THAI NATURAL LANGUAGE SEARCH ENGINE USING SYNONYM APPROACH : CASE STUDY FOR GOOGLE SYNTAX

CHOTIKA POUNGMALI

MASTER OF SCIENCE

IN STRATEGIC MANAGEMENT INFORMATION SYSTEM

MAE FAH LUANG UNIVERSITY

2009

THAI NATURAL LANGUAGE SEARCH ENGINE USING SYNONYM APPROACH : CASE STUDY FOR GOOGLE SYNTAX

CHOTIKA POUNGMALI

THIS THESIS IS A PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN STRATEGIC MANAGEMENT INFORMATION SYSTEM

MAE FAH LUANG UNIVERSITY

2009

THAI NATURAL LANGUAGE SEARCH ENGINE USING SYNONYM APPROACH: CASE STUDY FOR GOOGLE SYNTAX

CHOTIKA POUNGMALI

THIS THESIS HAS BEEN APPROVED
TO BE A PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF MASTER OF SCIENCE
IN STRATEGIC MANAGEMENT INFORMATION SYSTEM
2009

EXAMINING COMMITTEE

.....CHAIRPERSON

(Assoc. Prof. Dr. Kosin Chamnongthai)

.....MEMBER

(Prof. Dr. Chidchanok Lursinsap)

.....MEMBER

(Gp. Capt. Dr. Thongchai Yooyativong)

.....MEMBER

(Dr. Punnarumol Temdee)

ACKNOWLEDGEMENT

I would like to thank my advisor, Gp. Capt. Dr.Thongchai Yooyativong, for helpful comments, valuable suggestion, reviewing, encouragement and guidance in making this thesis a successful one. Moreover, I also wish to thank the members of the committee Prof. Dr.Chidchanok Lursinsap, Assoc. Prof. Dr.Kosin Chamnongthai, and Dr.Punnarumol Temdee for their valuable comments, discussions and suggestions. I also wish to extend my appreciation to all of the members in SMIS group for a friendship, enjoyment, and encouragement.

Finally, I would like to express my deepest gratitude to my parents for supporting everything and my sister, my brother, grandmother, and my best friends for their love, and endless patience to make a success for this work.

Chotika Poungmali

Thesis Title	Thai Natural Language Search Engine using Synonym Approach : Case Study for Google Syntax
Author	Chotika Poungmali
Degree	Master of Science (Strategic Management Information System)
Supervisory Committee	Gp. Capt. Dr. Thongchai Yooyativong Prof. Dr. Chidchanok Lursinsap

ABSTRACT

This thesis proposes the method used to convert Thai natural language phrase into search engine keyword pattern in order to list up the desired websites that are most matched to the query. The proposed method extracts user intention from the phrase, and converts it into keyword pattern with Google syntax so that the appropriate alternative search is obtained. The keyword pattern includes nouns, synonyms, and Boolean operators. The proposed method firstly determine keywords from Thai Natural Language Phrase input, and then find synonyms by using short semantic distance in the synonyms database in order to expand a chance for related websites. The keywords and weighted synonyms are finally converted into keyword pattern for search engine. To evaluate the performance of the proposed method, the comparison between the searched results using keyword pattern and without keyword pattern with 95 queries were conducted. This result has shown that the proposed system provides the average precision of 62.33%, whereas the search without keyword pattern provides the average precision of 36.1%.

Keywords: Search engine / Natural language / Semantic / Synonym

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES (continued)

CHAPTER 1

INTRODUCTION

1.1 Principle and Motivation

Nowadays, internet is a collection of information from the entire world. So it is not easy for searching information to match with the demand. A program called “Search Engine” therefore has been developed to support documents searching on the internet. Intelligent search engine is expected to assist people to efficiently find desired websites. Since users currently need to input keywords in search engine format for extracting desired websites, it is not friendly for normal people to understand the format and it is also difficult to get appropriate website lists. Actually, the interface with human should be friendly as the same level as talking with human, and the system has to understand Natural Language (NL). To develop this kind of search function, the thesis reports exactly concentrating on natural language understanding for search engine have not been found yet. Search engine usually has the problem of providing un-desired results from many matched website.

Generally, the work of search engine consists of 3 main parts (Learn the Net “How search engine work”, online).

1. Robots or spider or crawler: These are looking for documents and their Web addresses. The documents and Web addresses are collected and sent to indexing software.
2. Indexer: The indexing software extracts information from documents and storing it in database.

3. Search engine: The search engine performs a search by entering keywords, the database is searched for documents that match and list the results as hypertext links.

Normally, in using search engine, user inputs keywords and/or key phrase in the tool box for each search engine and after that search engine will return lists of results as hypertext links, as shown in Figure 1.1.

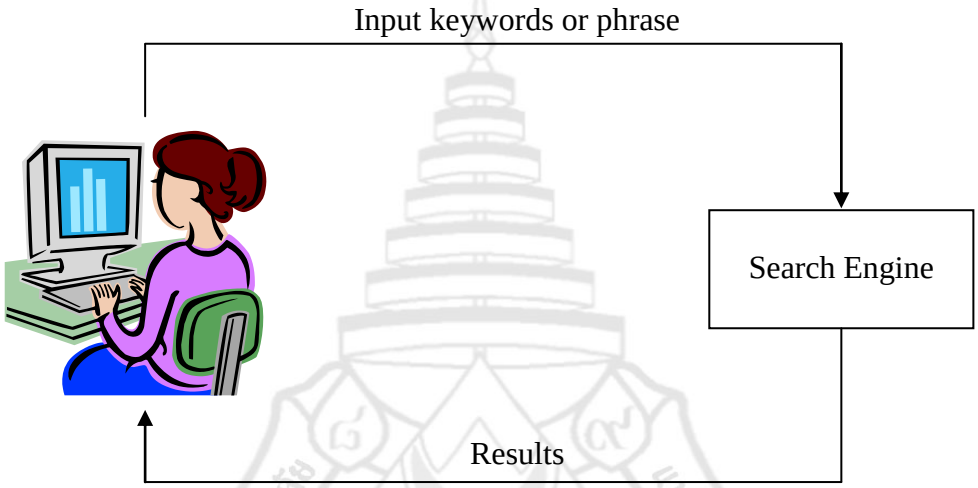

Figure 1.1 The Basic use of Search Engine

In using search engine for retrieving desired information, normally input keyword and/or key phrase in natural language such as Thai, English, Spanish, Japanese, and so on into search engine in order to extract the related information. In case that user inputs keywords, the search engine basically finds websites containing input keywords as related ones, and ranks them by using the order of appearance frequency. This assists search engine users to easily access to the desired websites. However, for key phrase which is radically written in natural language, it would find the related ones without understanding the meaning of the language. For instance, "แนะนำที่พักใกล้ถนนคนเดินเชียงใหม่ราคา 500-1000 (Recommended accommodation near Chiang Mai walking-street with the price 500-1000)" as shown in Figure 1.2.

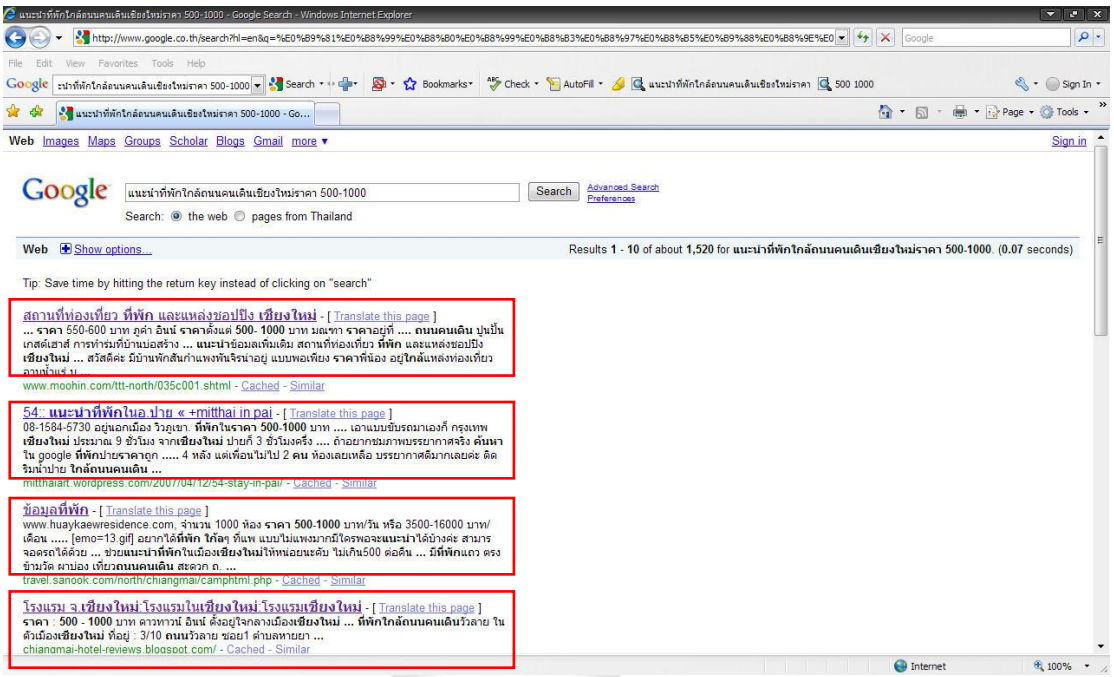

Figure 1.2 The Results of "แนะนำที่พักใกล้ถนนคนเดินเชียงใหม่ราคา 500-1000 (Recommended accommodation near Chiang Mai walking-street with the price 500-1000)"

Figure 1.2 shows the results which do not match the user's query. The websites related to recommendation as keywords are retrieved, but they are usually limited to find only accommodation with 500-1000, not the numbers of 500, 600, 700, etc. to 1000. They match only the format is 500-100 not the semantic of query though there are those with price 500-1000, which the user would like to have more chance to choose one. The normal search engine normally fails to provide the alternatives.

However, some language such as Thai and so on which have no symbol of border between neighboring words in the phrase are difficult for search engine to segment the words. To obtain the appropriate websites, the keywords with operations according to search engine regulation are required as search engine input. People generally prefer key phrase in natural language as search engine input to keywords with operations without learning search engine regulations. In order to make the search engine friendly for general users, the function of Thai natural-language key phrase and converting keywords into search engine format is needed to implement, as shown in Figure 1.3.

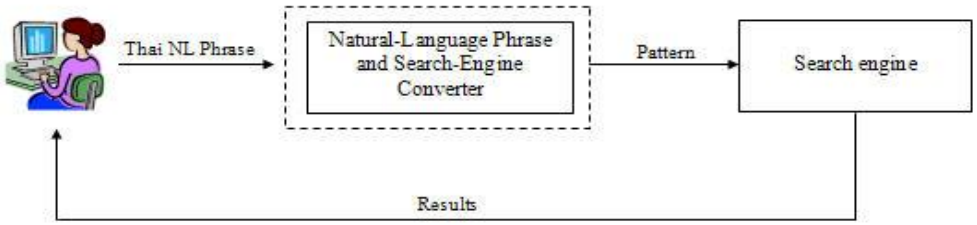

Figure 1.3 Thai Natural Language Search Engine using Synonym Approach

Therefore, this thesis proposes the method used to convert Thai natural language phrase into keyword pattern for search engine in order to list up the desired results that most match to the query. The method extracts from phrase and converts it into keyword patterns. Then, it converts to Google syntax so that the appropriate alternative search is obtained. The keyword pattern includes nouns, synonyms, and Boolean operators.

1.2 Objectives

The objectives of this thesis are to design and develop a system for Thai natural language search engine using synonym approach and converting to Google syntax. The keyword pattern includes nouns, synonyms, and Boolean operators. The system aims at solving problem of user that there are mis-matched results by converting Thai natural language phrase into search engine keywords pattern in order to list up the desired results that most match to the query.

1.3 Scopes

The proposed system supports only Thai natural language. The patterns length is not exceed 32 words. The proposed system is tested with tourism in the northern part of Thailand (Chiang Mai, Chiang Rai, Lamphun, Lampang, Mae Hong Son, Phayao, Phrae, and Nan).

