ABSTRACT
Automatic text summarization is a wide research area. There are several ways in which one can characterize different approaches to text summarization: extractive and abstractive from single document or multi document, goal of text summarization (intent, focus and coverage), characteristic of text summarization (frequency-based, knowledge-based and discourse-based), level of processing (surface level, entities level and discourse level) and kind of information (lexicon, structure information and deep understanding). Recently, most researchers for automatic text summarization have transferred their efforts from single document summarization to multi document summarization but they have to be aware of the issues of redundancy, sentence ordering, fluency, etc. Therefore, in this paper, we compare techniques that have done for multi document summarization. Next, we describe evaluation method for automatic text summarization to weigh the quality and performance of system through manual and/or automatic system. In conclusion, we illustrate our future work for multi document summarization.

Categories and Subject Descriptors
A.1 [General Literature]: Introductory And Survey

General Terms
Algorithms, Documentation, Performance

Keywords
Multi document summarization, Text summarization

1. INTRODUCTION
People widely use internet to find information through efficient information retrieval (IR) tools such as Google, Yahoo, AltaVista, and so on. However with the sharply growth of the information on the internet, the abstraction of information from the results of the IR becomes necessary for user to find out really information. So the huge number of information returned by IR system need to be summarized.

Automatic text summarization is the summary of the source text by machine to present the most important information in a shorter version of the source text while still keeping its main semantic content and helps the user to quickly understand large volumes of information. A number of researchers have proposed techniques for automatic text summarization which can be classified into two categories: extraction and abstraction. Extraction summary is a selection of sentences or phrases from the original text with the highest score and put it together to a new shorter text without changes the source text. Abstraction finds new concepts and generates a new shorter text that describes the most important information from the original document. Most of the current automated text summarization system use extraction methods to produce summary. Automatic text summarization works best on well-structured documents, such as news, reports, articles, and scientific papers.

Input to the summarization process could be a single document or multiple documents, text or multimedia information such as image, audio, or video [1]. Single document summarization uses only one document to build summary while multi document summarization uses more than one document that related topic to create summary. After 2002, the single-document summarization task was dropped [10]. Recently, most researchers for automatic text summarization have transferred their efforts from single documents to multiple documents but they have to aware with the issues of redundancy, sentence ordering, collocation, etc.

The rest of the paper is organized as follows, Section 2 describes concept of text summarization. Section 3 shows evaluation method for automatic text summarization. Section 4 presents related work in the area of multi document summarization. Section 5, details conclusion and future work.

2. TEXT SUMMARIZATION
There are several ways in which one can characterize different approaches to text summarization. We present two possible classifications of text summarization systems can be found from literature but not all. The first classification is based on the goal of text summarization, follows [7]. The second proposed in [8], is based on characteristic of text summarization. The third and the last classification summarized by [9] are based on the level of processing and the kind of information

2.1 Goal of Text Summarization
Usually describe in terms of certain key features which relate to the concepts of intent, focus, and coverage:

Intent describes the potential use of the summary. It can be classified into three types:
**Indicative:** Indicative summaries, provide just enough information to judge the relevance of the full text, use to alert the user as to what the source is about and decide to continue reading the full source.

**Informative:** Informative or substantiative summaries serve as acting for the full documents, keeping all important details

**Evaluation:** Evaluative summaries express the point of view of the author on a given topic.

**Focus** refers to the scope of the summary, either generic or user-directed. A generic summary is based on the main concept of a document, while directed summary is based on the topic of interest by the recipient of the summary.

**Coverage** indicates the summary is based on a single document or multiple documents.

### 2.2 Characteristic of Text Summarization

**Frequency-based:** frequency of words or key terms, proximity, and location within the text.

**Knowledge-based:** generally depend on rich knowledge sources to interpret the conceptual structure of the text. Knowledge-based approaches are usually very knowledge-intensive and domain-specific.

**Discourse-based:** theories of text cohesion and coherence of sentence in how much they push the limits of text understanding and the complexity as well as automation of that processing.

### 2.3 Level of Processing

Summarization can be characterized as level of processing at the surface, entity, or discourse level.

