A SEMI-AUTOMATIC CONSTRUCTION OF THAI WORDNET LEXICAL DATABASE FROM MACHINE READABLE DICTIONARIES

PATANAKUL SATHAPORNRUNGKIJ

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR
THE DEGREE OF MASTER OF SCIENCE
(COMPUTER SCIENCE)
FACULTY OF GRADUATE STUDIES
MAHIDOL UNIVERSITY
2004

ISBN 974-04-5256-6
COPYRIGHT OF MAHIDOL UNIVERSITY

Copyright by Mahidol University

Thesis
Entitled

A SEMI-AUTOMATIC CONSTRUCTION OF THAI WORDNET LEXICAL DATABASE FROM MACHINE READABLE DICTIONARIES

Patanakul Sathapornrungkij
.....
Miss Patanakul Sathapornrungkij
Candidate

Charnyote Pluempitiwiriyawej
.....
Lect. Charnyote Pluempitiwiriyawej, Ph.D.
Major-Advisor

Thanwadee Sunetnanta
.....
Asst. Prof. Thanwadee Sunetnanta, Ph.D.
Co-Advisor

Rass Hoon
.....
Assoc.Prof. Rassmidara Hoonsawat, Ph.D.
Dean
Faculty of Graduate Studies

Supachai Tangwongsan
.....
Assoc.Prof. Supachai Tangwongsan, Ph.D.
Chair
Master of Science Programme
in Computer Science
Faculty of Science

Copyright by Mahidol University

Thesis
Entitled

A SEMI-AUTOMATIC CONSTRUCTION OF THAI WORDNET LEXICAL DATABASE FROM MACHINE READABLE DICTIONARIES

was submitted to the Faculty of Graduate Studies, Mahidol University
For the degree of Master of Science (Computer Science)
on
May 21, 2004

Patanakul Sathapornrungkij
.....
Miss Patanakul Sathapornrungkij
Candidate

.....
Lect. Charnyote Pluempitiwiriyawej, Ph.D.
Chair

Thanwadee Sunetnanta
.....
Asst. Prof. Thanwadee Sunetnanta, Ph.D.
Member

Nuttanart Muansuwan
.....
Lect. Nuttanart Muansuwan, Ph.D.
Member

.....
Assoc.Prof. Rassmidara Hoonsawat, Ph.D.
Dean
Faculty of Graduate Studies
Mahidol University

.....
Prof. Amaret Bhumiratana, Ph.D.
Dean
Faculty of Science
Mahidol University

Copyright by Mahidol University

ACKNOWLEDGEMENTS

I am most grateful to my thesis advisor, Dr. Charnyote Pluempitiwiriyawej, for his guidance, invaluable suggestions and helpful discussions given to me throughout the course of this thesis.

I am also grateful to Dr. Thanwadee Sunetnanta, Assoc.Prof. Damras Wongsawang and Dr. Nuttanart Muansuwan for their invaluable comments and for serving as my thesis committee.

My special thanks are extended to Mr. Javier Farreres for his academic advices, and Ms. Kaewchai Chancharoen, Mr. Lappakorn and Mr. Dumrong for their help. I also would like to thank my good friend, Ms. Jitratda, Ms. Thida, Ms. Siri-orn, Ms. Lalida, Mr. Smith, Mr. Ananta and many friends at Merlin's Solutions International for their love and help especially Mr. Sutthisak Inthawadee.

Finally, I am grateful to my family and Mr. Nipon Mankarn for their constant love, encouragement and support. I would like to dedicate this thesis to my father, my mother and all the teachers who have taught me.

Patanakul Sathapornrungkij

Copyright by Mahidol University

A SEMI-AUTOMATIC CONSTRUCTION OF THAI WORDNET LEXICAL DATABASE FROM MACHINE READABLE DICTIONARIES

PATANAKUL SATHAPORNRUNGKIJ 4237668 SCCS/M

M.Sc. (COMPUTER SCIENCE)

THESIS ADVISORS: CHARNYOTE PLUEMPITIWIRIYAWEJ, Ph.D., THANWADEE SUNETNANTA, Ph.D.

ABSTRACT

A significant increase in the use of lexical databases has led WordNet to become one of the most widely used lexical information source for Natural Language Processing applications. WordNet is a lexical database that, apart from lexical information, also provides semantic information and is available in many languages. Despite its popularity, the database still lacks support for Thai language.

This research presents a semi-automatic method for constructing Thai WordNet lexical databases from machine readable dictionaries (MRDs). The method starts with extracting lexical information from available MRDs: the Princeton WordNet, also known as a semantic network, and the LEXiTRON, a Thai⇔English dictionary. Both MRDs contain hundreds of thousands of English and Thai words. Then sets of candidate translation links are generated with respect to the extracted lexical information. These translation links are the representations of the semantic relationships among the words. A subset of these candidate links is later on verified and used to generate a model, which is finally used to automatically construct the Thai WordNet.

Our Thai WordNet has a coverage of 44,844 links, 19,582 synsets, and 13,730 words. The logistic regression is used to choose translation links among candidate links. It shows an accuracy of 76%. This semi-automatic method takes less time than the manual. The result shows that our method is effective.

KEY WORDS: THAI WORDNET/ LEXICAL DATABASE/ MACHINE READABLE DICTIONARIES

112 pp. ISBN 974-04-5256-6

Copyright by Mahidol University

การสร้างฐานข้อมูลคำศัพท์ไทยเวิร์ดเน็ตจากพจนานุกรมอิเล็กทรอนิกส์แบบกึ่งอัตโนมัติ(A SEMI-AUTOMATIC CONSTRUCTION OF THAI WORDNET LEXICAL DATABASE FROM MACHINE READABLE DICTIONARIES)

พัฒณกุล สถาพรรุ่งกิจ 4237668 SCCS/M

วท.ม. (วิทยาการคอมพิวเตอร์)

คณะกรรมการควบคุมวิทยานิพนธ์ : ชาญยศ ปลื้มปีติวิริยะเวช, Ph.D., ธันวดี สุเนตนันท์, Ph.D.

บทคัดย่อ

จากความสำคัญของฐานข้อมูลคำศัพท์ที่เพิ่มขึ้นนั้นทำให้เวิร์ดเน็ต (wordnet) กลายเป็นฐานข้อมูลคำศัพท์ที่ใช้กันอย่างแพร่หลายสำหรับงานด้านการประมวลผลภาษาธรรมชาติเนื่องจาก WordNet มีข้อมูลเกี่ยวกับไวยากรณ์ซึ่งมีความสำคัญในงานด้านนี้มาก จากความแพร่หลายของเวิร์ดเน็ตทำให้เวิร์ดเน็ตได้ถูกนำมาสร้างในภาษาต่างๆ เว้นแต่ภาษาไทย ดังนั้นงานวิจัยนี้ได้นำเสนอการสร้างฐานข้อมูลคำศัพท์ไทยเวิร์ดเน็ตจากพจนานุกรมอิเล็กทรอนิกส์แบบกึ่งอัตโนมัติ ซึ่งมุ่งเน้นการสร้างฐานข้อมูลคำศัพท์เวิร์ดเน็ตภาษาไทยขึ้น

งานวิจัยนี้ได้นำเสนอแนวทางการสร้างเวิร์ดเน็ตซึ่งจะดึงข้อมูลเกี่ยวกับคำศัพท์ต่างๆ จากพจนานุกรมอิเล็กทรอนิกส์ที่มีอยู่ โดยจะเริ่มกระบวนการด้วยการเค้นข้อมูลจากพจนานุกรมอิเล็กทรอนิกส์คือ Princton WordNet ซึ่งเป็นฐานข้อมูลคำศัพท์เวิร์ดเน็ตภาษาอังกฤษที่มีโครงข่ายความหมายของคำ และ LEXiTRON ซึ่งเป็นพจนานุกรมอิเล็กทรอนิกส์ภาษาไทย ⇔ ภาษาอังกฤษ ฐานข้อมูลคำศัพท์ทั้งสองนี้มีคำศัพท์จำนวนมาก กระบวนการถัดมาจะเป็นการสร้างชุดความสัมพันธ์ระหว่างคำภาษาไทยและความหมายจากเวิร์ดเน็ตภาษาอังกฤษโดยใช้ข้อมูลที่ได้จากการเค้นข้อมูล ซึ่งความสัมพันธ์ระหว่างคำภาษาไทยและความหมายจากเวิร์ดเน็ตภาษาอังกฤษนี้จะถูกนำเสนอในรูปของความสัมพันธ์ระหว่างคำต่างๆ และความสัมพันธ์เหล่านี้จะถูกตรวจสอบและนำมาสร้างแบบจำลองทางสถิติที่มาช่วยในการตรวจสอบ candidate link ให้สร้างไทยเวิร์ดเน็ตได้รวดเร็วยิ่งขึ้น