1.4 Definition and Terminology

1.4.1 Search Engine is a tool and/or program designed to search information on the World Wide Web. A search engine works by sending robots (some call spider or crawler) to look for documents and their web addresses and then send to indexing software. Indexer software stores documents and web addresses in database, and the next is search engine, which supports query and shows the results.

1.4.2 Natural Language Processing is a field of computer science and linguistics concerned with the interactions between computers and human. Natural language processing is a very attractive method of human computer interaction.

1.4.3 Boolean and Operator is the symbol used to help for specific searching. The Boolean and operator used for searching are such as AND, OR, NOT, +, -, ".." and so on.

1.4.4 Synonym is different words with identical or similar meaning.

1.5 Expectations of This Thesis

This thesis proposes system for Thai natural language search engine using synonym approach and converting to Google syntax. The keyword pattern includes nouns, synonyms, and Boolean operators. The keyword pattern can help users who do not have knowledge or understand how to use Boolean operators for searching and to get results that match with their query. This system is developed to use with Thai natural language for asking about tourism in the northern part of Thailand (Chiang Mai, Chiang Rai, Lamphun, Lampang, Mae Hong Son, Phayao, Phrae, and Nan). Therefore, this system can be applied to other query and other language, but the part of word segmentation must be suitably improved for each language.

CHAPTER 2

LITERATURE REVIEW AND RELATED WORKS

2.1 Related Works

2.1.1 Query expansion

The query expansion is the process to expand query by using Boolean operator. The objectives of query expansion are to expand results and increase the accuracy of the results. Query expansion involves the techniques such as finding synonym; finding all the various morphological forms of words by stemming each word in the search query, etc. There are 2 techniques for query expansion. The first technique used Machine Readable Dictionary (MRD) (Ellen, 1994) and WordNet (Miller, 1995). Another technique used words from related document. The research of Buckley, Saltion, Allan, and Singhal (1994) used words from a document that's related with main query. Ishikawa, Satoh, and Okumura (1997) used words from paragraph that's related with main query, by using the similarity of meaning between paragraphs.

Salton and Lesk (1971) had studied the using of dictionary for a long time. Allen (1997) and Ellen (1994) used machine readable dictionary for query expansion by using synonym. The research of Rujira and Yuen (2003) used conceptual knowledge for keyword expansion and used Boolean OR for group words. The keywords used are the synonyms for the expansion of query. Hayder et al. (2006) studied the improvement of Arabic search engine by using search key expansion. They used synonym from Arabic WordNet to query expansion.

The technique used Machine Readable Dictionary and WordNet, and the technique used keywords from document or paragraph for indexing term by collect noun, verb, adverb, adjective, etc. The disadvantage of both techniques is that they can use only keywords search which cannot

use with natural language query and are not interested in the meaning of words and phrase. The natural language such as Thai and so on is difficult for search engine to understand and list up the desired results that are most matched to the query. This is because that language does not have symbol of border between neighboring words in the phrase. Therefore, firstly, the words must be segmented into keywords and take keywords for searching and then use Boolean operators for query expansion. The query expansion is a common process to expand query by using Boolean operators. These theses therefore aim at expanding Thai natural language query by using synonyms keyword and Boolean operators. Firstly, the expand natural query must be segmented into keywords, and then group the keywords and synonyms keyword by using Boolean operators.

2.2 Computational Theory

2.2.1 Natural Language Processing

Natural language processing (NLP) is the artificial intelligence and linguistics to study about problems in processing and use of natural language and to understand natural language, by computer can understand natural language. The main of research about Natural Language Processing (NLP) want to connect between computer and human by use Natural Language (NL). Natural language systems convert samples of human language into more formal representations. Natural language processing overlaps with the field of computational linguistics and is often considered a sub-field of artificial intelligence. The term natural language is used to distinguish human language from formal or computer language. Natural language processing (NLP) is a very attractive method of human computer interaction (Liddy, E.D., 2001). In natural language, when write or speak and put the word together, it will have meaning. So, computer must have the processes for understanding of sentence or phrase, the elements as following (The Process of Natural Language Processing, online)

Syntactic Analysis: will check the grammatical structures positioning of noun, verb, prepositions etc. by combine to sentence.

Semantic Analysis: will identify the correct meaning of sentence. The sentence written correctly with grammatical structure will have certain meaning, buy sometime the sentence will have ambiguous meaning or no meaning.

The process of NLP starts by taking the input to check the structure of sentence, it is called “Phrase Tree”, and then checks the meaning and relation between each part of sentence. The analysis in structure and meaning must have electronic dictionary containing keyword, and each keyword should have complete information about meaning and relationship of words. The analysis of NLP in computer will write in form of facts and rules by artificial intelligence process as checking program. The output after analysis of the correct sentence structures and grammar will be changed or translated into command for database or structure of desired language. How to create a command or to create sentences must follow the rules or the main goal by considering the structure and meaning of language.

Thai language is written left-to-right. Thai language has no symbol of border between neighboring words in the phrase. So it is difficult for search engine to segment the words (Hugh et al., 2009) and understand Thai language. The search engine will match with words. Therefore, Thai language must segment phrase into keywords. The word segmentation will be explained in the next section.

2.2.2 Word Segmentation

Word segmentation is the process of dividing written text into meaningful units. Thai word segmentation is important for Thai language. If word segmentation has mistake, it will affect the meaning of words. There are 3 main types of word segmentation, i.e., Rule based approach, Dictionary approach, and Corpus based approach (Charoenpornsawat, 1998).

The rule based approach creates a rule based Thai grammar. Yupin (1981) created rule based for Thai word segmentation by considering characters appeared in syllable or word, which can be divided into 5 groups: consonant; vowel; tone mark; numeral; and special character. The research of Surin (1983) used rule based approach from Thai grammar same as Yupin, and analyses character of Thai syllable. The rule based approach can be separated into 2 types: front boundary recognition rule and tail boundary recognition rule.

The dictionary approach, Yuen Poovarawan and Vivon Amarom (1986) collected all syllables in dictionary and get 18 rule based to help in the case that it cannot find a syllable in dictionary. Duangkaew Sawamipak (1990) created rule based and used with dictionary approach. The reason to mix rule based and dictionary approach is to solve problems caused by only used dictionary approach. Samphan Raruenrom (1991) collected words in dictionary instead of syllable. Virach Sornlertlamvanich (1993) developed word segmentation by Maximal Matching approach.

The research of Kawtrakul et al. (1997) used statistic and Trigram model to solve problem of word segment, and indicate role and meaning of words. Meknavin, Charoenpornsawat, and Kijsirikul (1997) used statistic by considering continuation of word. For word segmentation, the probability of sentence was taken into consideration.

The word segmentation is very important for language with no symbol of border between neighboring words in the phrase. The word segmentation for this thesis used dictionary approach. The dictionary approach is to collect all words such as noun, verb, adverb, adjective etc. The dictionary approach is suitable for this thesis because it's will compare word with database. The database will collect words and the principle of comparison short words for word segmentation.

2.2.3 Synonym

A synonym is one or two or more words, which have the same meaning (Illinois Mathematics and Science Academy, 2006). The synonym can be used for searching. If the first search query doesn't retrieve what the users are looking for, so they should think of synonyms of the terms in their query. Each synonym might have the same meaning. The search engines determine relevancy by matching the keywords in search query. The words will appear in the documents they have indexed. Synonym will make chances for a search engine to find other keywords that the users are looking for. Another way to find strong synonyms is to use a thesaurus. For search engine, it uses WordNet (Diab, 2004) to help finding synonym. The WordNet is an electronic lexical database developed to be as monolingual English Language online lexical resource by George Miller, Princeton University. WordNet comprises four parts of speed databases corresponding to nouns, verbs, adjectives, and adverbs.

2.2.4 Boolean Logic

The Boolean model is a simple retrieval model. The Boolean is the expansion of query by using AND, OR, and NOT. Each Boolean and Operator can be explained (Tanarangrak & Monsanit, 2006) as follows:

1. AND: The query (a1 AND a2) is used between the keywords to focus the only intended keywords.

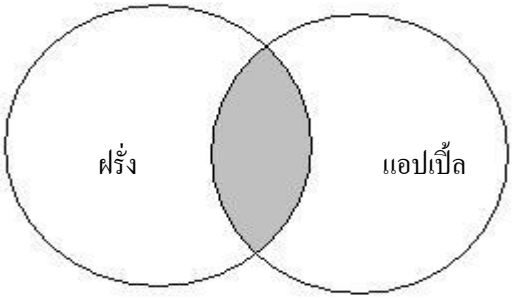

Figure 2.1 The Retrieved Records in Which BOTH of the Search Terms are Present in using Boolean "AND"

The Boolean "AND" is used between keywords, for example, ฝรั่ง AND แอปเปิ้ล. This Boolean will search websites that contain keywords, ฝรั่ง and แอปเปิ้ล by disregarding where the keywords appear in any part of websites.

2. OR: The query (a1 OR a2) is used between keywords to select all documents which satisfy a1 or a2.

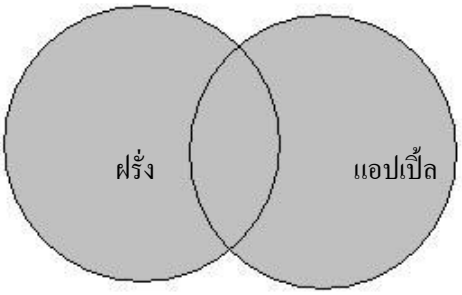

Figure 2.2 The Retrieved Records in Which AT LEAST ONE of the Search Terms are Present in using Boolean "OR"

The Boolean “OR” is used between keywords, for example ฝรั่ง OR แอปเปิ้ล. This Boolean will search websites that contain any keyword in a websites. From Figure 2.2, searching ฝรั่ง OR แอปเปิ้ล, the results will show the websites in that one or another word appears.

3. NOT: The query (a1 NOT a2) is used between keywords to select all documents which satisfy a1 but not a2.

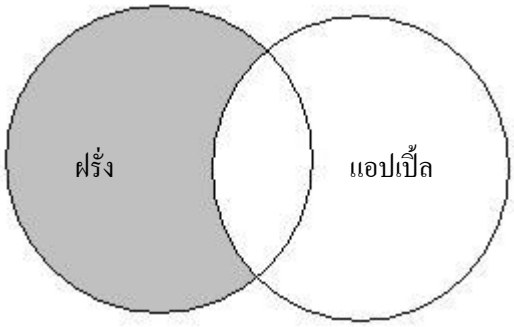

Figure 2.3 The Retrieved Records in Which ONLY ONE of the Search Terms are Present in using Boolean “NOT”

The Boolean “NOT” is used before a keyword that the users do not want to search. For example, ฝรั่ง NOT แอปเปิ้ล. From the example, the results will retrieve all ฝรั่ง but do not retrieve fruit from แอปเปิ้ล.

The Boolean model is easy to use because it is easy to understand Boolean logic (Internet Tutorial, online). The Boolean model is not complex and not flexible because its expansions are only AND, OR, and NOT. This model will retrieve specific result. Normally, the Boolean is used with keyword search. The expansion queries used the most is the synonyms. The Boolean model predicts that each document is either relevant or non relevant.

The principle of using Boolean operators for this is that the main keyword will use “AND” between keywords, but between keywords and synonyms keyword, it will use “OR”. Boolean “AND” will search by disregarding where the keywords appear in any part of websites. For Boolean “OR”, the results will show the websites in that one or another word appears.

2.2.5 Google Logic

Google search is a web search engine. Google receives several hundred million queries each day through its various services. Google search was originally developed by Larry Page and Sergey Brin in 1997. Google is the most popular web search engine. Google search provides more features such as: news, maps, weather forecasts, synonyms, sports scores, etc. Google is available in many languages and has been localized for many countries in their own languages, such as Thai, Chinese, French, Japanese, Turkish, etc. Google search also supports general and advanced search.

General search user inputs keyword and/or key phrase and uses Boolean or operators. The Boolean search can use AND, OR, and NOT. The operator search can use " + ", " – ", " * ", " ~ ", and " .. " (Tanarangrak & Monsanit, 2006).