**Surface level approaches:** represent information in terms of shallow feature to yield a salience function to extract the most important information. These features include: **Term frequency** (statistics provide a salient terms, the important sentences are the ones that contain words that occur frequently), **Location** (position in text, position in paragraph such as the lead method, the title-base method), **Cue words and phrases** (e.g., cue: “in summary”, “in conclusion”, “the paper describes”, “our investigation”; or emphasizes: “significantly”, “important”, “hardly”, “impossible”)

**Entity level approaches:** try to model text entities (simple word, compound nouns, name entities, etc) for creating an internal representation of text. These approaches include: **Similarity** (similarity between the sentence and the rest of the document, similarity between the sentence and the title of the document), **Proximity** (distance between text units), **Cohesion** (terms of connectivity such as word co-occurrence, local salience, co-reference), **Logical relations** (such as agreement, contradiction, entailment, and consistency) and **meaning representation-based relations** (establish relation between entities in text)

**Discourse level approaches:** model the global structure of the text and its relation to communicative goals that include: **Format of the document** (e.g., hypertext markup, document outlines), **Threads of topics** (topic signature, can be displayed in text) and **Rhetorical structure of the text** (argumentation or narrative structure)

### 2.4 Kind of information

Summarization can be characterized as kind of information at the lexical, structure, or deep understanding.

**Lexical:** exploit the information associated in text depend on the frequency of words. These approaches include: **Word Frequency** (Most frequency words in text are the most representative of its content), **Domain Frequency** (Domain specific words have a first relevance score), **Concept Frequency** (each word in the text is associated to a more general concept), **Cue words and phrases and Chains** (conceptual similarity according to a lexical resource related with chains that are built from lexical items)

**Structure Information:** These approaches include: **Document Structure** (such as heading, section), **Textual Structure** (position in text, paragraph in text), **Conceptual Structure** (chains in lexical approach), and **Discursive Structure** (linear or narrative, hierarchical or rhetoric)

**Deep Understanding:** try to achieve understanding of the text in order to build a summary. There are two approaches: **Top-down** (predefined knowledge structure to texts such as templates or frames) and **Bottom-up** (represent text as highly conceptual construct)

### 3. EVALUATION METHODS

Evaluation for automatic text summarization is important and difficult task. Since one can compare the performance of various tools and methods. It is important to find out the performance of the various tools and techniques that have many criteria to evaluation summary quality such as information coverage, grammatical and discursive coherence, readability etc.

Evaluation [16] of summarization can be intrinsic or extrinsic: **Intrinsic methods** measure a system's quality based on analysis in terms of some set of norms. **Extrinsic methods** measure a system's performance task based on how it affects the completion of other application task.

Evaluation can be evaluated by manual and/or automatic:

**Manual:** Human judgment of the quality of a summary varies from person to person and chooses sentences from document to create manual-extract summaries. Human who evaluated is specialization in each topic.

**Automatic:** Generated summaries and evaluated by computer such as ROUGE [15] (Recall Oriented Understudy for Gist Evaluation) is the official scoring technique for Document Understanding Conference (DUC) 2004, TIPSTER [17]

### 4. DISCUSSION OF MULTI DOCUMENT SUMMARIZATION

Recently year most researchers for automatic text summarization have transferred their efforts from single documents to multiple documents. So this paper, we only deal to compare technique for extraction summarization from multi document. There are many techniques that people use for multi document summarization from the part to present.

Multi document summarization became more interested by the mid 1990s, Summary from multi document must include the important ideas in each document, comparing ideas across document, reducing the size of each document and ordering in new sentence. The first stated of multi document by Radav and McKeown (1995) [8] developed SUMMONS to generate summaries of multiple documents on the same or related events, presenting similarities and differences, contradictions, and generalizations among sources of information from realized as English sentences. In 1995 [17], they improved their SUMMONS to combines it into a conceptual representation of the summary which selects information from underlying knowledge base. The structured conceptual representation of the summary, where
information that appears in only one article is given a lower rating and information that is synthesized from multiple articles is rated more highly.

Kathleen R. McKeown. et al.(2001) [19] presented MultiGen and DEMS for Columbia multi-document summarization system built on the observation that depending on the intended purpose of the summary and on the types of document summarized. This technique focused on the summarization of sets of documents that all describe the same event or news. They used an enhanced version of MultiGen to summarize the document. They used alternative system DEMS( Dissimilarity Engine for Multi document Summarization) for biographical documents. While processing stage, the input articles are transformed into a uniform XML format. After that, the router components of the system determined the type of each input document set and direct the input texts to the summarizers.

Fukumoto J.(2004) [14] proposed a summarization system which automatically classified type of document set and summarized a document set with its appropriate summarization mechanism. This system classified a document set into three types, a series of events, a set of the same events and related events, by using information of high frequency nouns and named entity. The unnecessary parts are deleted after summarized each document and generated multi-document summary. They used single document summarization mechanism for each document of a document set and removed similar parts between summarized documents for generation of a target summary. They applied a TF/IDF based sentence extraction for single document summarization and used of single document summarization for multi-document summarization. Their mechanism of document set classification does not work well in the evaluation because their current implementation has some system bugs in classification mechanism.