จากการนำ Logistic Regression ช่วยในการตรวจสอบ ชุดความสัมพันธ์ที่ได้สร้างขึ้นนั้นจะได้ความแม่นยำของ translation link เท่ากับ 76% ด้วยความครอบคลุมที่ 44,844 ความสัมพันธ์, 19,582 ความหมาย, and 13,730 คำ ซึ่งจะใช้เวลาในการสร้างพจนานุกรมน้อยกว่าการสร้างพจนานุกรม จากผลที่ได้รับนี้ถือได้ว่ามีประสิทธิผลดี

112 หน้า ISBN 974-04-5256-6

Copyright by Mahidol University

CONTENTS

	Page
ACKNOWLEDGEMENTS	iii
ABSTRACT	iv
CONTENTS	vi
LIST OF TABLES	ix
LIST OF FIGURES	x
CHAPTER 1 INTRODUCTION	1
1.1 Challenges	1
1.2 Goal of the thesis	2
1.3 Thesis Organization	3
CHAPTER 2 LITERATURE SURVERY	4
2.1 Lexical Database	4
2.2 Lexical database Construction	5
2.3 WordNet	8
2.3.1 EuroWordNet	9
2.3.2 Spanish WordNets	10
2.3.3 MultiWordNet (Italian Wordnet)	11
2.3.4 Chinese-English WordNet	11
2.3.5 Comparative Study of WordNet Construction Approach	11
CHAPTER 3 THAI WORDNET CONSTRUCTION ARCHITECTURE	14
3.1 MRD Extractor	16
3.2 Link Analyzer	17
3.3 WordNet Constructor	23
CHAPTER 4 SYSTEM DESIGN	26

Copyright by Mahidol University

CONTENTS (CONT.)

Page

4.1 Structure Chart 26
4.2 Top-Level Data Flow Diagrams and Process Description 27
4.3 Low-Level Data Flow Diagrams and Module Specifications 33
4.4 File Structure 41
4.4.1 Source Table 41
4.4.2 Intermediate Table 45
4.4.3 Target Table 47
4.5 Input-Output Design 48

CHAPTER 5 SYSTEM IMPLEMENTATION 54
5.1 MRD Extractor 55
5.1.1 Conversion 56
Monolingual Dictionary Conversion 56
Bilingual Dictionary Conversion 57
5.1.2 Cleansing 59
Monolingual Dictionary Cleansing 59
Bilingual Dictionary Cleansing 60
5.1.3 Alignment 64
5.1.4 Selection 65
5.1.5 Result of MRD Extractor 66
5.2 Link Analyzer 66
5.2.1 Candidate Link Construction 66
Class criteria group 67
Structural criteria group 69
5.2.2 Sampling 72
5.2.3 Verification 73
5.2.4 Model Construction 75
5.2.5 Model Evaluation 77
5.2.6 Model Employment 78

Copyright by Mahidol University

CONTENTS (CONT.)

Page

5.3 WordNet Constructor 80
5.3.1 Synset Attachment 80
5.3.2 Relation Inference 80
5.3.3 Glossary Attachment 80
5.4 Summary 80
5.4.1 Difficulties 81
5.4.2 Performance 81
Accuracy 81
Coverage 82
5.4.3 Limitations 82

CHAPTER 6 CONCLUSION AND FUTURE WORK 83
6.1 Conclusion 83
6.2 Suggestion for future work 84

REFERENCES 86

APPENDIX 90

BIOGRAPHY 112

Copyright by Mahidol University

LIST OF TABLES

Page

Table 2.1 Compares approaches and accuracy of construction of WordNet variations. 12

Table 3.1 Matrix of existence of translation link in each criteria 21

Table 3.2 The result of evaluation Matrix 22

Table 5.1 The example of filling missing data in monolingual dictionary 59

Table 5.2 The example of filling missing data in bilingual dictionary 60

Table 5.3 The example of removing wrong spelling 61

Table 5.4 The example of manipulating words with parenthesis 62

Table 5.5 The example of removing duplicate record 64

Table 5.6 Number of data from each process in MRD Extractor 66

Table 5.7 The examples of candidate links created from monosemic 1:1. 70

Table 5.8 Number of link created in each criteria 71

Table 5.9 The matrix of existence of translation link in all criteria. 72

Table 5.10 Correctness of each criteria (sample size=400) 74

Table 5.11 The matrix of boolean variables of each criteria 75

Table 5.12 Coefficient of each factor in the first model 76

Table 5.13 Coefficient of each factor in the second model 77

Table 5.14 The example of E(P(OK)) of each link 78

Table 5.15 Coverage comparison 82

Copyright by Mahidol University

LIST OF FIGURES

Page

Figure 2.1 The global architecture of the EuroWordNet database. 9

Figure 3.1 WordNet Builder (WNB) system 14

Figure 3.2 MRD Extractor 16

Figure 3.3 Link Analyzer 18

Figure 3.4 Multiple Criteria 18

Figure 3.5 Class criteria group 20

Figure 3.6 Structure criteria group 21

Figure 3.7 WordNet Constructor 23

Figure 3.8 Inference Internal relation 25

Figure 3.9 The example of inference the internal relations 25

Figure 4.1 Structure Chart of WNB 26

Figure 4.2 Context Diagram: WordNet Builder 27

Figure 4.3 Data Flow Diagram Level 1: WordNet Builder. 28

Figure 4.4 Data Flow Diagram Level 2: MRD Extraction 29

Figure 4.5 Data Flow Diagram Level 2: Link Analysis 31

Figure 4.6 Data Flow Diagram Level 2: WordNet Construction 32

Figure 4.7 Data Flow Diagram Level 3: MRDs Extractor 34

Figure 4.8 Data Flow Diagram Level 3: Candidate Link Construction 36

Figure 4.9 The relational model for wordnet2sql. 42

Copyright by Mahidol University

LIST OF FIGURES (CONT.)

Page

Figure 4.10 The relational model for Thai WordNet. 47

Figure 4.11 Candidate Link Construction (13 criteria) 49

Figure 4.12 Verify candidate link 50

Figure 4.13 Count Verified link 50

Figure 4.14 Create matrix 51

Figure 4.15 Input Coefficients 51

Figure 4.16 Model Employment 52

Figure 4.17 Thai WordNet (definitions and relations) 52

Figure 4.18 Thai WordNet (hypernym relation) 53

Figure 4.19 Thai WordNet (part meronym relation) 53

Figure 5.1 Overview of WNB 54

Figure 5.2 The example data of wordnet2sql 56

Figure 5.3 The example of converted wordnet2sql (table name: synset) 57

Figure 5.4 The example data of LEXiTRON in XML file format 58

Figure 5.5 The example of converted LEXiTRON (table name: E-T) 58

Figure 5.6 Alignment module 65

Figure 5.7 Monosemic Criteria 68

Figure 5.8 Polysemic Criteria 68

Figure 5.9 Structural Criteria 69

Figure 5.10 Correctness of each criteria (sample) 74

Figure 5.11 Coverage V.S. E(P(OK)) (Before remove C08) 79

Copyright by Mahidol University

LIST OF FIGURES (CONT.)

Page

Figure 5.12 Coverage V.S. E(P(OK)) (After remove C08) 79

Copyright by Mahidol University

CHAPTER 1

INTRODUCTION

A lexical database is a fundamental repository of information about language components (e.g. words meaning, grammars and pragmatic knowledge). It is one of the most significant information resources for a variety of applications such as speech transcription, information retrieval, information extraction, text summarization, message classification, question answering, discourse understanding, and machine translation. Different applications use information that is stored in the lexical database on different levels, ranging from a rough approximation triggered by keywords to a deep understanding that applies all information. For example, in speech recognition, the information about sound features (e.g. phonetic information) is used whereas meanings of words and grammars as well as pragmatic knowledge are applied to machine translation.

Due to the significant increase in the use of lexical databases, WordNet becomes one of the lexical databases that is widely used as a lexical information source in NLP applications. Unlike other lexical databases, which are normally specific to a particular application, WordNet is designed to support general use. WordNet is available in many languages including English [20], Dutch[34], Italian[25] and Spanish[1]. It is not yet available in Thai, however.