1. The operator " + " : this operator is typically used in front of stop words that Google would ignore. For example [+Two +on +a Tower], can be used instead of " + " operator before each stop word in the query, put ALL the terms including stop words within double quotes. Querying this way will return similar results.

2. The operator " – " : this operator excludes unwanted word in the resulting pages. For example, [Apple -Computer], the results will show everything about Apple but do not show that Apple is computer.

3. The operator " * " : this operator matches one or more words in the query. Google treats the " * " as a placeholder for a word or more than one word. For example, 1: [ministry*] the results will show the first word is ministry and the words after ministry can be any.

4. The operator " ~ " : this operator is called the fuzzy operator or synonym operator. It searches for page that contains the specified term as well as synonyms for the term. For example, [~generous] would return pages on which the word generous appears, as well as pages on which the word unselfish appears.

5. The operator " .. " : this operator is for range of number. For example, [Price 500..1000], the example returns the pages on price between 500 and 1000.

This thesis tests the developed keyword pattern with Google search engine because Google search is very popular the most people use it. Search of Google will have general and special search. In the general search, user inputs keywords in tool box and then Google will show results. The special search user input keywords and Boolean operators such as ฝรั่ง AND แอปเปิ้ล , ฝรั่ง OR แอปเปิ้ล etc. in tool box and then Google will show results. To use the special search user must know about Boolean operators of Google search. User must to know and understand how to use each Boolean operator, sometime not easy for people to understand how to use. Thus, this thesis test use general search and special search make to know if used special search results are correct with demand or match with query more than use normal search.

2.3 Proposed Method

This thesis proposes the method used to convert Thai natural language phrase into keywords pattern for search engine in order to expand the chance for the right search in the form of natural language by listing up the desired websites as the best ones. The method extracts user intention from the phrase, and converts it into keywords pattern. Then, it is converted to Google syntax. The keyword pattern proposed here includes nouns, synonyms, and Boolean operators. The detail of keyword pattern is shown in Chapter 3. Figure 2.4 show the basic system of this thesis by user using inputs Thai keyword and/or Thai phrase, the next process is to covert Thai NL phrase into Google syntax, and after that send the group words of Google syntax to Google search. Google search will return results to user.

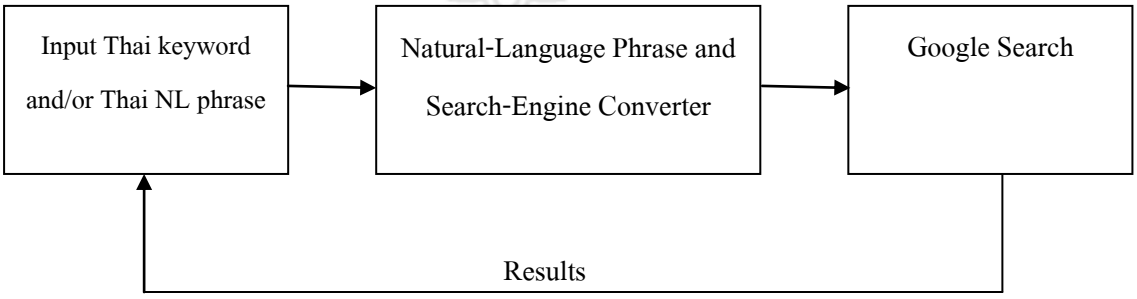

Figure 2.4 The System of Thai Natural Language Search Engine

CHAPTER 3

METHODOLOGY

This chapter describes methodology for Thai natural language search engine using synonym. This methodology is separated into 2 steps as follows:

1. Analysis and Design Phase
2. System Development Phase

3.1 Analysis and Design Phase

This section describes the system overview and detail of each process of the proposed system.

3.1.1 System Overview

This system translates Thai natural language key phrase into Google search engine format. The sequence of processes is set up as shown in Figure 3.1. There are four parts, including 1) keywords or NL phrase input, 2) word segmentation, 3) Thai natural language to Google syntax conversion, and 4) searching (Google). For the processes surrounded by the dash line in Figure 3.1, This thesis propose search synonyms of these noun keywords with Thai noun keyword are determined using synonym keyword database at A, and then the weights of all synonyms are computed at B, and keyword pattern is setup by using keyword pattern rules at C. The database for this system collected Thai noun keywords and synonyms about tourism in the northern part of Thailand (Chiang Mai, Chiang Rai, Lamphun, Lampang, Mae Hong Son, Phayao, Phrae, and Nan) such as accommodation, places, hotel etc.

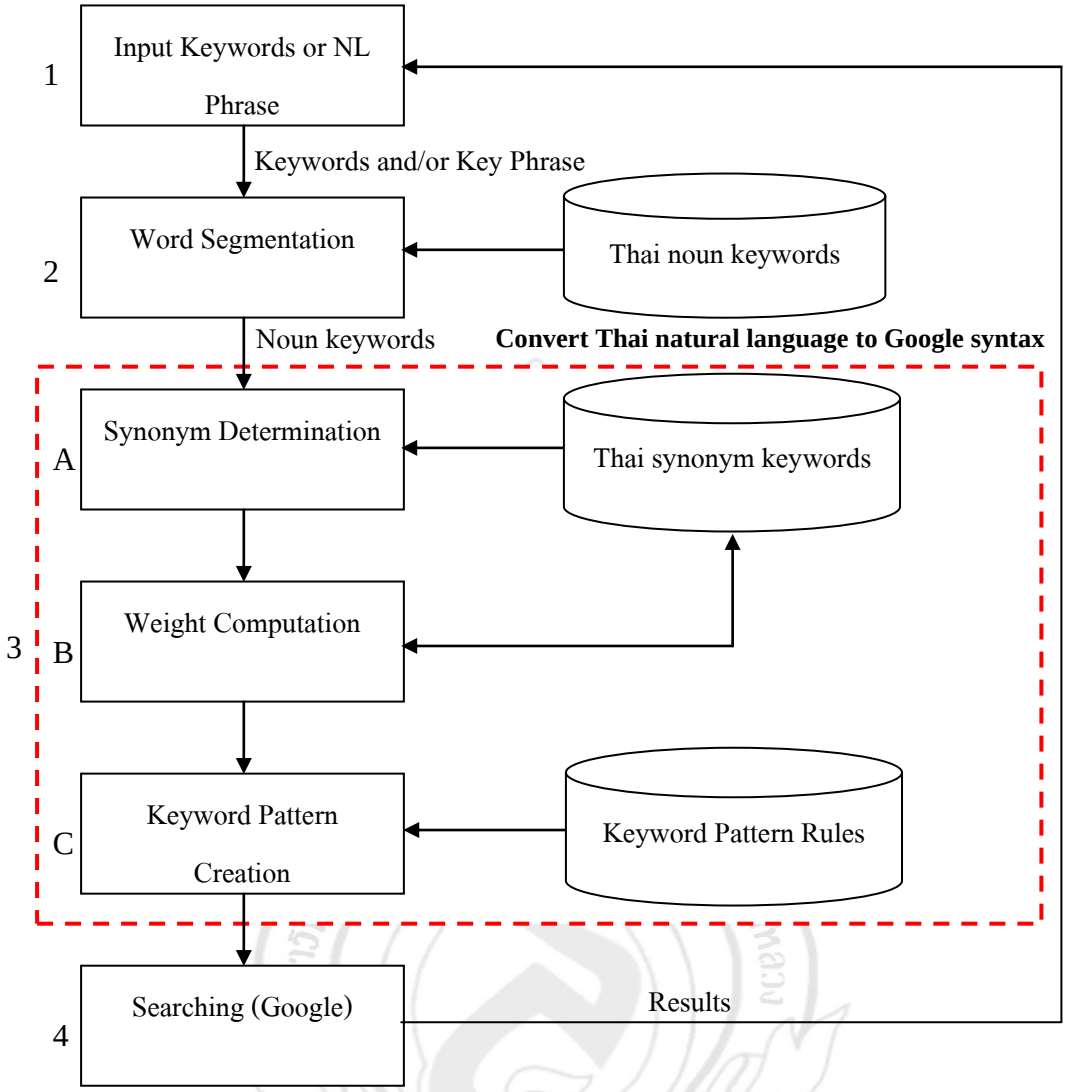

Figure 3.1 The System Overview

3.1.2 Input Keywords or NL Phrase

In this process, user inputs Thai natural language query. The queries are about tourism in the northern part of Thailand (Chiang Mai, Chiang Rai, Lamphun, Lampang, Mae Hong Son, Phayao, Phrae, and Nan). For example, Thai natural language phrase input by user is "แนะนำที่พักใกล้ถนนคนเดินเชียงใหม่ราคา 500-1000 (Recommended accommodation near Chiang Mai walking-street at the price between 500-1000)".

3.1.3 Word Segmentation

This process is to separate Thai natural language into words. This process uses Dictionary approach and the word is chosen by comparing to the possible shortest word in the database. From example phrase is “แนะนำที่พักเชียงใหม่ (Recommended accommodation Chiang Mai)”. This process will separate the phrase into characters as follows:

แ	น	ะ	น	ำ	ท	ี	่	พ	ั	ก	เ	ช	ก	ย	ง	ใ	ห	ม	่

From above phrase can separate word into characters are 20 characters. This process will take each character compare with database until match with word in database. Firstly, first character is “แ” compared with all words in database. If it does not match with any word in database, it will take the next character to combine with the last character until the word is matched with any word in database. The first not have a word matched in database. The second will start the next character is “น” for compare with database until match with word in database. Thus, in each round of segmentation will start a next character. From above phrase can separate 20 characters, so will segmentation 20 times. The first word matched in database is ที่พัก (Accommodation). After matching a word, it will take the next character to be the first character to compare with database again until the last character. In this case, it is “ใ”. Then follow the same process as the above until the last character of phrase.

3.1.4 Synonym Determination

A group of words related to nouns obtained from word segmentation is searched in Thai synonyms database to identify the synonym candidates, as shown in Figure 3.2.

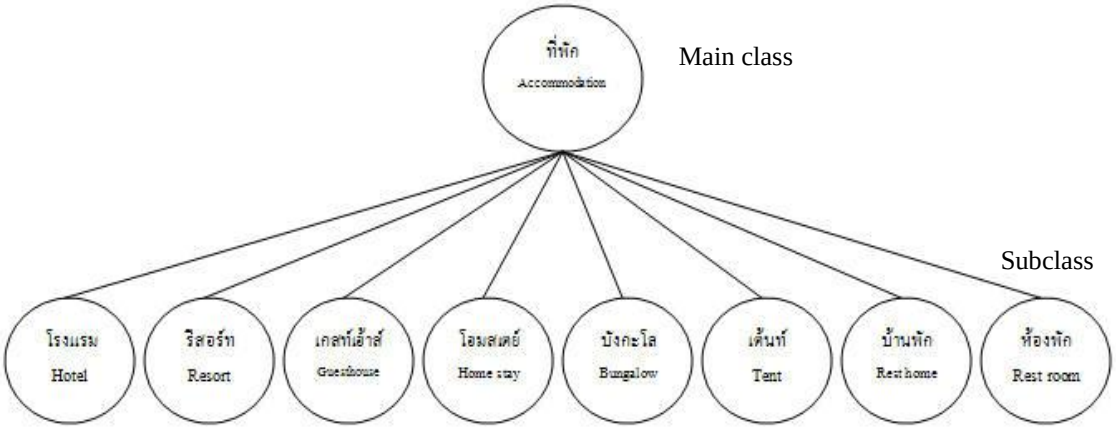

Figure 3.2 The Synonym Candidate

From Figure 3.2, the synonym candidates of "ที่พัก (Accommodation)" are โรงแรม (Hotel), รีสอร์ท (Resort), เกสท์เฮ้าส์ (Guesthouse), โฮมสเตย์ (Homestay), บังกะโล (Bungalow), เต้นท์ (Tent), บ้านพัก (Villa), and ห้องพัก (Lodge). Words and synonyms keywords were collected from LEXiTRON dictionary (NECTEC : online), and Thai homonym dictionary (Yuen, et al., 2000) to create database. After collecting the words by considering the wideness or narrowness of the words meanings for ranking, the words that cover the meaning of other words will be a main class. Other words within a category that has narrower meaning than the main class is called subclass, and the subclass must have relation within the group. From Figure 3.2, the main class is "ที่พัก (Accommodation)", the subclass is โรงแรม (Hotel), รีสอร์ท (Resort), เกสท์เฮ้าส์ (Guesthouse), โฮมสเตย์ (Homestay), บังกะโล (Bungalow), เต้นท์ (Tent), บ้านพัก (Villa), and ห้องพัก (Lodge). In subclass, it can have many words.