Junlin Zhang. et al. (2005) [12] proposed a new approach for multi-document summarization under the Hub/Authority framework. This approach combines the text content with some cues such as cue phrase, sentence length and first sentence and explores the sub-topics in the multi-documents by bringing the features of these sub-topics into graph-based sentence ranking algorithms. They improved two main different points from the old graph-based method: (1) combines the text content with some cues such as cue phrase, sentence length and first sentence. (2) explores the sub-topics in the multi-documents and brings the features of these sub-topics into graph-based sentence ranking algorithms. Also, they used the Markov Models to order the sub-topics that the final summarization should contain and output the text summarization according to the sentence ranking score of all sentences within one sub-topics as user' requirement. The result of evaluation this method on DUC 2004 data proved that the idea of combining the surface and content features under the Hub/Authority framework is an effective graph-ranking schema in multi-document generic text summarization.

Yan-Min C. et al. (2005) [18] used lexical chain for multi-document summarization in Chinese document based on Hownet knowledge database. In their algorithm start from pre-process the text: remove redundant similarities and remain differences in information content among multiple documents, then constructs lexical chains and identifies strong chains. Then significant sentences are extracted from each document and ordered sentences, then recognized and removed redundant information. Finally, the summary is generated in chronological order, and the anaphora resolution technology is applied to improve the fluency of the summary.

De-Xi Liu et al. (2006) [13] proposed a cluster-based method for Chinese multi-document summarization which two steps: sentence clustering and sentence selection. Sentence clustering has two strategies to determine the number of clusters automatically: one strategy makes full use of the summary length fixed by the user while the other one is stability based, which can infer the optimal cluster number automatically. From each Sentence selection, select one sentence to represent the topic by weight the sentence based on the terms included in the sentence.

Judith D. Schlesinger et al. (2008) [11] proposed CLASSY system. This text summarization is an automatic multi-lingual document summarization. It used CLASSY system architecture to summarize document. CLASSY (Clustering, Linguistics, And Statistics for Summarization Yield) is an automatic, extract-generating, summarization system that uses linguistic trimming and statistical methods to generate generic or topic(query)-driven summaries for single documents or clusters of documents. CLASSY used trimming rules to shorten sentences, identify sentences, select sentences and organize the selected sentences for the final summary. The main approach of this research is to generate a multi-lingual summarization document based on summaries document from Machine Translate (MT) document. CLASSY architecture consists of five steps: document preparation, sentence trimming, sentence scoring, redundancy reduction, and sentence ordering. The CLASSY will generate very good summaries when using signature term that computed from English document and machine translated version of Arabic document. The quality of machine translation affected directly to the quality of this summary architecture.

Discussion from researches of multi document summarization can be illustrated by Table 1
<table>
<thead>
<tr>
<th>Researcher(s), Year, Reference</th>
<th>Language(s)</th>
<th>Features(s)</th>
<th>Technique(s)</th>
<th>Evaluation Method(s)/Tool(s)</th>
<th>Strength(s) and Weakness(es)</th>
</tr>
</thead>
<tbody>
<tr>
<td>McKeown, Barzilay, Evans, Hatzivassiloglou, Teufel, Yen Kan, and Schiffman 2001 [19]</td>
<td>Mono-Lingual English</td>
<td>Entity Level</td>
<td>MultiGen, DEMS</td>
<td>- Columbia system did well on grammatically but did not fare as well on cohesion and organization</td>
<td></td>
</tr>
<tr>
<td>Yan-Min, Xiao-Long and Bing-Quan. 2005 [18]</td>
<td>Mono-Lingual Chinese</td>
<td>Discourse Level</td>
<td>Lexical chain</td>
<td>Intrinsic - Lexical chains are effective for Chinese texts summarization</td>
<td></td>
</tr>
</tbody>
</table>

5. CONCLUSION

In this paper we review concept of text summarization which can characterize different approaches to text summarization. Section 2, we present four possible classifications of text summarization systems can be found from literature. The first classification is based on the goal of text summarization. The second propose is based on characteristic of text summarization. The third and the last classification summarized are based on the level of processing and the kind of information. Section 3, we compare and discuss method that people used for multi document summarization. In the last section, we describe evaluation method for automatic text summarization. In the future, we will try to develop an algorithm or new model that supports multi document summarization area combine Lexical Chain with cluster-based method because from strength of [13][18], it well efficient for Chinese texts, that are explicit concatenations of characters, and words are not bounded by spaces as that in English same as Thai Language therefore we will apply this technique to text summarization in Thai Language in our future work.

6. REFERENCES