1.1 Challenges

Constructing the WordNet lexical database for Thai language, called *Thai WordNet* (TWN), is a challenging problem because of the specification and complexity of the particular language. A lexical database can be constructed from natural language corpora or from machine-readable dictionaries (MRDs). The corpora are collections of words, texts and sample sentences which are collected from particular materials such as newspapers and scripts of radio broadcast. The MRDs are

Copyright by Mahidol University

basically electronic representations of standard printed dictionaries. They can be monolingual (e.g., COLLIN[7]), bilingual (e.g., LEXiTRON[24]) or multilingual (e.g., Babylon[2]). One may consider a lexical database to be a type of MRD since lexical entries in the database are typically similar to that in MRD but the structure of the lexical database is more precise and predictable. In addition, a lexical database that is constructed from the corpora is typically more application-specific than those constructed from MRDs.

Two approaches have been presented to construct a lexical database. One is a manual construction, and the other is an (semi-) automatic construction. The manual construction requires a large number of lexicographers to hand-build the lexical database. This approach is applied to the construction of the Princeton WordNet [20], which is the WordNet lexical database of English. The (semi-) automatic construction, on the other hand, uses computer to extract and manipulate the information from existing resources with probably very little human intervention. This approach is applied to the construction of many WordNet databases, one of which is the Spanish WordNet [1]. Manual construction is a cumbersome, labor-intensive, expensive and time-consuming process, whereas the (semi-) automatic construction is more manageable, less labor-intensive, less expensive and faster.

1.2 Goal of the thesis

To construct the Thai WordNet lexical database, we adapt the semi-automatic method that is presented in construction of the Spanish WordNet. The method is used in our construction framework, called WordNet Builder (WNB). The WNB extracts lexical information from available MRDs, which include the Princeton WordNet and the LEXiTRON. Both MRDs contain totally hundred thousands of English and Thai words which include nouns, verbs, adjectives and adverbs. We, in this thesis, focus exclusively on the extraction and analysis of nouns since the semi-automatic approach has shown an interesting result for nouns in the construction of the Spanish WordNet.

Based on the lexical information extracted, WNB generates a set of candidate translation links, which represent the semantic relationships between Thai and English

Copyright by Mahidol University

words. Then, a subset of the candidate links are verified and used to generate a statistical model. The model is used to automatically verify the remainders of translation links. After all, the accepted translation links are used to construct the Thai WordNet by using the Princeton WordNet as a skeleton.

The contributions of our research include the WordNet Builder framework, the collection of translation links and the Thai WordNet. The framework presents the semi-automatic construction which is adaptable for other languages in order to overcome the problems of manual construction. The collection of translation links can be consider as a semantic network of words. Finally, the Thai WordNet is a new state for Thai lexical database, which can be used in many NLP applications, e.g. semantic analyser, machine translation and information retrieval/extraction.

1.3 Thesis Organization

The rest of this thesis is organized as follows:

Chapter II	*Literature Survey* provides the basic concepts, components and processes of WordNet construction. In addition, related researches, and comparison among them is summarized.
Chapter III	*Approach* explains the methodology of constructing Thai WordNet from the existing Machine Readable Dictionaries.
Chapter IV	*System Design* presents detailed design and model development methods, which include system overview as well as system design.
Chapter V	*System Implementation* shows detail of our implementation and the result of our implementation.
Chapter VI	*Discussion and Conclusion* discusses and concludes the study with an analysis of the performance, problem difficulties and that found throughout the development phase and including suggestions for future work

Copyright by Mahidol University

CHAPTER 2

LITERATURE SURVERY

In this section, we provide a brief introduction and background information about Lexical Database, Lexical Database Construction and WordNet. Then, we present an overview of the research works that are relevant to ours. Finally, we compare such research works to our work.

2.1 Lexical Database

A lexical database is a repository of data dealing with words together with additional information. It is one of the most significant components for NLP systems, which require different kinds of information Typically, the information includes phonological information (e.g. phonemes, stress, etc.), morphological information (e.g. parts of speech, concordance patterns, etc.), syntactic information (e.g. syntactic category, subcategorization, predicate/argument structure, valences, cooccurrence patterns, etc.), semantic information (e.g. semantic class, properties of the class, selectional restrictions, etc.), pragmatic information (e.g. usage, registers, topic domains, etc.) and translation links (translation link across languages).

With respect to the number of languages, lexical databases can be classified into three groups: monolingual, bilingual and multilingual. The monolingual lexical databases contain information about components of only one language. The Oxford Advanced Lerner's Dictionary (OALD) [23], Longman Dictionary of Contemporary English (LDOCE) [18], Collins Cobuild English Language Dictionary (COBUILD) [7] are the well known English lexical databases. The bilingual lexical databases contain information about language component involving two languages. The LEXiTRON and the Thaisoft So Sethaputra are two samples of the Thai⇔English lexical databases. The multilingual lexical databases contain information about language components involving more than two languages. The samples of multilingual

Copyright by Mahidol University

lexical database are the Babylon [2] which is 13-language online lexical database, and the Papillon [21], which is an English-French-Japan lexical database.

With respect to the structure, lexical databases can be classified into two types: *semantic-base* and *descriptive.* The *semantic-base* lexical databases, e.g. WordNet [19] and Thai Conceptual classification [30], organizes words with reference to relationships among them. The *descriptive* lexical databases, e.g. Oxford Advanced Lerner's Dictionary, Royal Institute Dictionary [28], and LEXiTRON [17], organize words with reference to their description.

2.2 Lexical database Construction

In the past, a lexical database is constructed manually. It has been estimated that the average time needed to manually construct a lexical entry in a lexical database is about 30 minutes[22]. If we have 1,000 lexical entries, it will take 21 days to obtain the lexical database. We can see that the manual construction is difficult, messy, laborious and time consuming.

This problem has inspired the development of various methods to construct lexical databases. Nowadays, there are many resources: Machine readable dictionaries, thesauri, encyclopedias, and corpora. The machine readable dictionaries, also known as Electronic Dictionaries, are the conventional dictionaries for human use on a computer support [27]. The thesauri are dictionaries of synonym. The encyclopedias are book or set of books containing alphabetically arranged information on many subjects. The corpora which are the collection of writings. These resources are available for lexical database construction tasks which can be computer-aided manual, semi-automatic or automatic acquisition.

In order to deal with the lexical database construction three central problems must be considered: a) What is lexical information/knowledge needed for a concrete NLP system? b) Where is this information/knowledge located? and c) Which procedure can be applied to extract/handle this information/knowledge from its sources? For the first problem, the information needed for NLP systems in general includes phonology, morphology, syntax, semantics, pragmatic knowledge as well as

Copyright by Mahidol University

translation links. However for a specific NLP system, a specific lexical database does not need all of information.

Regarding the second problem, multiple sources contain lexical information. Those sources can be classified into three types: Introspection resources, structured lexical resources (e.g. dictionaries, thesauri and encyclopedias) and unstructured lexical resources (e.g. corpora). The construction of lexical database from introspection resources has been introduced in [29], [16], [32] and [20]. The construction of lexical database from structured lexical resources has been described in [1], [3], [8], [36], [26], [27], [10], [25] (by using dictionaries) [37], [12] (using thesauri) [11] (by using encyclopedias.) The construction of lexical database from structured lexical resources has been presented in [6] and [38].

According to the third problem, the lexical acquisition approaches can be classified into three types: Machine-aided manual construction, (semi)automatic extraction from pre-existing lexical resources and combination of both. The machine-aided manual construction has been used in CYC [16], EDR [38], WordNet [20], Dutch WordNet [34], Italian WordNet [25], and LEXiTRON [24]. The (semi)automatic extraction from pre-existing lexical resources has been used Chinese-English WordNet [5]. The hybrid approach has been used in Spanish and Catalan WordNet [9]&[10].

Lexical acquisition is also dividable into two main approaches: Prescriptive and descriptive approaches. In prescriptive approach, several tools have been developed to make quicker and easier manual construction and maintenance of the lexical databases. Although manual construction is the most reliable technique, it suffers from time-consuming problem. CYC [16], EDR [38], WordNet [20], Dutch WordNet [34], Italian WordNet [25], and LEXiTRON [24] are the lexical databases that use prescriptive approach. The descriptive approach is intended to obtain lexical knowledge in an automatic or semi-automatic way from pre-existing texts. Spanish and Catalan WordNet [9]&[10], Chinese-English WordNet [5] are the lexical databases that use descriptive approach.