3.1.5 Weight Computation

This process is to create the weight of each synonym are considered. Input each synonym in Google search, then select the pages from Thailand and look at the numbers, as shown in Figure 3.3.

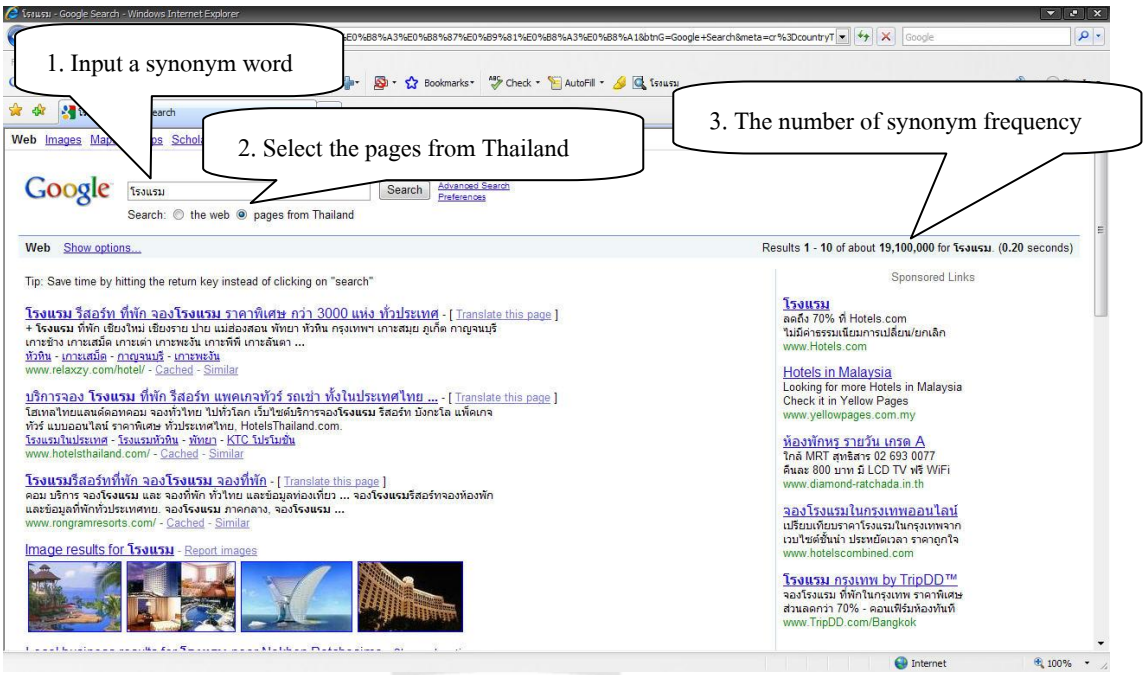

Figure 3.3 The Step to find Synonym Frequency

The calculation of synonym weight is as equation (3.1)

$$\text{Weight of synonym} = \frac{n_i}{\sum n_i} \times 100 \qquad (3.1)$$

Where n_i is a number of synonym, $\sum n_i$ is the total number of synonym.

Noun keyword	Synonyms keyword	Real frequency of synonym	Weight of synonym
ที่พัก	โรงแรม	8860000	38.6
	ห้องพัก	4960000	21.61
	รีสอร์ท	4480000	19.51
	บ้านพัก	2080000	9.06
	เต้นท์	1770000	7.71
	โฮมสเตย์	372000	1.62
	เกสท์เฮ้าส์	219000	0.95
	บังกะโล	210000	0.91
Sum		22951000	99.97

Figure 3.4 The Calculation of Synonym Weight

Table 3.1 The Weight of Synonym

ที่พัก / Synonym	Weight of synonym
โรงแรม (Hotel)	38.6
ห้องพัก (Lodge)	21.61
รีสอร์ท (Resort)	19.51
บ้านพัก (Villa)	9.06
เต้นท์ (Tent)	7.71
โฮมสเตย์ (Homestay)	1.62
เกสท์เฮ้าส์ (Guesthouse)	0.95
บังกะโล (Bungalow)	0.91

Figure 3.4 shows how to calculate weight of synonyms keyword. The eight synonyms keywords are โรงแรม (Hotel), ห้องพัก (Lodge), รีสอร์ท (Resort), บ้านพัก (Villa), เต้นท์ (Tent), โฮมสเตย์ (Homestay), เกสท์เฮ้าส์ (Guesthouse), and บังกะโล (Bungalow). The Figure shows the real frequency and weight of the synonyms. The weight of synonym can be calculated by equation 3.1. From table 3.1, the weight can be computed as shown in Figure 3.4. This thesis used first five rank maximum weights as the representative search keywords. From Figure 3.4, โรงแรม (Hotel), ห้องพัก (Lodge), รีสอร์ท (Resort), บ้านพัก (Villa), and เต้นท์ (Tent) are selected.

ที่พัก	(Accommodation)	โรงแรม	(Hotle)	38.6
		ห้องพัก	(Lodge)	21.61
		รีสอร์ท	(Resort)	19.51
		บ้านพัก	(Villa)	9.06
		เต้นท์	(Tent)	7.71
		โฮมสเตย์	(Homestay)	1.62
		เกสท์เฮ้าส์	(Guesthouse)	0.95
		บังกะโล	(Bungalow)	0.91

Figure 3.5 The First Five Rank Maximum of Synonym Keywords

3.1.6 Keyword Pattern Creation

In order to create the search pattern, the operations are added into the synonym keywords. This research uses three basic Boolean operators as follows:

1. “AND” is used between the keywords to focus only the intended keywords.
2. “OR” is used between keywords and their synonyms to increase the alternatives.
3. “..” is used for searching the range of numbers, for example “x1..x2” means any value between x1 and x2.

3.1.7 Searching with Google

This process is a process of Google search. After grouping the keywords by using Boolean operators from keyword pattern creation, send them to Google search. From above input example, the keywords group sent to Google search is [ที่พัก OR โรงแรม OR ห้องพัก OR รีสอร์ท OR บ้านพัก OR เต้นท์] [ถนนคนเดิน] [เชียงใหม่] [ราคา] 500..1000 [Accommodation OR Hotel OR Lodge OR Resort OR Villa OR Tent] [Walking-Street] [Chiang Mai] [Price] 500..1000.

3.2 Implementation Phase

This section describes about system requirements and web implementation of Thai natural language search system.

3.2.1 System Requirements

This thesis proposes the minimum requirements for both hardware and software as follows.

1. Hardware

1) CPU: Intel (R) Core(TM) 2 2.00GHz
2) Ram: 1.00 GB
3) Hard disk: 120 GB
4) Monitor: Aspire 5050
5) Peripheral: Keyboard, mouse USB

2. Software

1) Operating system: Microsoft windows XP Professional

2) System development: Macromedia Dreamweaver 8, Navicat for MySQL, Apache2.2 server and PHP 5

3.2.2 Web implementation

These sections describe the detail of web implementation as follows:

1. Database Design

We design database for this system by separating into 2 tables and rules: noun table; synonym table and keyword pattern rules. The details are as follows:

1) Noun Table

Noun table is a table that contains noun keywords. This table collects noun keywords from tourism in the northern part of Thailand. It composes of 4 attributes including id, id_noun, word, and number. The detail of noun table is shown in Table 3.2.

Table 3.2 The Detail of Noun Table

Attribute Name	Type	Primary Key
Id	Int	Yes
id_noun	Int	No
Word	Varchar	No
Number	Varchar	No

From Table 3.2, id attribute is the primary key and runs automatic number. The id_noun confine length is 5. The word attribute is the Thai noun keyword about tourism. The last attribute is number (e.g. 1, 2, 3, and etc.). Figure 3.5 shows the details of Noun Table.

id	id_noun	word	number
1	1001	เรื่อง	(Null)
2	1002	ที่พัก	(Null)
3	1003	ตัวเมือง	(Null)
4	1004	การเดินทาง	(Null)
5	1005	รถเช่า	(Null)
6	1006	สถานีรถไฟ	(Null)
7	1007	สนามบิน	(Null)
8	1008	ล่องแก่ง	(Null)
9	1009	บริษัททัวร์	(Null)
10	1010	เส้นทาง	(Null)
11	1011	ฮันนีมูน	(Null)
12	1012	อุทยาน	(Null)
13	1013	สถานที่ท่องเที่ยว	(Null)
14	1014	ราคา	(Null)
15	1015	คำแนะนำ	(Null)
16	1016	จังหวัด	(Null)
17	1017	อำเภอ	(Null)
18	1018	ตำบล	(Null)
19	1019	แถว	(Null)
20	1020	ร้านอาหาร	(Null)
21	1021	วิธี	(Null)
22	1022	สายการบิน	(Null)
23	1023	เชียงใหม่	(Null)
24	1024	เชียงราย	(Null)
25	1025	ลำพูน	(Null)
26	1026	ลำปาง	(Null)
27	1027	แพร่	(Null)

Figure 3.6 The Details in Noun Table

2) Synonym Table

Synonym table is a table that contains synonym keywords. This table collects synonym keywords. The synonym keywords must relate with noun keywords. This table composes of 4 attributes including id_syn, id_noun, word, and frequency. The details of table synonym is shown in Table 3.3

Table 3.3 The Details of Synonym Table

Attribute Name	Type	Primary Key
id_syn	Int	Yes
id_noun	Int	No
Word	Varchar	No
Frequency	Varchar	No

From Table 3.3, id_syn attribute is the primary key and runs automatic number. The id_noun is referring from noun table in attribute id_noun. The word is the Thai synonym keywords that have the same meaning with words from noun table. The last attribute is weight, which keeps the weight of synonyms keyword obtained from weight computation process. The words in this table are only synonym keywords. The synonym keywords must relate with noun keyword. For example "ที่พัก" have 8 synonyms, which are, โรงแรม (Hotel), ห้องพัก (Lodge), รีสอร์ท (Resort), บ้านพัก (Villa), and เต้นท์ (Tent), โฮมสเตย์ (Homestay), เกสท์เฮ้าส์ (Guesthouse), and บังกะโล (Bungalow). The noun keyword "ที่พัก" has id_noun 1002. The detail of Synonym Table is shown in Figure 3.6.

id_syn	id_noun	word	frequency
1	1002	โรงแรม	46
2	1002	รีสอร์ท	18
3	1002	เกสท์เฮ้าส์	1
4	1002	โฮมสเตย์	2
5	1002	บังกะโล	1
6	1002	บ้านพัก	9
9	1002	เต้นท์	4
10	1002	ห้องพัก	19
13	1003	ในเมือง	8
14	1003	ใจกลางเมือง	4
15	1003	ชุมชน	3

Figure 3.7 The Details in Synonym Table

3) Keyword Pattern Rules

The keyword pattern rules is principle use Boolean operators. This thesis used three Boolean operators are AND, OR, "..". The Boolean operators "AND" used between keywords such as ที่พัก AND เชียงใหม่. The Boolean operators "OR" used between keyword and synonyms keyword such as ที่พัก OR โรงแรม OR รีสอร์ท OR เกสท์เฮ้าส์. The Boolean operators ".." used between number such as 500..1000.

2. Web User Interface

The web user interface is the function to get Thai natural language phrase from user. User inputs keyword and/or key phrase through the web shown in Figure 3.7.

