rice	0.1684	0.1874	0.2743	0.0707	0.2519	0.3383	1.0	0.0707	0.3767	0.1767	0.1602	0.1936
Nitrogen fertilizer	0.0599	0.6482	0.015	0.6667	0.15	0.177	0.0707	1.0	0.0023	0.0667	0.0687	0.0787
scar	0.1961	0.0279	0.682	0.0024	0.4554	0.3144	0.3767	0.0023	1.0	0.1758	0.1969	0.1872
Pyricularia	0.2878	0.2593	0.3216	0.0667	0.5085	0.299	0.1767	0.0667	0.1758	1.0	0.696	0.7055
conidia	0.3953	0.1619	0.3924	0.0687	0.4177	0.2724	0.1602	0.0687	0.1969	0.696	1.0	0.6749
Blast disease	0.2209	0.3341	0.3406	0.0788	0.4816	0.32	0.1936	0.0787	0.1872	0.7055	0.6749	1.0

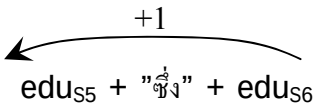
Coherent Direction value by Using Discourse Markers

Discourse Markers Thai/English	**Discourse Markers Pattern**	**Coherent Direction value {-1,0,+1}**
“ฉะนั้น”/ “thus”	edu_{left} + ฉะนั้น + edu_{right}	+1
“โดยทั่วไป” / “generally”	edu^{+} + โดยทั่วไป + edu^{+} ^ foc (edu_{left}^{+})<> foci(edu_{right}^{+})	-1
“โดยทั่วไป” / “generally”	edu_{left}^{+}+โดยทั่วไป + edu_{right}^{+} ^ foc(edu_{left}^{+}) = foci(edu_{right}^{+})	+1
“เมื่อ”/ “when”	edu_{left} +เมื่อ+ edu_{right}	-1
“เมื่อ..จะ”/ “when…will be”	เมื่อ + edu_{left} + จะ + edu_{right}	+1
“ถ้า”/ “if”	edu_{left} +ถ้า+ edu_{right}	-1
“ถ้า ..จะ”/ “if..then”	ถ้า + edu_{left} + จะ + edu_{right}	+1
“ในที่สุด”/” finally”	edu_{left} + ในที่สุด + edu_{right}	+1
“ดังนั้น”/”thus”	edu_{left} +ดังนั้น+ edu_{right}	+1
“นอกจากนี้”/ / “therefore”	edu_{left} +นอกจากนี้+ edu_{right}	+1
“และ”/ “and”	edu_{left} + และ + edu_{right}	+1

For the example, discourse marker “ซึ่ง”/ “which” in Text-A;

S5: ต้นข้าวแสดงอาการใบเหลืองแห้ง คล้ายถูกน้ำร้อนลวก

S6: **ซึ่ง** เรียกว่า "อาการไหม้เป็นหย่อม"



Text segment [5,6], discourse marker “ซึ่ง” play the role as coherent direction device of two texts which has a meaning that consequence text S6 has a coherence to precede text S5 with coherent value +1.

Appendix B

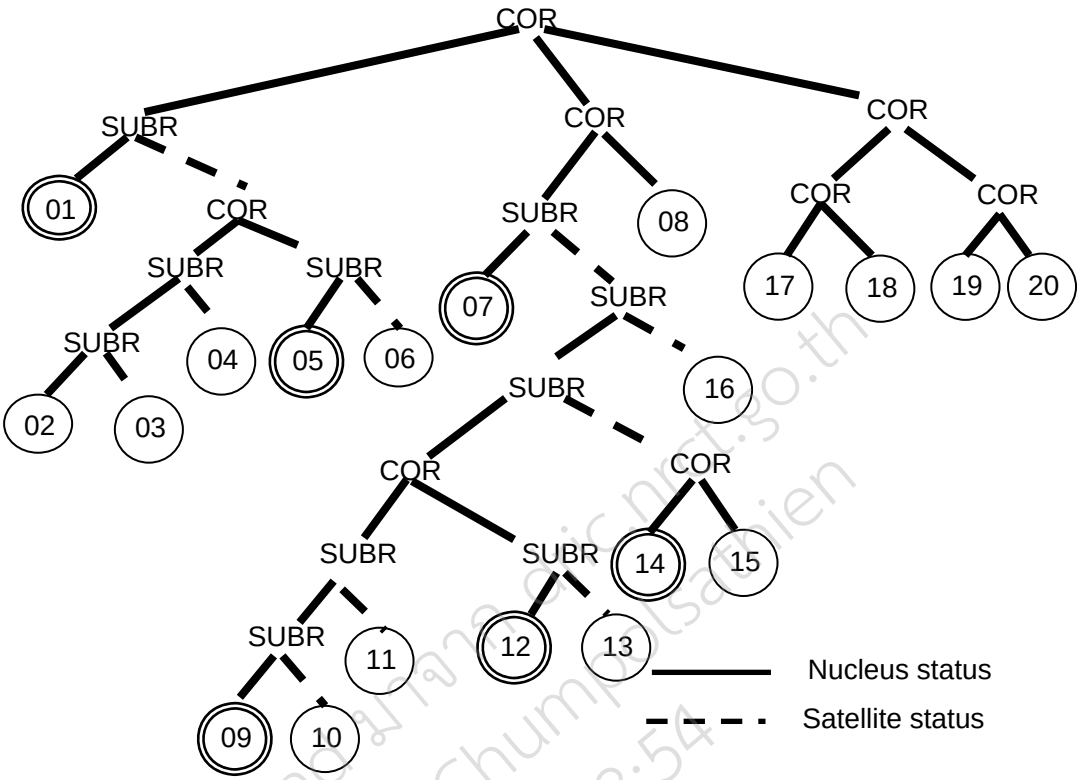
Outputs

Computing the Attachment Point Identification by Using the ARF from Text-B.