Copyright by Mahidol University

After the lexical database is constructed, evaluations of the result are considered. The way to evaluate the lexical database are the coverage, and the accuracy. The coverage refers to the number of the words in lexical database. The higher volume of words signify the higher coverage. The accuracy can be calculated in the form of *precision* and *recall.* According to [27], precision is the percentage of actual answers that are correct, where as the recall is the percentage of possible answers that are correct. In our work, precision is used for the measurement of accuracy.

For Thai language, most of lexical databases are constructed machine-aided manually and prescriptively. And the resources for lexical database construction are introspection, structured lexical resources (e.g. Royal Institute Dictionary [28], NAiST Lexibase [1], So Sethaputra dictionary, Saikam [30], and LEXiTRON [24]), and unstructured lexical resources(e.g. ORCHID Corpus).

An example of Thai lexical database construction is LEXiTRON, which is bilingual Thai-English and English-Thai online machine readable dictionary. It is developed by Information Research and Development Division, National Electronics and Computer Technology Center (NECTEC). LEXiTRON is a corpus-based dictionary and uses machine-aided manual construction approach. A statistical model is developed to assist lexicographers in order to construct LEXiTRON from COBUILD and English/Thai ORCHID Corpus, which contain 400,000 words. LEXiTRON contains 21,090 words from Thai-English dictionary and around 60,000 words, 80,000 meanings from English-Thai dictionary

For Thai language, it seems that no standardization for Thai lexical database used in NLP application. Most Thai lexical databases are custom-made for each specific NLP applications and they require lexicographical knowledge, take time and labor for construction. Thus we use the semi-automatic construction for Thai lexical database. English taxonomy from English WordNet is used for constructing our Thai lexical database. Thai WordNet is a new state of Thai lexical database.

Copyright by Mahidol University

2.3 WordNet

WordNet [19], more specifically *Princeton WordNet* (PWN), is an on-line lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory. The initial idea was to provide an aid to use in searching dictionaries conceptually, rather than merely alphabetically. The problem with an alphabetical thesaurus is redundant entries. English nouns, verbs, and adjectives are organized into synonym sets, each representing one underlying lexical concept. Different relations link the synonym sets (synsets). WordNet is the semantic organization, which supports synonymic, antonymic, hyponymic–hypernymic, and meronymic–holonymic relations. And not only internal part of speech semantic relations, but also cross-part of speech semantic relation are defined in the latest version of WordNet (WordNet 2.0.) WordNet is both semantic-base and descriptive lexicon, it has both semantic relation between words and provides descriptive meaning for each word. WordNet also qualifies as an ontology, which is set of concepts, such as things, events, and relations that are specified in some way in order to create an agreed-upon vocabulary for exchanging information [13].

The most ambitious feature of WordNet, however, is its attempt to organize lexical information in terms of “word meanings” (the lexicalized concept that a form can be used to express), rather than “word forms” (the physical utterance or inscription.) WordNet distinguishes between semantic relations and lexical relations; the emphasis is still on semantic relations between meanings, but relations between words are also included. WordNet is manually constructed by lexicographers. Although the manual construction can lead to the best results in terms of accuracy, it has the important drawback of its cost.

The significant increase of using wide coverage of ontologies for NLP tasks drives WordNet become a *de-facto* standard for a wide range of NLP applications. WordNet is applied to many languages such as Dutch, Italian, Spanish, French, and etc. The following subsections provide a summary of EuroWordNet , Spanish WordNet, Italian WordNet, and Chinese-English WordNet.

Copyright by Mahidol University

2.3.1 EuroWordNet

Apart from the original English version developed at Princeton University, WordNets are implemented for Dutch, Italian, Spanish, German, French, Czech and Estonian in the EuroWordNet project [33]. The most significant feature of WordNet is the semantic organization. It supports synonymic, antonymic, hyponymic–hypernymic, and meronymic–holonymic relations.

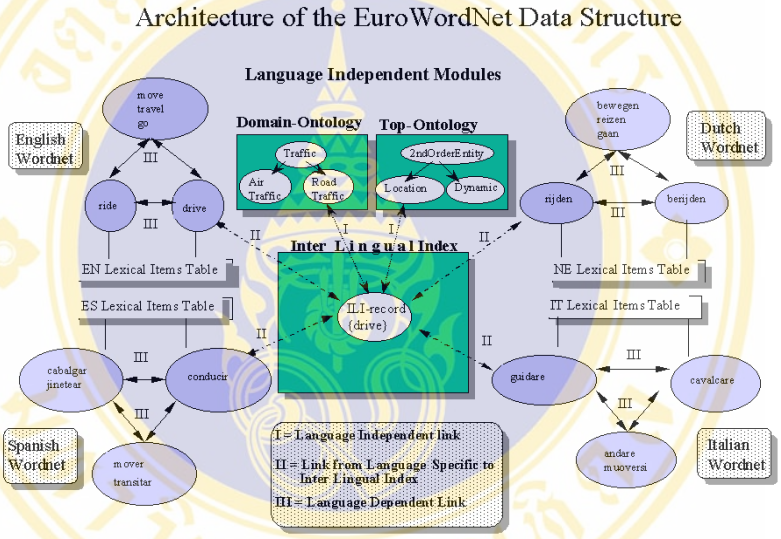


Figure 2.1 The global architecture of the EuroWordNet database.

Each of the European WordNets is a network of relations between word meanings in a specific language. The semantic relations are therefore considered as language-internal relations (Figure 2.1). In addition to the language-internal relations, each synset is linked to the closest synset in the Princeton WordNet1.5. Such a multilingual database is useful for cross-language information retrieval, for transfering of information from one resource to another or for simply comparing the different wordnets. Equivalent relations between the synsets in different languages and WordNet1.5 are made explicit in the so-called Inter-Lingual-Index (ILI). Each synset in the monolingual wordnets has at least one equivalence relation with a record in this ILI, either directly or indirectly via other related synsets. Language-specific synsets

Copyright by Mahidol University

linked to the same ILI-record should thus be equivalent across the languages, as illustrated in Figure 2.1 for the language-specific synsets linked to the ILI-record drive.

2.3.2 Spanish WordNets

Spanish WordNet is a portion of EuroWordNet project. There are many attempts to construct Spanish WordNet (including Catalan and Basque), using both semi-automatic and automatic approaches, e.g. [27], [1], and [9].

Rigau [27] combines multiple methods and uses Spanish-Spanish, Spanish-English, English-Spanish dictionary, and WordNet English-English dictionaries to construct Spanish WordNet. The Spanish-Spanish dictionary is used for constructing Spanish taxonomy. It is used with the English WordNet, the Spanish – English and the English – Spanish dictionaies to generate candidate links, which describe relationships between Spanish words and English concepts.

Eleven methods, which can be classified into 2 categories of methods Class method and Conceptual Distance method, applied to generate candidate links. Class method includes 4 monosemic (1:1, N:1, 1:M, M:N), 4 polysemic (1:1, N:1, 1:M, M:N) And Conceptual Distance methods includes using Co-occurrence words collected from DGILE (CD1), using headword and genus of DGILE (CD2) and using Spanish entries with multiple translations in the bilingual dictionary (CD3).

Two hybrid methods and four Structural methods were added in [1] to generate candidate links. Hybrid methods include Variant method and Field method. Structure methods comprises Intersection, Parent, Brother, and Distance Hyperonymy method. The result of 17 methods were intersected and added to expand the coverage of Spanish WordNet. The verification of the result of each methods both in [27] and [1] are labor-intensive task.

Logistic regression was used in [9] to construct a statistical model assist in verification of the result of the combined multiple methods [1]. This approach has overall precision amounts 86%, the highest precision and lesser labor-intensive comparing with others. This research adapt these approaches for constructing Thai WordNet from existing resources.

Copyright by Mahidol University

2.3.3 MultiWordNet (Italian Wordnet)

MultiWordNet (MWN) [25], so-called Italian WordNet is a part of EuroWordNet project. MultiWordNet is built with new synsets in correspondence with the PWN synsets. The semantic relations of English synsets has been imported to Italian language, assuming that if there are two synsets in PWN and a relation holding between them, the same relation holds between the corresponding synsets in the Italian language.

The resources used in construction of MultiWordNet are Italian ⇔ English dictionaries and English WordNet. To help the lexicographers in constructing MultiWordNet, two automatic procedures are introduced: Assign-procedure and lexical-gap procedure. In assign procedure, it takes as input one of the senses of the Italian-to-English section of the Collins dictionary and it procedures as output a set of candidates. Each candidate is described as *<PWN synset, confidence score>* pair, where *confidence score* measures the degree of confidence in the link between the Italian word sense and the *PWN synset*. Only candidates with a confidence score greater than a certain threshold are proposed to the lexicographer. The construction of MultiWordNet provides 70% precision and 63% recall.