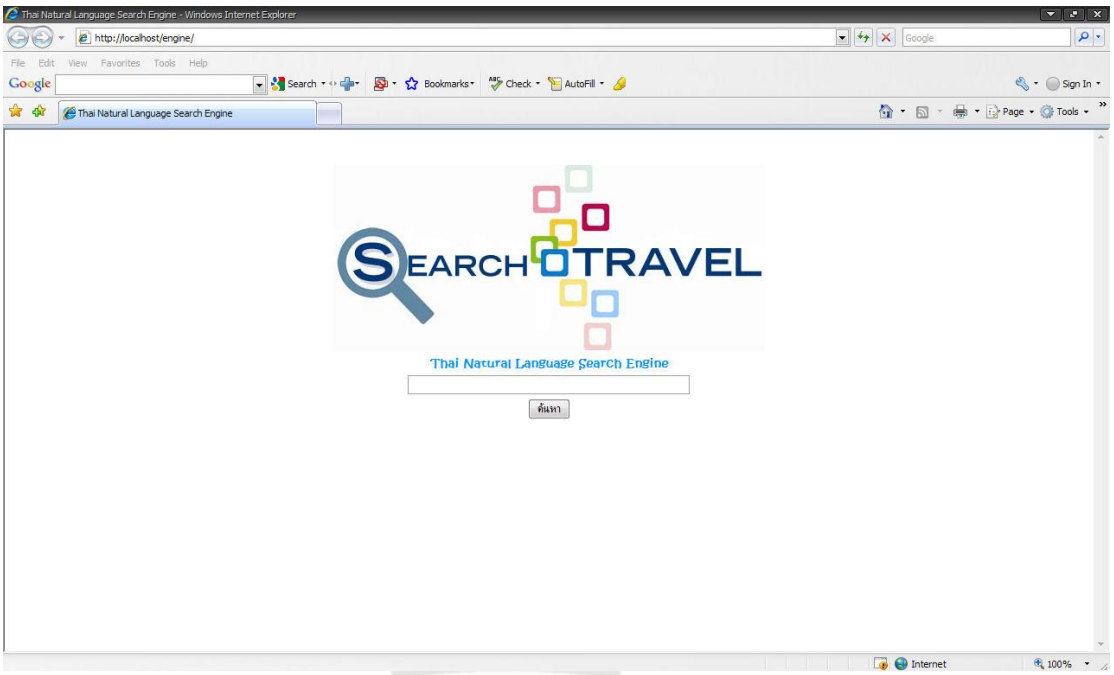

Figure 3.8 The Web User Interface

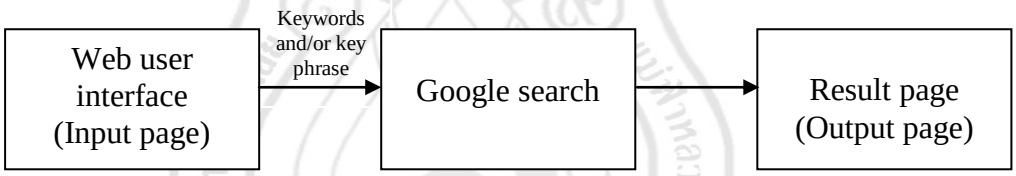

Figure 3.9 The System Overview Contact with Google Search

Figure 3.8 show the system overview how to contact with Google search. Web user interface is that the user inputs keywords and/or key phrase, and then web user interface sends that input to Google search by written code, and after that Google will search and show the results on results page. The detail all about this processes as follows.

1) Web user interface (input page)

This process is to design an input page by user for easy use. User inputs keywords and/or key phrase in tool box as shown in Figure 3.7, and after that will forward keywords and/or key phrase into Google search

2) Google search

This process is connecting between web user interface (input page) and result pages (output page). The input page can connect with Google search by writing this code in the program. The code to contact between input page and output page is as follows:

"http://www.google.com/custom?q=$groupWord&client=pub-0585777730843386&forid=1&ie=tis-620&oe=tis-620&cof=GALT:#008000;GL:1;DIV:#336699;VLC:663399;AH:center;BGC:FFFFFF;LBGC:336699;ALC:0000FF;LC:0000FF;T:000000;GFNT:0000FF;GIMP:0000FF;FORID:1;&hl=en";

3) Result page (output page)

This process is to design an output page. All the results from Google will be shown in this page as in Figure 3.9.

3. Result page

The result page shows the results from the search of query. From Figure 3.9, A is a box for query in the form of Thai natural language where user inputs keyword and/or key phrase. B is a box that shows a group of keywords and can use tool box for searching without search pattern (Google search). C shows the results from the search using search pattern and without search pattern.

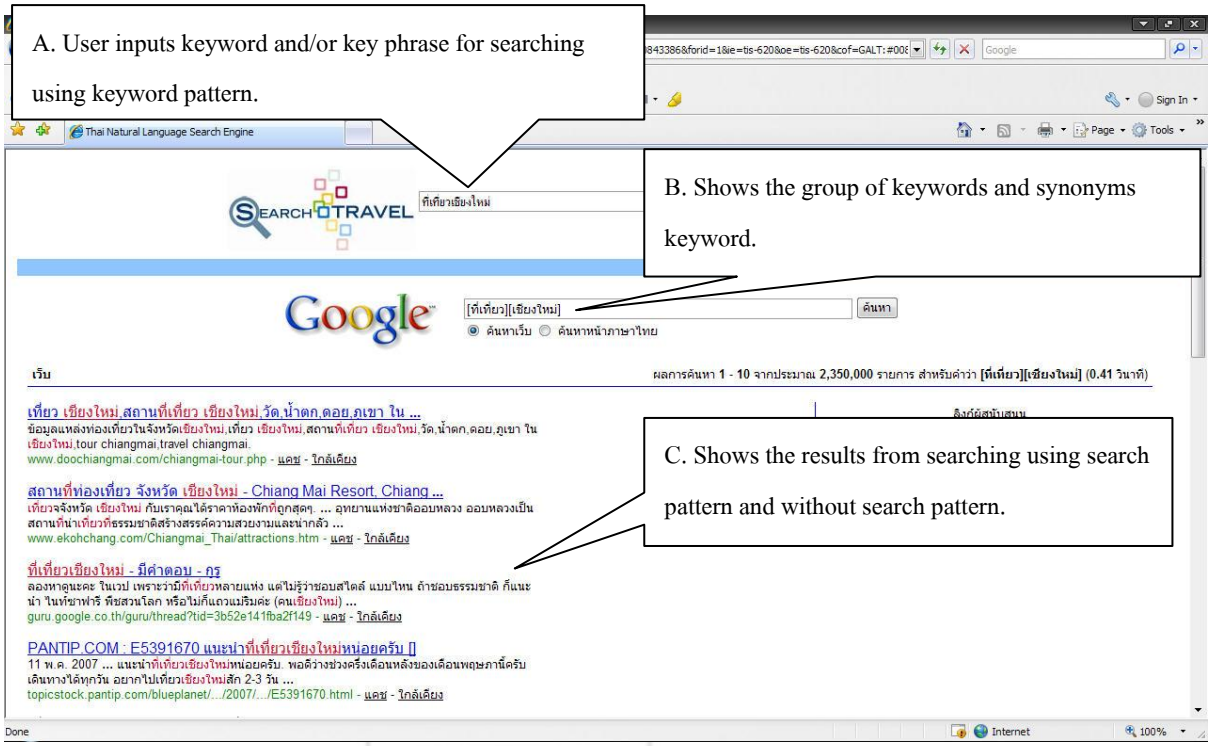

Figure 3.10 The Page Showing the Results

CHAPTER 4

EXPERIMENT AND RESULT

This chapter describes the experiment and result of Thai Natural language search engine using synonym and Google syntax. This chapter is separated into three sections: web user interface; experiment detail; and experimental result.

4.1 Web User Interface

This system creates web user interface by user's input of Thai natural language for searching appropriate website. This web user interface is a part of communication between user and system, and displays results, as shown in Chapter 3 in the Figures 3.7 and 3.9, respectively.

4.2 Experiment Detail

The experiment is set in order to compare the searched result between the search with keyword pattern and without keyword pattern. The details of experiment are described in this section.

4.2.1 Query Phrase

The query tested in this thesis is only Thai natural language and the length is not exceed 32 words. The queries ask about tourism in the northern part of Thailand (Chiang Mai, Chiang Rai, Lamphun, Lampang, Mae Hong Son, Phayao, Phrae, and Nan) such as accommodation, tourist attraction, journey, etc. This thesis used 95 queries for testing.

4.2.2 Testing and Result

The testing used Google search to compare between searches using keyword pattern and without keyword pattern. This thesis evaluation by used the first five pages results.

1. Search using keyword pattern

User inputs Thai natural language such as "แนะนำที่พักใกล้ถนนคนเดินเชียงใหม่ ราคา 500-1000 (Recommended accommodation near Chiang Mai walking-street with the price 500-1000)", as shown in Figure 4.3. After word segmentation as the followings:

Thai nouns: ที่พัก ถนนคนเดิน เชียงใหม่ ราคา 500-1000

English nouns: Accommodation Walking-Street Chiang Mai Price 500-1000

After the process of word segmentation, each noun keyword is taken to compare with database for synonym keyword, for example the synonyms candidate of "ที่พัก (Accommodation)" From Figure 3.2, many synonyms candidate of "ที่พัก (Accommodation)" are shown. This thesis uses the first five maximum weights, as shown in Figure 3.4. The noun keywords above can be separated as follows:

"ที่พัก (accommodation)", the synonym keywords are โรงแรม (hotel), ห้องพัก (lodge), รีสอร์ท (resort), บ้านพัก (villa), and เต้นท์ (tent).

"ถนนคนเดิน (walking-street)"

"เชียงใหม่ (Chiang Mai)"

"ราคา (price)"

"500-1000"

After the process of synonym determination and weight computation, words are grouped by using Boolean operators of Google, as shown in Figure 4.4 (Tanarangrak & Monsanit, 2006) as follows:

Thai: [ที่พัก OR โรงแรม OR ห้องพัก OR รีสอร์ท OR บ้านพัก OR เต้นท์] [ถนนคนเดิน] [เชียงใหม่] [ราคา] 500..1000

English: [Accommodation OR Hotel OR Lodge OR Resort OR Villa OR Tent] [Walking-Street] [Chiang Mai] [Price] 500..1000.