Seq. no	Node no.	PDT			INC no.	INC			Coherence Value			DDF	RFC	ARFC	AP	Result
		CV	DM	AE		CV	DM	AE	CV	DM	AE					
1	1	Blast							0.0000	0.0000	0.0000			0.0000	0.0000	[1]
2	1	Blast			2	Blast		(Blast -2,1)	1.0000	0.0000	1.0000	2.0000	1.0000	3.0000	3.0000	[1,2]
3	1 on	Blast							0.6749	0.0000	0.0000	0.6749	0.0000	0.6749		
	2 on	Blast		(Blast -2,1)					0.6749	1.0000	1.0000	2.6749	1.0000	3.6749	3.6749	[1,[2,3]]
	[1,2]	Blast			3	conidia	which	(fungus-3,2)	0.6749	0.0000	1.0000	1.6749	0.5000	2.1749		
4	1 on	Blast							1.0000	0.0000	0.0000	1.0000	0.0000	1.0000		
	2 on	Blast		(Blast -2,1)					1.0000	0.0000	1.0000	1.0000	0.0000	1.0000		
	3 on	conidia	which	(fungus -3,2)					0.6000	0.0000	1.0000	1.6000	1.0000	2.6000		
	[2,3]	Blast							1.0000	1.0000	2.0000	4.0000	0.6600	4.6600	4.6600	[1,[[2,3],4]]
	[1,3]	Blast			4	Blast	therefore	(Blast -4,2)	1.0000	0.0000	2.0000	3.0000	0.3300	3.3300		
5	1 on	Blast							0.6749	-1.0000	0.0000	0.6749	0.0000	0.6749		
	2 on	Blast		(Blast -2,1)					0.6749	-1.0000	0.0000	0.6749	0.0000	0.6749		
	3	conidia	which	(fungus-3,2)					1.0000	-1.0000	1.0000	2.0000	0.0000	2.0000		

Seq. no	Node no.	PDT			INC no.	INC			Coherence Value			DDF	RFC	ARFC	AP	Result
		CV	DM	AE		CV	DM	AE	CV	DM	AE					
	4	Blast	therefore						0.6749	-1.0000	0.0000	0.6749	1.0000	1.6749		
	[2,3]	Blast							0.6749	-1.0000	1.0000	1.6749	0.0000	1.6749		
	[2,4]	conidia	which						0.6749	-1.0000	1.0000	1.6749	0.6600	1.8349		
	[1,4]	Blast	therefore		5	conidia	when		0.6749	-1.0000	1.0000	1.6749	0.3300	2.0049	2.0049	[5]
6	5 on	conidia	when		6	conidia	will be	(conidia - 6,5), (rice plant-6,5)	1.0000	1.0000	2.0000	4.0000	1.0000	5.0000	5.0000	[5,6]
7	5 on	conidia	when						0.6749	-1.0000	0.0000	-0.3251	0.0000	-0.3251		
	6 on	conidia	will be	(conidia - 6,5), (rice plant-6,5)					0.6749	-1.0000	0.0000	-0.3251	1.0000	0.6749		
	[5,6]	conidia	will be	conidia]- 5	7	Blast	normally	(Blast,1,7)	0.6749	-1.0000	0.0000	-0.3251	0.5000	0.1749	new segment	[7]
8	7 on	Blast	normally, will cause	(Blast,1,7)	8	(Φ scar)	and sometime	(Φ scar,7,8)	0.1872	1.0000	1.0000	2.1872	1.0000	3.1872		[7,8]
9	7 off	Blast	normally, will cause	(Blast,1,7)					0.1936	-1.0000	0.0000	-0.8064	0.0000	-0.8064		
	8 on	(Φ scar)	and sometime	(Φ scar,7,8)					0.3767	-1.0000	0.0000	-0.6233	1.0000	0.3767		
	[7,8] on	(Blast)+(Φ scar)		(Blast,1,7) (Φ แผล,7,8)	9	leaf	when		0.2852	-1.0000	0.0000	-0.7148	0.5000	-0.2148	new segment	[9]
10	9	leaf	when		10	leaf	will be	(leaf,9,10)	1.0000	1.0000	1.0000	3.0000	1.0000	4.0000		[9,10]

Text-B representing in COR&SUBR tree structure.



Result: Summary from the source Text-B

S1: โรคไหม้ระบาดทั่วไปในทุกภาคของประเทศไทย

S7: ปกติโรคไหม้จะทำให้ใบของต้นกล้าเกิดเป็นแผลรูปกลมหรือคล้ายรูปตาคน

S8: และบางครั้ง จะมีขอบของแผลเป็นสีน้ำตาลด้วย

S17: ต้นข้าวที่ได้รับเชื้อแล้ว

S18: จะมีอาการต้นเตี้ยแคระแกรน และไม่ออกรวงหรือออกรวงน้อย

S19: ถ้าสภาพแวดล้อมเหมาะสม

S20: ปริมาณเพลี้ยกระโดดสีน้ำตาลจะเพิ่มขึ้นตามอายุข้าว จากระยะกล้าถึงระยะออกรวง