2.3.4 Chinese-English WordNet

Chinese-English WordNet [4], [5], is constructed by using semi-automatic approach. In the construction process, it takes a Chinese monolingual dictionary called Cilin, an English WordNet, a Chinese corpus called ASBC, an English corpus called SemCor, and a set of Chinese-English dictionaries (e.g. LDC, BDC, MTIR, and Denisowski's CEDICT). Chinese-English WordNet is evaluated upon the accuracy with 69.23% precision. It has been used in information retrieval applications.

2.3.5 Comparative Study of WordNet Construction Approach

As aforementioned, due to the wideness of using WordNet as one of NLP resources has inspired many attempts in constructing WordNet in other

Copyright by Mahidol University

languages such as Chinese, English, Italian, and Spanish. Various approaches and resources are applied for constructing WordNet. The comparison of approaches and accuracy of construction of various WordNet is shown in Table 2.1.

Table 2.1 Compares approaches and accuracy of construction of WordNet variations.

	Resources	Approach	Accuracy
Chinese-English WordNet [5], [6]	**Structure&Unstructure** - Cilin(Chinese dictionary), LDC,Denisowski's CEDICT, BDC, MTIR(Chinese-English dictionary), WordNet (English dictionary), ASBC (Chinese corpus), SemCor(English corpus)	Semi-Automatic, descriptive	P-69.23% (using in IR app.)
Italian Wordnet [26]	**Structure** - Collins bilingual (Italian-English, English-Italian dictionary), WordNet (English dictionary)	Computer aided manual construction, prescriptive	P-70%, R-63%
Spanish WordNet [2], [10]	**Structure** - DGILE (Spanish dictionary), WordNet (English dictionary) , LDOCE (Spanish-English, English-Spanish dictionary)	Semi-Automatic, descriptive	85.80%
WordNet [21]	**Introspection**, existing English dictionary	Computer aided manual construction, prescriptive	-
Thai WordNet	**Structure** - LEXiTRON (English-Thai, Thai-English dictionary), WordNet (English dictionary)	Semi-Automatic, descriptive	76%

According to Table 2.1, Spanish WordNet was the highest accurate method (85.80% precision) follow by Italian WordNet (70% precision and 63% recall) and Chinese-English WordNet (69.23% accuracy in IR application). Resources for constructing WordNet in Spanish and Italian language are structure resources. Chinese WordNet uses both structure and unstructured resources. Italian WordNet uses computer aided manual construction approach. Spanish and Chinese WordNet use semi-automatic approach in constructing. Our research adapts Spanish WordNet methodology for construct Thai WordNet, because Spanish WordNet provides the highest accurate method and it is a semi-automatic construction, which use less laborious and time consume. And this methodology appropriate for our existing resources.

In summary, we focus on a semi-automatic construction of WordNet lexical database. We use existing resources which are machine readable dictionaries e.g. bilingual dictionary (LEXiTRON) and monolingual dictionaries (WordNet). We employ multiple criteria to create candidate sets of translation link between Thai⇔English words, and use stratified sampling to sample translation link for

Copyright by Mahidol University

verification. Afterward logistic regression is used to assist in evaluation of translation links. When the translation links are generated, English WordNet is translated into Thai using the generated translation links. Finally all English semantic relations are imported to Thai WordNet lexical database.

Copyright by Mahidol University

CHAPTER 3

THAI WORDNET CONSTRUCTION ARCHITECTURE

In this section, we present an operational approach to construct the Thai WordNet lexical database from existing machine readable dictionaries (MRDs). Our approach is operated via a system, called *WordNet Builder* (WNB). An overview of the WNB system is shown in Figure 3.1. System components can be classified into two groups: storage and operational components. Storage components include the MRDs and the Thai WordNet. Operational components include the MRD extractors, the link analyzer and the WordNet constructor.

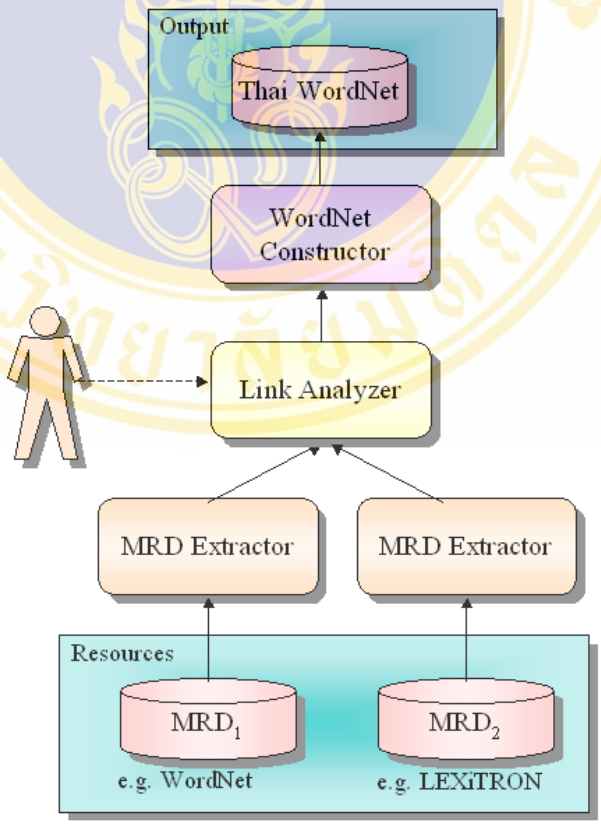


Figure 3.1 WordNet Builder (WNB) system

Copyright by Mahidol University

MRDs. We relate the MRDs as the electronic version of standard dictionaries which may contain other lexicographic information that does not appear in the printed version. MRDs can be monolingual, bilingual and multilingual. Any number of MRDs can be taken into account. However, one of monolingual MRDs and one of bilingual MRDs are current focus of this research. The focal monolingual MRD is the Princeton WordNet[20]. It provides a lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory. The focal bilingual MRD is the LEXiTRON[24] which provides a bilingual corpus-based lexical reference system.

Thai WordNet (TWN). We relate the TWN as a collection of Thai words and their relationships. In this research, we modeled the TWN after the Princeton WordNet which organizes words by meaning of the words and identifies the relationships among the words via reference links. Our TWN can be considered as a lexical database which contains information about words for a particular set of languages. We use the Database Management System (DBMS) to help manage the lexical database.

MRD Extractor. This component converts the data from multiple resources with different format into a common format. It cleans the noisy, erroneous, missing, irrelevant and duplicate data. It joins and aligns the scattered data for smoothly access and selects the relevant data that will be analyzed by the Link Analyzer.

Link Analyzer. This component analyzes, classifies and generates candidates of translation link between Thai⇔English word, samples candidate of translation link, and then verifies candidates by taking advantage of the statistic approach to reduce human intervention and time consuming.

WordNet Constructor. This component construct Thai WordNet from the verified Translation Link and WordNet. WordNet Constructor attaches Thai word to English SynSet, infers relations of Thai WordNet from English WordNet, and then attaches glossary from Bilingual dictionary to Thai WordNet. The detail of each operational components is described in the following sections..

Copyright by Mahidol University

3.1 MRD Extractor

The MRD Extractor perform several processes to extract and transform data from two main machine readable dictionaries: WordNet and LEXiTRON. We divides the tasks of MRD Extractor into four processes: Conversion, Cleansing, Alignment and Selection as shown in Figure 3.2.

Figure 3.2 MRD Extractor

Conversion – converts data from various encoding formats to a common format. The existing resources encoding in different formats, WordNet is SQL script and LEXiTRON is XML format. The Conversion transforms data from different format into the same common format, relational database, for more convenient in accessing and retrieving data.

Cleansing – handles noisy, erroneous, missing or irrelevant data that usually come with the existing resources. LEXiTRON has missing and noisy records, and duplicated records. And WordNet also has missing fields. Therefore for the correctness and accuracy of the result, the data cleansing is performed to eliminate such kind of data.