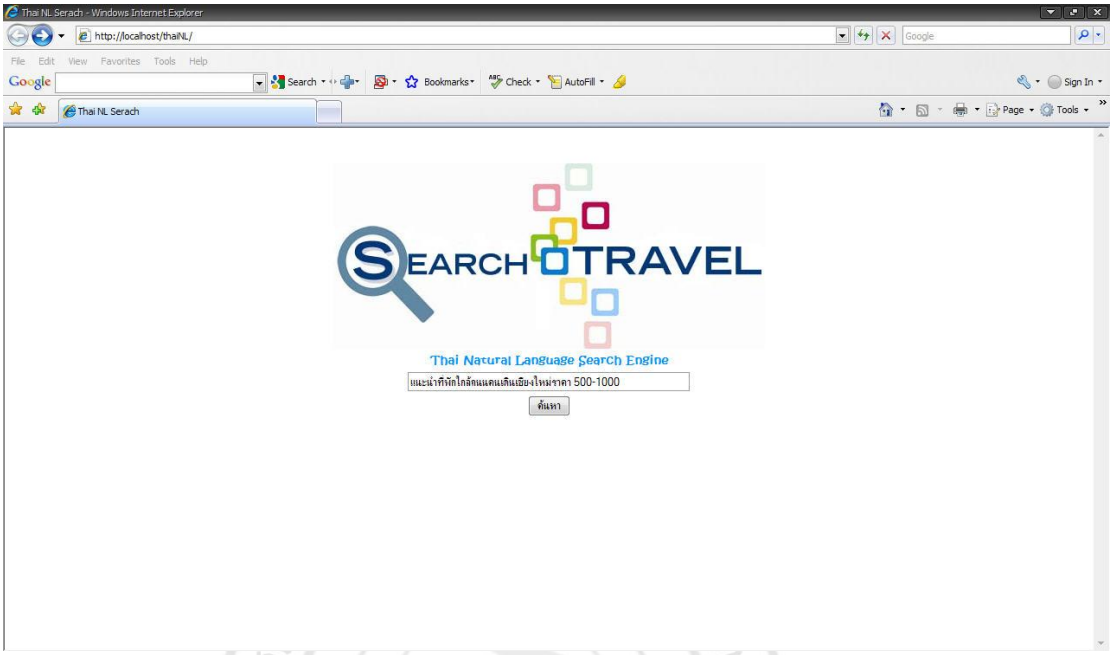

Figure 4.1 User Inputs Thai Natural Language

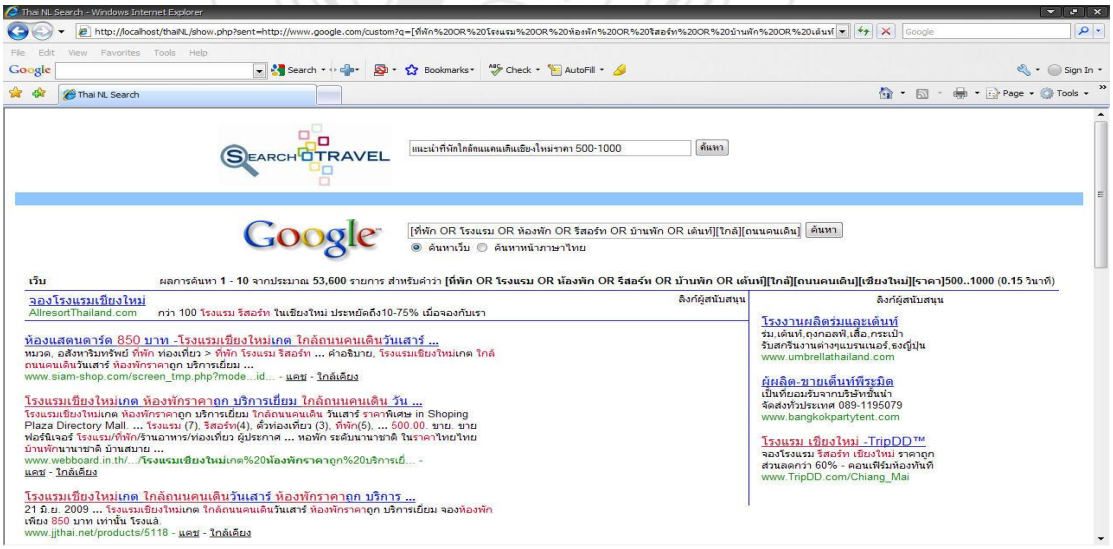

Figure 4.2 The Keyword Pattern Word Groups in Google Search

4.2.2.1 Searching without keyword pattern

User inputs Thai natural language phrase query in Google search. A query for search is “แนะนำที่พักใกล้ถนนคนเดินเชียงใหม่ราคา 500-1000 (Recommended accommodation near Chiang Mai walking-street with the price 500-1000)”, as shown in Figure 4.5.

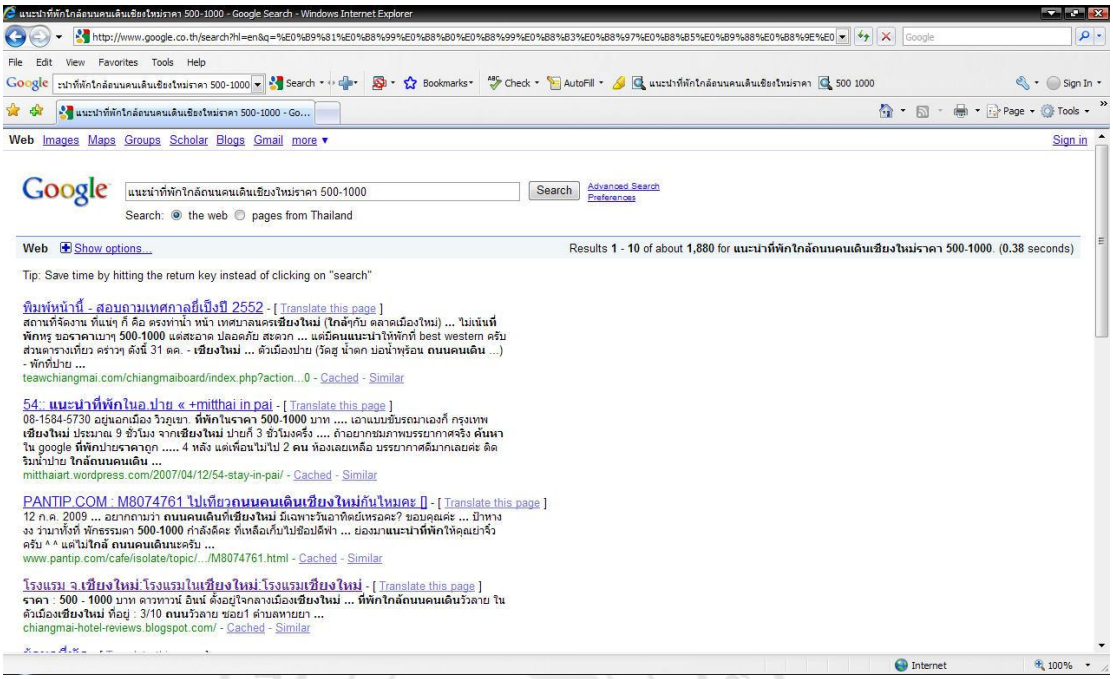

Figure 4.3 User Inputs Thai Natural Language in Google Search

The total 95 queries used in this thesis are shown in Appendix A.

4.3 Experimental Result

The proposed system was tested with 95 query phrases in order to compare search using keyword pattern and without keyword pattern. The evaluation from the fifty first results by check each result. The results should match with keywords, synonym, and content with query. That’s all should related with query. So, assume that’s results correct. Table 4.1 show the example precision of input query phrase by compare between search using keyword pattern and without keyword pattern.

Table 4.1 Example of Input Query

No	Query phrase
1	แนะนำเช่ารถมอเตอร์ไซต์ที่เชียงใหม่
2	แนะที่พักราคาถูกในเชียงใหม่
3	แนะนำโรงแรมหรือรีสอร์ทสำหรับจัดสัมมนาในเชียงใหม่
4	แนะนำสถานที่ท่องเที่ยวที่พะเยา
5	แนะนำที่กินที่เที่ยวที่เชียงใหม่
6	แนะนำที่พักในเชียงใหม่ราคา 500-1000
7	แนะนำโรงแรมในลำปาง
8	แนะนำโรงแรมแนวบูติกในเชียงใหม่
9	หาโฮมสเตย์ในเชียงราย
10	ขอรายชื่อร้านอาหารในเชียงใหม่

Table 4.1 shows the examples of input query. The query can be all about tourism in the northern part of Thailand.

Appendix A shows the score of precision by comparing between search without keyword pattern and search using keyword pattern. The precision is calculated as equation (4.1)

$$\text{Precision} = \frac{x}{y} \times 100 \qquad (4.1)$$

Where x is a number of the satisfied results, y is a number of total searched results. In this experiment, the results of the first 50 displayed websites are considered.

Table 4.2 The Example of Real Score before Calculate Precision

No	Query phrase	The precision of Search using Keyword Pattern	The precision of Search without Keyword Pattern
1	แนะนำเช่ารถมอเตอร์ไซต์ที่เชียงใหม่	18	8
2	แนะที่พักราคาถูกในเชียงใหม่	38	22
3	แนะนำโรงแรมหรือรีสอร์ทสำหรับจัดสัมมนาในเชียงใหม่	19	9
4	แนะนำสถานที่ท่องเที่ยวที่พะเยา	45	34
5	แนะนำที่กินที่เที่ยวที่เชียงใหม่	23	22
6	แนะนำที่พักในเชียงใหม่ราคา500-1000	35	28
7	แนะนำโรงแรมในลำปาง	34	28
8	แนะนำโรงแรมแนวบูติกในเชียงใหม่	35	34
9	หาโฮมสเตย์ในเชียงราย	27	24
10	ขอรายชื่อร้านอาหารในเชียงใหม่	28	26

Table 4.2 shows the example of the real score between search using keyword pattern and without keyword pattern by checking every result if it relates with query. If it matches or relates with query, it is counted as 1. The 50 results or the first five pages of the retrieved from the search using keyword pattern and without keyword pattern are checked.

Table 4.3 The Example of Precision of Both Process

No	Query phrase	The precision of Search using Keyword Pattern	The precision of Search without Keyword Pattern
1	แนะนำเช่ารถมอเตอร์ไซต์ที่เชียงใหม่	36	16
2	แนะที่พักราคาถูกในเชียงใหม่	76	44
3	แนะนำโรงแรมหรือรีสอร์ทสำหรับจัดสัมมนาในเชียงใหม่	38	18
4	แนะนำสถานที่ท่องเที่ยวที่พะเยา	90	68
5	แนะนำที่กินที่เที่ยวที่เชียงใหม่	46	44
6	แนะนำที่พักในเชียงใหม่ราคา500-1000	70	56
7	แนะนำโรงแรมในลำปาง	68	56
8	แนะนำโรงแรมแนวบูติกในเชียงใหม่	94	66
9	หาโฮมสเตย์ในเชียงราย	60	34
10	แนะนำโรงแรมราคา 500-1000 ในเชียงใหม่	56	52

After checking the real score in Table 4.2, the real score is taken to calculate for the precision in each query by equation 4.1. Table 4.3 shows the example of precision between search using keyword pattern and without keyword pattern. From the calculation of precision, it is found that the maximum and minimum of search using keyword pattern are 98% and 18%, respectively, and the maximum and minimum of search without keyword pattern are 90% and 0%, respectively.

The average of precision is as follows:

$$\text{Average of precision} = \frac{\sum \frac{x}{y} \times 100}{n} \qquad (4.2)$$

Where n is the total numbers of tested sentence. The average of precision is shown in Table 4.1

Table 4.1 shows the average of precision of 95 queries, comparing search using search pattern with search without search pattern. Appendix A shows the precision of 95 queries.

Table 4.4 The Average of Precision from 95 Queries.

	Search using Keyword Pattern	**Search without Keyword Pattern**
Average of precision	**62.33%**	**36.1%**

CHAPTER 5

DISSCUSION

This thesis proposed the system used to convert Thai natural language phrase into search engine format language in order to list up the desired websites that most match to the query. The system extracts user's intention from the phrase, and converts it into search engine pattern, then convert to Google syntax so that the appropriate alternative search is obtained. The keyword pattern includes nouns, synonyms, and Boolean operators. From the experiment, the search by using keyword pattern was compared with the search without keyword pattern. The result has shown that the proposed system gives the average precision of 62.33%, whereas search without keyword pattern gives the average precision of 36.1%.

5.1 General Matching Search

From the experiment, the proposed method gives the average precision of 62.33%. The search without keyword pattern has average precision of 36.1%. Obviously, search using keyword pattern increases the number of results related to the query. However, there are some results that match with keywords but not related with query. This is because keywords used for searching normally depend on the ranking of the keywords. Keywords ranking is dynamic, it changes all the time. For example, today frequency, which determines ranking, of โรงแรม (hotel) is 17 million, after one to two weeks later, the ranking maybe more or less. Therefore, ranking of keywords has effect on search results.

This proposed method has the limitation based on the keywords frequency. Therefore, the keywords need to be updated periodically.

5.2 Content Related Matching Search

From the experiment, the content of the search results were considered in detail to make sure that the content actually related to the meaning of the query of 62.33%. The search without keyword pattern has as average precision of 36.1%. From the experiment, 95 queries were performed, in each query its precision was calculated, then the maximum, the minimum and the average precision of all the queries were also determined. As the result, the maximum and minimum precision of search using keyword pattern is 98% and 18% respectively and maximum and minimum precision of search without keyword pattern is 90% and 0% respectively. The maximum between of search using keyword pattern and search without keyword pattern has a similar score. This is because the proposed method uses keywords that have specific meaning such as ถนนคนเดิน (walking-street), เชียงใหม่ (Chiang Mai), ที่พัก (accommodation), etc. However, it can be additional of words in database for help to expanding the keyword such as “แนะนำวัดที่สำคัญในตัวเมืองน่าน (Suggestion of an important temple in Nan)”. In this example, วัด (temple), ตัวเมือง (city), น่าน (Nan) are specific keywords and the word สำคัญ (important) is the expansion part of query. If using only specific keywords such as วัด (temple), the entire temple in Nan will show up whether the temple is important or not. But when using the word สำคัญ (important), the search will expand opportunity of results related to the query.

CHAPTER 6

CONCLUSION

6.1 Conclusion

Intelligent Search Engine is expected to assist people to efficiently find desired websites. Since we currently need to input keywords in search engine format for extracting desired websites, it is not friendly for normal people to understand the format and it is also difficult to get appropriate website lists. Actually, the interface with human should be friendly as the same level as talking with human, and the system has to understand natural language (NL). To develop this kind of search function, this research reports the model for Thai natural language search engine. Search engine usually has the problem of providing un-desired results from many matched websites. This research therefore proposed the method used to convert Thai natural language phrase into search engine in order to list up the desired websites as the most desired results. The methods extract user's intention from the phrase, and convert it into search engine pattern keywords, then convert to Google syntax so that the appropriate alternative search is obtained. The keyword pattern includes nouns, synonyms, and Boolean operators.

The experiment was conducted by comparing the search using keyword pattern and the search without keyword pattern. The proposed system gives the average precision of 62.33% with the maximum of precision 98% and minimum of precision 18%, whereas Google search has the average precision of 36.1% with maximum of precision 90% and minimum of precision 0%. From the experiment, the results show that there are two types of the search, i.e., general matching search and content related matching search.

6.2 Future Work and Suggestion

The two types of search are general matching search and content related matching search. Both types search have the limitations of the system because database for this system collects only the keywords that are nouns. It should collect other word classes such as verb, adverb, adjective, etc. to make more chance for searching. Moreover, the synonym frequency changes every day because each time user searches by using different keyword, so the frequency of synonym should be updated all the time. Therefore, in future work, the database should collect other word classes and be automatically updated. The synonym frequency should also make the system automatically update by creating a program called spider to look for the keywords that the user used for search. This is because those words will have effect on the search and make more chance for search. However, this system is able to use only with other languages which have no border between words like Thai language.

REFERENCES

REFERENCES

Achawanantakun, R. & Poovarawa, Y. (2003). **A performance improvement of search engine using conceptual knowledge.** Master's thesis. Computer Engineering. Kasetsart University, Bangkok.

Allen, B.P. (1997). WordWeb-Using the Lexicon for www.available source. Retrieved August 1,2003. from http://www.inference.com

Buckley, C., Saltion G., Allan, J., & Singhal, A. (1994). Automatic query expansion using SMART: TREC3 In **The third text retrieval conference.** (pp.69-81).

Charnyapornpong, S. (1983). **A Thai syllable separation algorithm.** Master's thesis. Asian Institute of Technology, Bangkok.

Charoenpornsawat, P. (1998). **Feature-Based Thai word segmentation.** Master's thesis. Computer Engineering. Chulalongkorn University, Bangkok.

Diab, M., Hacioglu, K., & Jurafsky, D. (2004). Automatic tagging of Arabic text: from raw text to base phrase chunks. In **Proceedings of HLT-NAACL.**

Hayder, K. Al Ameed, Shaikha, O. Al Ketbi, Amna A. Al Kaabi, Khadija, S. Al shebli. Naila, F.Al Shamsi, Noura, H. Al Nuaimi, and Shaikha, S. Al Muhairi. (2006). Arabic search engines improvement: A new approach using search key expansion derived from arabic synonyms structure. In **Proceeding of IEEE.**

Illinois Mathematics and Science Academy. (2006**). What are Synonyms.** Retrieved August 9,2006, from http://21cif.imsa.edu/tutorials/micro/mm/synonyms/index_html?b_start:int=3

Internet Tutorials. (2008). Boolean searching on the internet A primer in Boolean logic. Retrieved August 19,2009, form http://www.internettutorials.net/boolean.html

Ishikawa, K., Satoh, K., & Okumura, A. (1997). Query term expansion based on paragraphs of the relevant documents. In **The sixth text retrieval conference.** (pp.557-585).

Kawtrakul, A., Thumkanon, C., Poovorawan, Y., Varasrai, P., & Suktarachan, M., (1997). Automatic Thai unknow word recognition. In **Proceeding of the natural language processing pacific rim symposium 1997 (NLPRS' 97).**

Lerner, M. (2009). All Rights Reserved. **How Search Engine Work.** Retrieved 2009 from http://www.learnthenet.com/english/animate/search.html

Liddy, E.D. In Ecyclopedia of Library and Information Science. 2nd Ed. Marcel Decker, Inc.

Meknavin, S., Charoenpornsawat, P., & Kijsirikul, B., (1997). Feature-Based Thai word segmentation. In **Proceedings of the natural language processing pacific rim symposium 1997 (NLPRS' 97).**

Miller, G.A., Leacock, C., Randee, T., and Bunker, R., (1995). WordNet: A lexical database. **In Communication of the ACM**, (pp.39-41).

NECTEC (National Electronics and Computer Technology Centre). **LEXiTRON dictionary.** from http:// lexitron.necter.or.th/downloadLex_detail.html

Poovarawan, Y. & Amarom, V. (1986). **การแบ่งแยกพยางค์ไทยด้วยดิกชันนารี.** In Processing of Technical Electrical Engineering, 9.

Poovarawan, Y. et al. (2000). **Thai homonym dictionary.** Thailand: Science engineering & education.

Raruenrom, S. (1991). **Dictionary-Based Thai Word Separation.** In Senior Project Report. Department of Computer Engineering, Chulalongkorn University, Bangkok.

Salton, G. & Lesk M.E., (1971). **Computer evaluation of indexing and text processing.** In The SMART retrieval system: experiments in automatic document processing. (pp. 143-180).

Sawamipak, D. (1990). **Construction of Thai Syntax Analysing Software Under UNIX.** Master's thesis. Thammasart University Press, Bangkok.

Sornletlamvanich, V. (1993). **Word Segmentation for Thai in Machine Translation System.** In Machine Translation, pp. 50-56. NECTEC, Bangkok.

Tanarangrak, N & Monsanit, P. (2006). **Google.** KTP COMP & CONSULT, Bangkok.

Thairatananond, Y. (1981). **Towords the design of a Thai text syllable analyzer.** Master's thesis. Asian Institute of Technology, Bangkok.

Voorhees, E.M, (1994). **Query expansion using Lexcial-Sementic relations.** In Proceedings of the 17th Annual International ACM SIGIR, Conference on Research and Development in Information. (pp.61-69).

APPENDIXES

APPENDIX A

LIST OF INPUT QUERY PHRASE AND PRECISION

Table A.1 List of Input Query Phrase and Precision

No.	Thai Natural Language Phrase	Search Using Search Pattern	Search Without Search Pattern
1	แนะนำเช่ารถมอเตอร์ไซต์ที่เชียงใหม่	36	16
2	แนะนำที่พักใกล้สนามบินไนท์บาร์ซ่าเชียงใหม่	90	68
3	แนะนำที่พักราคาถูกในเชียงใหม่	76	44
4	แนะนำโรงแรมหรือรีสอร์ทสำหรับจัดสัมมนาในเชียงใหม่	38	18
5	แนะนำสถานที่ท่องเที่ยวที่พะเยา	90	68
6	แนะนำที่กินที่เที่ยวเชียงใหม่	46	44
7	แนะนำที่พักในตัวเมืองเชียงใหม่	68	60
8	แนะนำที่พักในเชียงใหม่ราคา 500-1000	70	56
9	แนะนำโรงแรมมีบริการรถรับส่งสนามบินเชียงใหม่	68	52
10	แนะนำโรงแรมในลำปาง	68	56
11	แนะนำโรงแรมแนวบูติกในเชียงใหม่	94	66
12	หาโฮมสเตย์ในเชียงราย	60	34
13	ขอรายชื่อร้านอาหารในเชียงใหม่	56	52
14	แนะนำโรงแรมใกล้ถนนคนเดินเชียงใหม่ราคา 500-1000	88	8
15	แนะนำเกสท์เฮ้าส์ใกล้ถนนคนเดินวันอาทิตย์เชียงใหม่	74	68
16	แนะนำโรงแรมหรือรีสอร์ทในเชียงใหม่ราคา 500-1000	44	40

Table A.1 (continued)

No.	Thai Natural Language Phrase	Search Using Search Pattern	Search Without Search Pattern
17	แนะนำรีสอร์ทในตัวเมืองเชียงใหม่	64	32
18	แนะนำรีสอร์ทที่มีสระว่ายน้ำที่ปาย	84	36
19	ราคาเช่ารถมอเตอร์ไซต์ในเชียงใหม่	40	38
20	จะไปเที่ยวน่าน แนะนำสถานที่ท่องเที่ยวด้วยครับ	88	22
21	แนะนำที่พักติดริมน้ำปาย	78	74
22	ที่พักแนวธรรมชาติที่เชียงราย	64	38
23	แนะนำที่พักที่ปายราคา 500-1000	82	30
24	แนะนำที่พักแถวไนท์บาร์ซ่าเชียงใหม่	74	46
25	แนะนำที่พักใน อ.เมือง ลำปาง	58	36
26	แนะนำที่พักในเชียงใหม่	98	90
27	แนะนำที่พักสำหรับหมู่คณะในเชียงใหม่	88	72
28	แนะนำที่พักสำหรับฮันนีมูนในเชียงใหม่	36	30
29	แนะนำที่พักรายเดือนในเชียงใหม่	94	40
30	แนะนำโฮมสเตย์ในเชียงราย	72	52
31	แนะนำวัดสำคัญในตัวเมืองน่าน	82	32
32	แนะนำที่พักสำหรับหมู่คณะและจัดสัมมนาในเชียงใหม่	60	40
33	แนะนำจัดทัวร์ สัมมนา หมู่คณะเชียงใหม่	76	52
34	แนะนำรีสอร์ทสำหรับจัดสัมมนาในเชียงราย	42	20
35	แนะนำที่เที่ยวที่น่าน	50	14
36	แนะนำการเดินทางไปน่าน	52	20
37	แนะนำที่พักในตัวเมืองน่าน	64	24
38	แนะนำเทศกาลประเพณีสำคัญของเชียงใหม่	96	86
39	แนะนำร้านอาหารที่น่าน	78	14
40	แนะนำร้านอาหารที่เที่ยวที่น่าน	46	6

Table A.1 (continued)

No.	Thai Natural Language Phrase	Search Using Search Pattern	Search Without Search Pattern
41	แนะนำร้านอาหารที่เที่ยวที่น่าน	54	38
42	แนะนำของฝากในเชียงใหม่	72	36
43	แนะนำสถานที่ท่องเที่ยวในลำพูน	92	48
44	แนะนำโรงแรมในจังหวัดลำพูน	74	66
45	แนะนำที่พักใกล้สนามบินเชียงใหม่	56	38
46	แนะนำกิจกรรมผจญภัยในเชียงใหม่	60	52
47	แนะนำบริษัทนำเที่ยวเชิงผจญภัยอนุรักษ์ที่เชียงใหม่	64	34
48	แนะนำร้านกาแฟในเชียงราย	50	36
49	แนะนำร้านเค้กในเชียงใหม่	96	88
50	แนะนำโฮมสเตย์ที่น่าน	46	8
51	แนะนำโฮมสเตย์ที่แม่ฮ่องสอน	68	48
52	แนะนำท่องเที่ยวผจญภัยที่แม่ฮ่องสอน	72	54
53	แนะนำโรงแรมในเชียงใหม่ราคา 300-800	50	12
54	แนะนำเกสท์เฮ้าส์ในเชียงใหม่ราคา 300-500	72	26
55	แนะนำสถานที่ท่องเที่ยวในพะเยา	92	60
56	แนะนำแหล่งท่องเที่ยวธรรมชาติในเชียงใหม่	54	38
57	แนะนำสถานที่ท่องเที่ยวแบบน้ำตกในเชียงใหม่	72	26
58	แนะนำการท่องเที่ยวผจญภัยเดินป่าในเชียงใหม่	76	58
59	แนะนำสถานที่ท่องเที่ยวธรรมชาติที่น่าน	54	12
60	ลำพูนมีอะไรน่าท่องเที่ยวบ้าง	76	16
61	แนะนำที่พักแบบมีสปาในเชียงใหม่	86	28
62	แนะนำอุทยานแห่งชาติและน้ำตกในแม่ฮ่องสอน	78	78
63	แนะนำรีสอร์ทในตัวเมืองเชียงราย	60	24
64	แนะนำร้านอาหารในตัวเมืองเชียงราย	44	26
65	แนะนำที่พักร้านอาหารในเชียงใหม่	72	50
66	น่านมีสถานที่ท่องเที่ยวที่ไหนบ้าง	82	38