Alignment – joins and rearranges the data for more convenience access and retrieval. LEXiTRON, the bilingual dictionaries, contain heterogeneous data: English-Thai and Thai-English dictionaries. English-Thai and Thai-English dictionaries, the data are in separate data source which is inconvenience in access and retrieval. To avoid the redundant and erroneous data when access and retrieve data from both, then

Copyright by Mahidol University

heterogeneous data sources are combined into one. For example, the translation "water"-"น้ำ" in English–Thai dictionary and "น้ำ"-"water" in Thai-English dictionary, both are same meaning then one of them are selected and the redundant translation links are removed.

Selection – choose the relevant data for using in next component. In as much as there are irrelevant data, both in LEXiTRON and WordNet. In this research, we focus on noun category because in Candidate Link Construction module are proved for noun word. Then our research emphasizes only on nominal words then verbal, adjectival and adverbial words are ignored. Irrelevant data in LEXiTRON and WordNet are ignored. Furthermore, translation link in which English words are not exist in both LEXiTRON and WordNet are also discarded.

3.2 Link Analyzer

After the *MRD Extractors* prepared the resources, the *Link Analyzer* analyzes and classifies translation links between Thai-English words by using a set of criteria, samples of translation links from each criteria, manually verify sampled translation links is manually verified. Finally, a model is constructed and deployed to verify the remaining translation links. Then the result of this component is the disambiguated translation links between Thai words and English concepts. This component comprises 5 module includes *Candidate Link Construction, Sampling, Verification, Model Construction, Model Evaluation* and *Model Employment.* The illustration of the Link Analyzer is shown as in Figure 3.3

Copyright by Mahidol University

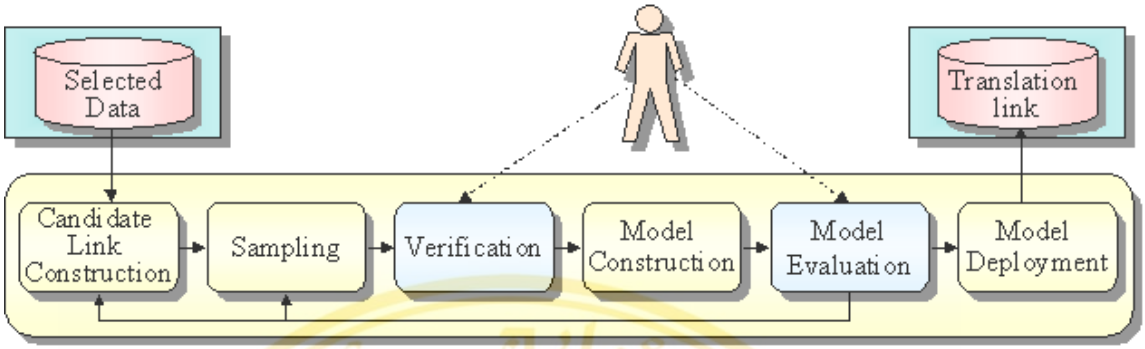


Figure 3.3 Link Analyzer

Candidate Link Construction – analyze and classify translation link between Thai and English based on 13 criteria as shown in Figure 3.4. These criteria can be categorized into two groups: Class and Structure. In the class group, link between Thai and English are used to analyze and classify translation links. In the structure group, WordNet structure is used to analyze and classify translation links.

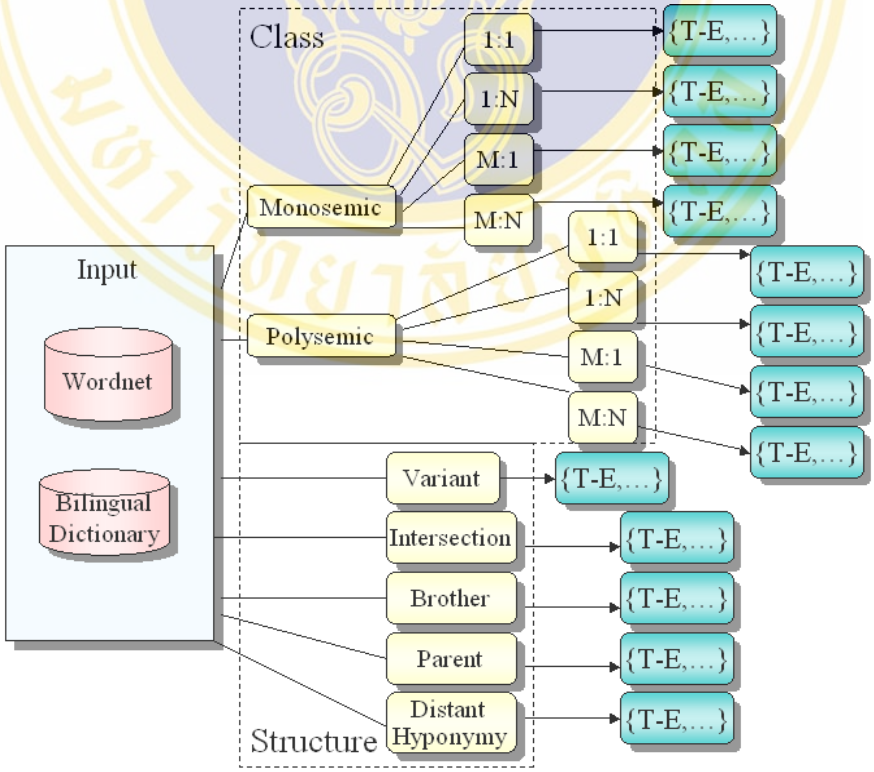


Figure 3.4 Multiple Criteria

Copyright by Mahidol University

Class criteria group classifies the translation between Thai and English words using two criteria: Monosemic criteria (English words which have one meaning) and Polysemic criteria (English words which have more than one meaning). Each can be divided into 4 sub-criteria: 1:1, 1:N, M:1, and M:N (considering the link between Thai–English words). The illustration of the Class criteria group is shown as in Figure 3.5 and the description of each criteria is explained as follows:

Monosemic criteria – These criteria apply only to monosemous English words with respect to WordNet1.7. As a result, this unique synset is linked to the corresponding Thai words.

- Monosemic (1:1) A Thai Word (TW) has only one English translation (EW); symmetrically, EW has TW as its unique translation.
- Monosemic (1:N) A TW has more than one translation; each EW has TW as its unique translation.
- Monosemic (M:1) Several TWs have the same translation; EW has several translations to Thai.
- Monosemic (M:N) Several TWs have different translations; EWs also have several translations.

Polysemic criteria – Applying the corresponding cases for those English polysemic nouns in WordNet1.7, criteria polysemic 1 to 4 have been obtained alike monosemic criteria.

Copyright by Mahidol University

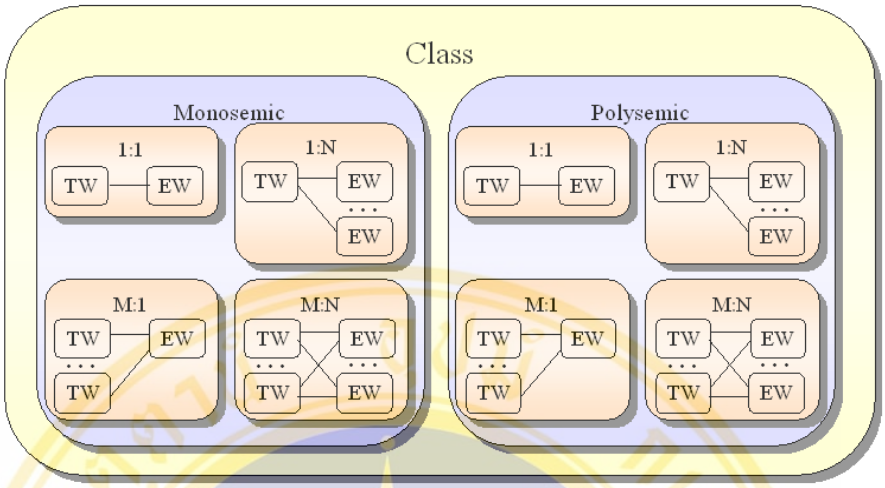


Figure 3.5 Class criteria group

In Structure criteria group, it takes profit of WordNet structure. It consists of 5 criteria :- Variant criteria, Intersection criteria, Parent criteria, Brother criteria, and Distant hyponymy criteria. The illustration of the Structure criteria group is shown as in Figure 3.6 and the detail of each criteria is described as follows:

Variant criteria – For a WN1.7 synset which contains a set of variants EWs1, if it is the case that two or more of the variants EWi have only one translation to the same TW, a link is produced for TW into the WN1.7 synset.

Intersection criteria – All EWs share at least one common synset in WordNet Link TW is linked to all common synsets of its translations

Parent criteria – A synset of an EW is direct parent of synsets corresponding to the rest of EWs Link. The TW is linked to all hyponym synsets

Brother criteria – All EWs have synsets which are brothers respecting to a common parent Link. The TW is linked to all co-hyponym synsets.

Distant hyponymy criteria – A synset of an EW is a distant hypernym of synsets of the rest of the EWs Link. The Thai Word is linked to the lower level (hyponym) synsets.

Copyright by Mahidol University

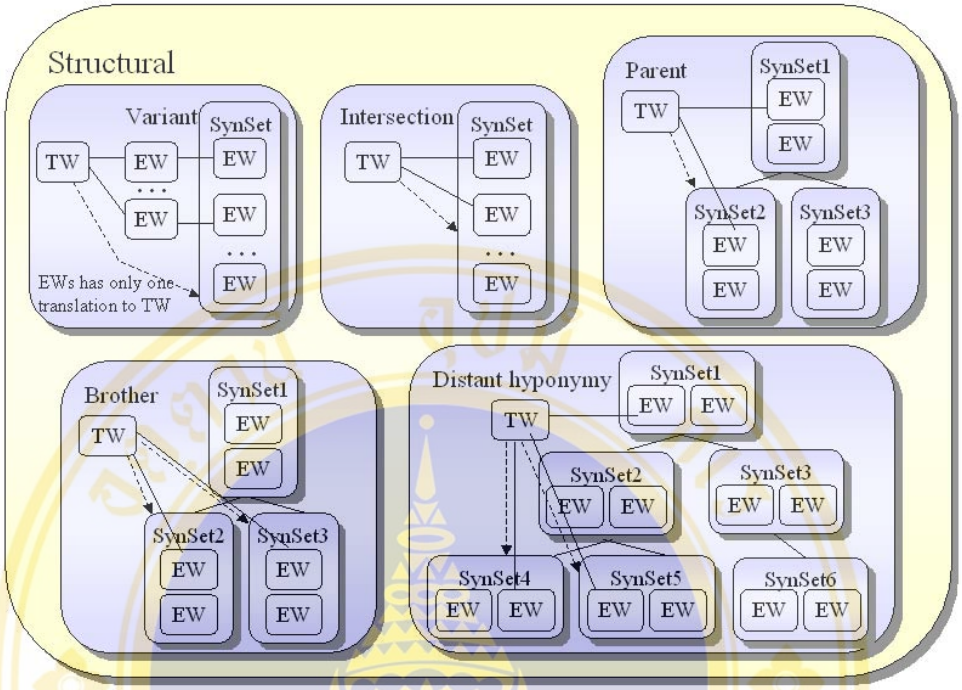


Figure 3.6 Structure criteria group

The result of this process are sets of translation link of each criteria and matrix of existence of translation link in each criteria as illustrated below (Table 3.1).

Table 3.1 Matrix of existence of translation link in each criteria

Translation Link	Criteria 1	Criteria 2	…	Criteria 13
กงสุล-consul	1	0	…	0
กล้วย- banana	0	1	…	1
รูปวงรี- ellipse	1	0	…	1
…	…	…	…	…

Sampling– sample candidate set of translation link using the random stratified sample, sampling candidate set of translation link from each criteria.

Verification – extracts and manually verifies the random stratified sample of links covering all the sets of candidate. And then the translation links are automatically evaluated by using existing resources (bilingual dictionary) and the remaining, which are not exist in existing resources, are manually evaluated by human. The specification

Copyright by Mahidol University

of evaluator is the expert in both Thai and English language, in our research the specification of evaluator is compromised to non-expert but using standard well-known bilingual dictionary in evaluation. And the result of the evaluation is the matrix of boolean variables of each criteria. as illustrated in Table 3.2.

Table 3.2 The result of evaluation Matrix

NOK	NTOT	Criteria 1	Criteria 2	…	Criteria 13
3	7	1	0	…	0
4	8	0	1	…	1
5	10	1	0	…	1
…	…	…	…	…	…

NOK (number of correct evaluations for the set of the results of every method) and NTOT (total number of evaluations.)

Model Construction – finds model for approximate the probability of that translation link would be correct, and finds the correlation and significant of the criteria. The result of evaluation matrix is used as a training set to construct model. Logistic regression, which used to predict a discrete outcome from a set of binary variables, is used for finding the model for approximating P(OK). The probability that a link would be correct can be estimate by P(OK)=NOK/NTOT, where OK is the Correct link. And the least squares criterion is used for estimating every value of β_i (unknown parameters) to fit the formula $\log(P(nok/ntot)) = \beta_0+\beta_1 C_1 +...+\beta_{13} C_{13}$. The result of this process is the factors resulting from regression, the example is shown as following:

C_i	β_I
C_0	0.62
C_1	1.39
C_2	1.20
C_3	1.39

C_i	β_i
C_4	0.62
C_5	1.39
C_6	1.20
C_7	1.39

C_i	β_i
C_8	0.62
C_9	1.39
C_{10}	1.20
C_{11}	1.39

C_i	β_i
C_{12}	0.62
C_{13}	1.39

Copyright by Mahidol University

Model Evaluation – evaluates the model. By using Pearson goodness-of-fit, which determines the significance of model. If the result is acceptable, the model is deployed for the remaining translation link.

Model Employment – applies the model to all of the remaining set of translation link, if the model is acceptable. The outcome of this module is the Thai-English Translation Links that have probability. Calculate the probability of each link from the results of regression model where C_i is the value of the i^{th} factors resulting from regression, and β_i is the coefficient i^{th} of the factor. For Spanish WordNet 86 % of P(OK) is used, also with Thai WordNet.

$$P(OK) = \frac{e^{\Sigma\, Ci\beta i}}{1+e^{\Sigma\, Ci\beta i}}$$

3.3 WordNet Constructor

The last component, WordNet constructor, uses evaluated Thai-English translation links (the result of the previous process) and the existing English WordNet to attach Thai words to English WordNet. This component comprises three processes: SynSet Attachment, Relations Inference, and Glossary Attachment. The illustration of WordNet Constructor is shown as in Figure 3.7

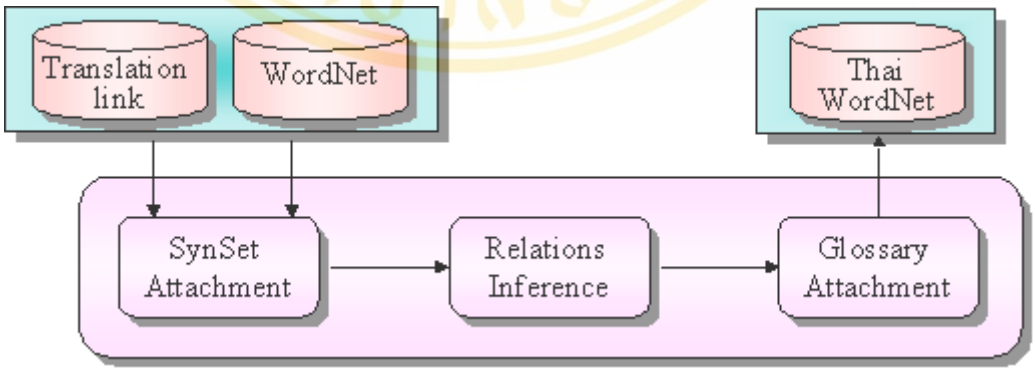


Figure 3.7 WordNet Constructor

SynSet Attachment – uses English WordNet as a skeleton of Thai WordNet, translates each English word into Thai, and attaches English synset to each Thai word.

Copyright by Mahidol University

Relations Inference – infers the internal relations (e.g. Hypernym, Hyponym, Meronym, and Holonym) and apply transitivity.

Glossary Attachment – attaches Glossary information to Thai WordNet, for example sample, definition, and etc., to fulfill Thai WordNet.