Table A.1 (continued)

No.	Thai Natural Language Phrase	Search Using Search Pattern	Search Without Search Pattern
67	แนะนำของฝากขึ้นชื่อประจำจังหวัดพะเยา	50	40
68	แนะนำรีสอร์ทในพะเยา	68	24
69	แนะนำรีสอร์ทในจังหวัดน่าน	72	0
70	แนะนำโรงแรมหรือเกสท์เฮ้าส์ในเชียงใหม่ราคา 400-800	54	24
71	แนะนำที่พักใกล้ถนนคนเดินเชียงใหม่ราคา 500-1000	64	10
72	แนะนำโรงแรมในตัวเมืองเชียงใหม่ราคา 300-800	50	10
73	แนะนำโรงแรมมีสปาสระว่ายน้ำในเชียงใหม่	66	44
74	แนะนำเกสทเฮ้าส์ใกล้ถนนคนเดินในเชียงใหม่ราคา 500-800	30	4
75	แนะนำที่พักติดคูเมืองเชียงใหม่	18	16
76	แนะนำรีสอร์ทในตัวเมืองเชียงใหม่ที่ให้บริการรถรับส่งสนามบินมีสระว่ายน้ำรวมถึงบริการนวด	42	10
77	หรือสปา แนะนำโรงแรมที่มีสระว่ายน้ำห้องซาวน่าและฟิตเนส	64	44
78	ในเชียงใหม่ หาโรงแรมที่มีสระว่ายน้ำใกล้ไนท์บาร์ซ่าเชียงใหม่	68	58
79	หาโรงแรมแนวบูติคอยู่ในใจกลางเมืองเชียงใหม่ราคา	46	18
80	500-1500 แนะนำที่พักติดแม่น้ำปิงในเชียงใหม่	62	22
81	แนะนำเกสท์เฮ้าส์ติดแม่น้ำปิงเชียงใหม่	30	14
82	แนะนำที่พักสำหรับหมู่คณะและจัดสัมมนาใน	52	34
83	เชียงใหม่ หารีสอร์ทที่มีนวดสปาสระว่ายน้ำในตัวเมืองเชียงใหม่	44	26
84	แนะนำโฮมสเตย์ในเชียงใหม่	56	36

Table A.1 (continued)

No.	Thai Natural Language Phrase	Search Using Search Pattern	Search Without Search Pattern
85	แนะนำเกสท์เฮ้าส์ที่มีบริการให้เช่ารถมอเตอร์ไซต์ในเชียงใหม่	28	24
86	แนะนำโรงแรมที่มีบริการทัวร์และรถรับส่งสนามบินเชียงใหม่	32	16
87	แนะนำรีสอร์ทที่มีบริการห้องพักสำหรับคู่ฮันนีมูนเชียงใหม่	46	18
88	แนะนำที่พักแบบมีสปาในเชียงใหม่	66	28
89	แนะนำโรงแรมหรือเกสท์เฮ้าส์ในเชียงใหม่ราคา 400-800	50	6
90	แนะนำเกสท์เฮ้าส์ที่มีบริการนวดในเชียงใหม่ราค 500-1500	34	4
91	แนะนำเกสท์เฮ้าส์หรือโรงแรมติดคูเมืองในเชียงใหม่	32	22
92	แนะนำรีสอร์ทสำหรับหมู่คณะในเชียงใหม่	48	20
93	แนะนำโฮมสเตย์สำหรับหมู่คณะในเชียงใหม่	34	26
94	แนะนำที่พักใกล้ไนท์บาร์ซ่ามีบริการนวดและสระว่ายน้ำในเชียงใหม่	60	52
95	แนะนำเกสท์เฮ้าส์ใจกลางเมืองเชียงใหม่	52	48

APPENDIX B

TABLE IN KEYWORD DATABASE

Table B.1 Table of Noun Keyword

id_noun	Word	id_noun	Word
1001	เรื่อง	1021	วิธี
1002	ที่พัก	1022	สายการบิน
1003	ตัวเมือง	1023	เชียงใหม่
1004	การเดินทาง	1024	เชียงราย
1005	รถเช่า	1025	ลำพูน
1006	สถานีรถไฟ	1026	ลำปาง
1007	สนามบิน	1027	แพร่
1008	ล่องแก่ง	1028	น่าน
1009	บริษัททัวร์	1029	แม่ฮ่องสอน
1010	เส้นทาง	1030	พะเยา
1011	ฮันนีมูน	1031	แก่งเจ็ดแคว
1012	อุทยาน	1032	ขุนขาน
1013	สถานที่ท่องเที่ยว	1033	ขุนแจ
1014	ราคา	1034	ขุนน่าน
1015	คำแนะนำ	1035	คลองตรอน
1016	จังหวัด	1036	คลองลาน
1017	อำเภอ	1037	คลองวังเจ้า
1018	ตำบล	1038	แจ้ซ้อน
1019	แถว	1039	เชียงดาว
1020	ร้านอาหาร	1040	ดอยขุนตาล

Table B.1 (continued)

id_noun	Word	id_noun	Word
1041	ดอยจง	1068	แม่ตะไคร้
1042	ดอยผากลอง	1069	แม่โถ
1043	ดอยภูคา	1070	แม่ปิง
1044	ดอยภูนาง	1071	แม่ฝาง
1045	ดอยเวียงผา	1072	แม่เมย
1046	ดอยสุเทพ	1073	แม่ยม
1047	ดอยปุย	1074	แม่วงก์
1048	ดอยหลวง	1075	แม่วะ
1049	ดอยอินทนนท์	1076	แม่วัง
1050	ตากสินมหาราช	1077	รามคำแหลง
1051	ตาดหมอก	1078	ลานสาง
1052	ถ้ำปลา-น้ำตกผาเสื่อ	1079	ลำน้ำน่าน
1053	ถ้ำผาไท	1080	เวียงโกศัย
1054	ถ้ำสะเกิน	1081	ศรีน่าน
1055	ทุ่งแสลงหลวง	1082	ศรีสัชนาลัย
1056	นันทบุรี	1083	สาละวิน
1057	น้ำตกชาติตระการ	1084	ห้วยน้ำดัง
1058	น้ำตกพาเจริญ	1085	ออบขาน
1059	น้ำตกแม่สุรินทร์	1086	ออบหลวง
1060	น้ำหนาว	1087	ปาย
1061	ป่าแม่ปืม	1088	ปางอุ๋ง
1062	ภูซาง	1089	ถนนคนเดิน
1063	ภูสอยดาว	1090	บ่อสร้า
1064	ภูหินร่องกล้า	1091	หนึ่งร้อย
1065	แม่กาษา	1092	สองร้อย
1066	แม่เงา	1093	สามร้อย
1067	แม่จริม	1094	สี่ร้อย

Table B.1 (continued)

id_noun	Word	id_noun	Word
1095	ห้าร้อย	1123	ของฝาก
1096	หกร้อย	1124	สปา
1097	เจ็ดร้อย	1125	บริษัท
1098	แปดร้อย	1126	บูติค
1099	เก้าร้อย	1128	ตั๋วเครื่องบิน
1100	หนึ่งพัน	1129	รถยนต์
1101	หนึ่งพันห้าร้อย	1130	รถประจำทาง
1102	สองพัน	1131	รถมอเตอร์ไซต์
1103	สองพันห้าร้อย	1132	รถไฟ
1104	สามพัน	1133	เครื่องบิน
1105	สามพันห้าร้อย	1134	เรือ
1106	สี่พัน	1135	รถทัวร์
1107	สี่พันห้าร้อย	1136	รถจักรยาน
1108	ห้าพัน	1137	การบินไทย
1109	ห้าพันห้าร้อย	1138	นกแอร์
1110	หกพัน	1139	แอร์เอเชีย
1111	หกพันห้าร้อย	1140	วันทูโก
1112	เจ็ดพัน	1141	วัด
1113	เจ็ดพันห้าร้อย	1142	ไหว้พระ
1114	แปดพัน	1143	กลางคืน
1115	แปดพันห้าร้อย	1144	รายชื่อ
1116	เก้าพัน	1145	หมู่คณะ
1117	เก้าพันห้าร้อย	1146	ธรรมชาติ
1118	หนึ่งหมื่น	1147	ตารางเวลา
1119	ราคาถูก	1148	รถตู้
1120	ราคาไม่เกิน	1149	แม่น้ำปิง
1122	สัมมนา	1150	คน

Table B.1 (continued)

id_noun	Word	id_noun	Word
1151	ริมน้ำ	1180	ไม่เกิน
1152	เช่า	1181	ซาวน่า
1153	ค่ารถ	1182	แพคเกจ
1154	ท่องเที่ยว	1183	มหาวิทยาลัยเชียงใหม่
1157	รายเดือน	1184	ครอบครัว
1158	เทศกาล	1185	มอเตอร์ไซต์
1159	ประเพณี	1186	วันเสาร์
1160	ร้านค้า	1187	วันอาทิตย์
1161	กิจกรรม		
1162	ผจญภัย		
1163	นำเที่ยว		
1164	อนุรักษ์		
1165	ร้านกาแฟ		
1166	ร้านเค้ก		
1167	น้ำตก		
1168	ภูเขา		
1169	เดินป่า		
1170	ข้อมูล		
1171	ห้องประชุม		
1172	สระว่ายน้ำ		
1173	นวด		
1174	อินเตอร์เน็ต		
1175	รถ		
1176	รับ		
1177	ส่ง		
1178	คูเมือง		
1179	ฟิตเนส		

Table B.2 Table of Synonym Keyword

id_noun	Word	id_noun	Word
1002	โรงแรม	1019	ระแวก
1002	รีสอร์ท	1019	แถบ
1002	เกสท์เฮ้าส์	1020	ภัตตาคาร
1002	โฮมสเตย์	1020	ที่กิน
1002	บังกะโล	1020	ห้องอาหาร
1002	บ้านพัก	1021	ขั้นตอน
1002	เต้นท์	1021	กระบวนการ
1002	ห้องพัก	1021	กรรมวิธี
1003	ในเมือง	1021	ขบวนการ
1003	ใจกลางเมือ	1021	แนวทางปฏิบัติ
1003	ชุมชน	1119	ราคาไม่แพง
1007	ท่าอากาศยาน	1119	ไม่แพง
1010	ทาง	1120	ราคาต่ำกว่า
1010	ทางสัญจร	1122	ประชุม
1010	ตรอก	1122	อบรม
1010	ซอกซอย	1123	ของที่ระลึก
1011	ดื่มน้ำผึ้งพระจันทร์	1125	กิจการ
1012	อุทยานแห่งชาติ	1125	องค์กร
1013	ที่เที่ยว	1125	องค์การ
1013	แหล่งท่องเที่ยว	1125	ห้างร้าน
1015	ข้อเสนอแนะ	1009	ทัวร์
1015	ข้อแนะนำ	1003	อ.เมือง
1015	คำเสนอแนะ	1003	อำเภอเมือง
1015	การเสนอแนะ	1149	น้ำปิง
1015	การแนะแนว	1137	จักรยาน
1015	การชี้แนะ	1162	เสี่ยงภัย
1019	บริเวณ	1163	เสี่ยงอันตราย

Table B.2 (continued)

id_noun	Word
1164	รักษา
1164	สงวน
1183	มช.
1019	ใกล้

CURRICULUM VITAE

NAME	Miss Chotika Poungmali
DATE OF BIRTH	10 May 1981
ADDRESS	88 Moo 2 T.Yangnung A.Saraphee, Chiang Mai 50140

EDUCATIONAL BACKGROUND

2003	Bachelor of Business Administration
	Major in Business Information Technology
	Maejo University