The following we will give more discussion on the transitivity which is used is Relations Inference. "Transitivity is a mathematical property of binary relations such that if A and B are related, and B and C are related, then it follows that A and C are also related, for all A, B, and C for which the relation may apply. The relation is then said to be *transitive*. In notation, this is: $aRb \wedge bRc \Rightarrow aRc$ [35]. The relations between words and synsets in English WordNet can also be applied this property, then relations in English WordNet will transitive to Thai WordNet. The internal relations of English WordNet will be inferred to Thai internal relations, as illustrated in Figure 3.8 (where S1 stands for SynSet1, S2-SynSet2, W1-Word1,TW1-Thai Word1, and TW2-Thai Word2). From the illustration, English Word1 in SynSet1, has internal relation with English Word1 in SynSet2, if both English Words has Translation Link to Thai words then the internal relation will infer to Thai as well. That means Thai Word1 and Thai Word2 will have internal relation similar to English Word1 and English Word2. All the possible relations will be inferred and constructed. The examples of inference the internal relations are illustrated in Meronym or HAS A relation

In Figure 3.9a shows the transitivity of the hypernym relation. The "Vehicle" can be translated into "ยานพาหนะ", the "Car" can be translated into "รถยนต์", and the "Vehicle" is hypernym of "Car". Therefore, the "ยานพาหนะ" is a hypernym of "รถยนต์". In Figure 3.9b "Hand" can be translated into "มือ", the "Finger" can be translated into "นิ้ว" and the "Hand" is meronym of "Finger". Therefore, "มือ" is meronym of "นิ้ว".

Copyright by Mahidol University

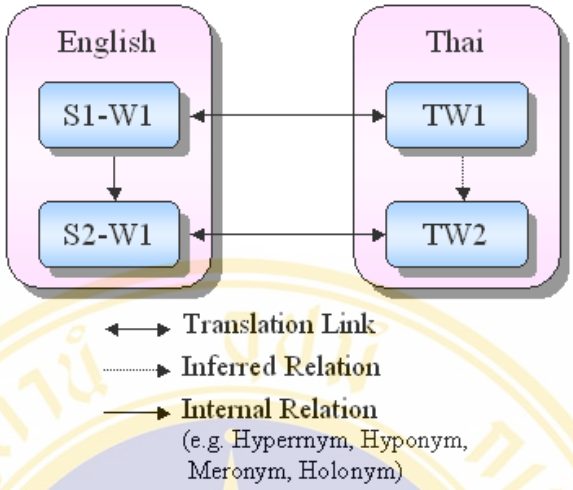
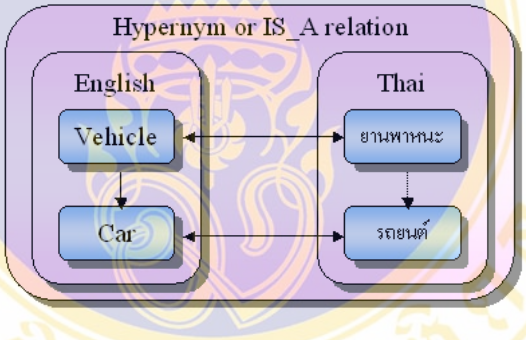
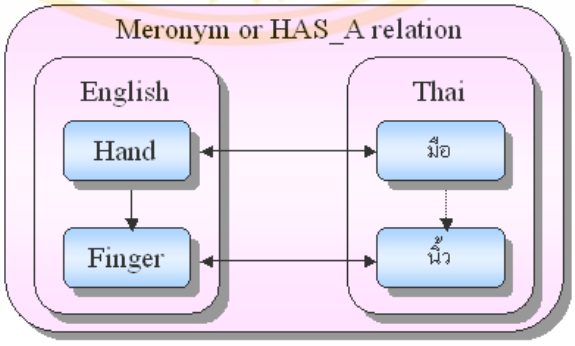


Figure 3.8 Inference Internal relation



a) Hyponym or IS A relation



b) Meronym or HAS A relation

Figure 3.9 The example of inference the internal relations

Copyright by Mahidol University

CHAPTER 4

SYSTEM DESIGN

In this section, we present system design for WordNet Builder, which include structure chart, data flow diagrams, module specifications, file structure and input-output design.

4.1 Structure Chart

The structure chart is commonly used to show how system modules are organized. Figure 4.1 displays the structure chart of Building Thai WordNet. With respect to The architecture of WordNet Builder presented in Chapter 3, Building Thai WordNet is composed of three main processes.

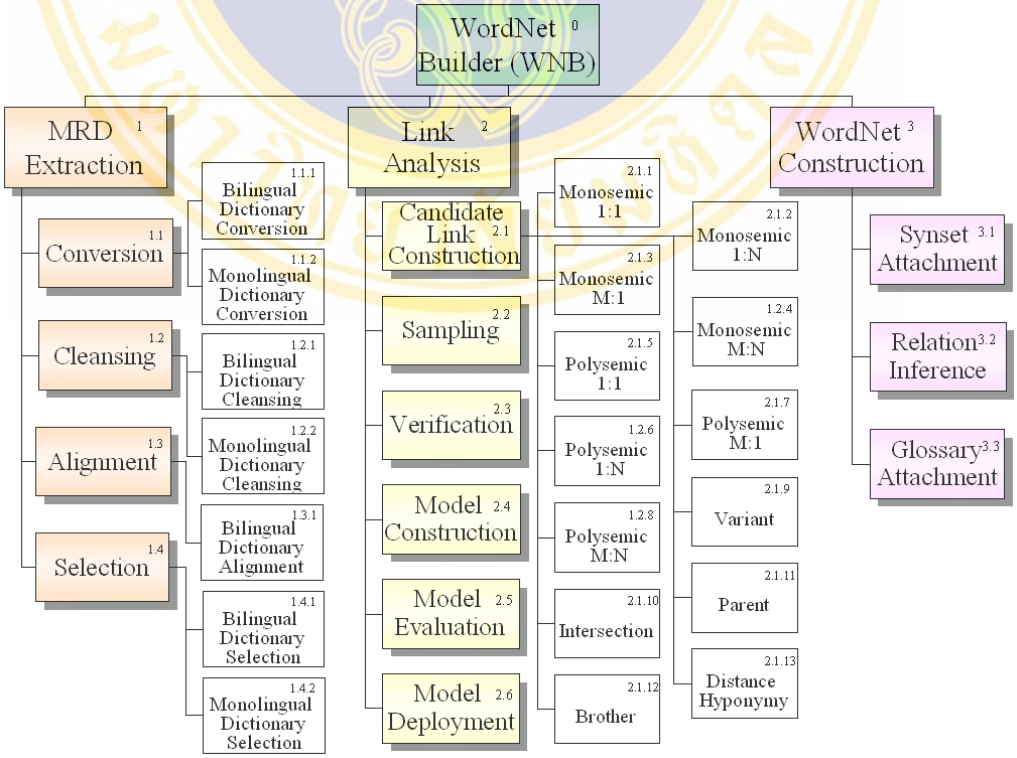


Figure 4.1 Structure Chart of WNB

Copyright by Mahidol University

MRD Extraction consists of four steps: Conversion, Cleansing, Alignment, and Selection. Conversion, Cleansing, and Selection have two procedures, the first procedure involves with bilingual dictionary and the last procedure involves with monolingual dictionary. But Alignment only concern with bilingual dictionary. The Link Analysis consists of six steps: Candidate link Construction, Sampling, Verification, Model construction, Model Evaluation, and Model Employment. The Candidate Link Construction comprise of 13 criterion including Monosemic 1:1, M:1, 1:N, M:N, Polysemic 1:1, M:1, 1:N, M:N, Variant, Intersection, Parent, Brother, and Distance Hyponymy The WordNet Construction comprise of three steps including Synset Attachment, Relation Inference, and Glossary Attachment.

4.2 Top-Level Data Flow Diagrams and Process Description

A Data Flow Diagram (DFD) is most commonly used to represent a conceptual model of systems. Figure 4.2 and Figure 4.3 show the Data Flow Diagram level 0 and 1, respectively, of WordNet Builder system.

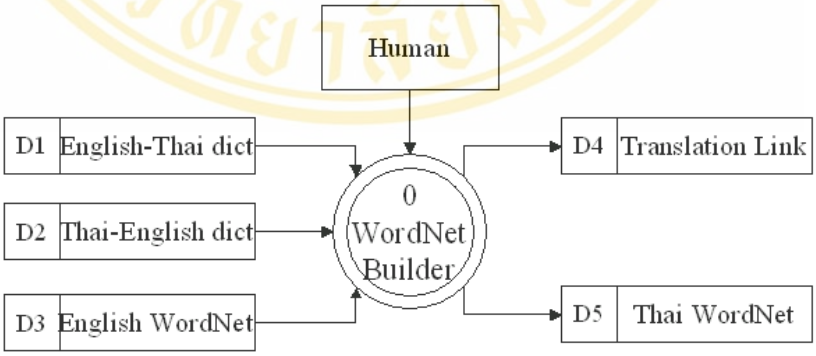


Figure 4.2 Context Diagram: WordNet Builder

Copyright by Mahidol University

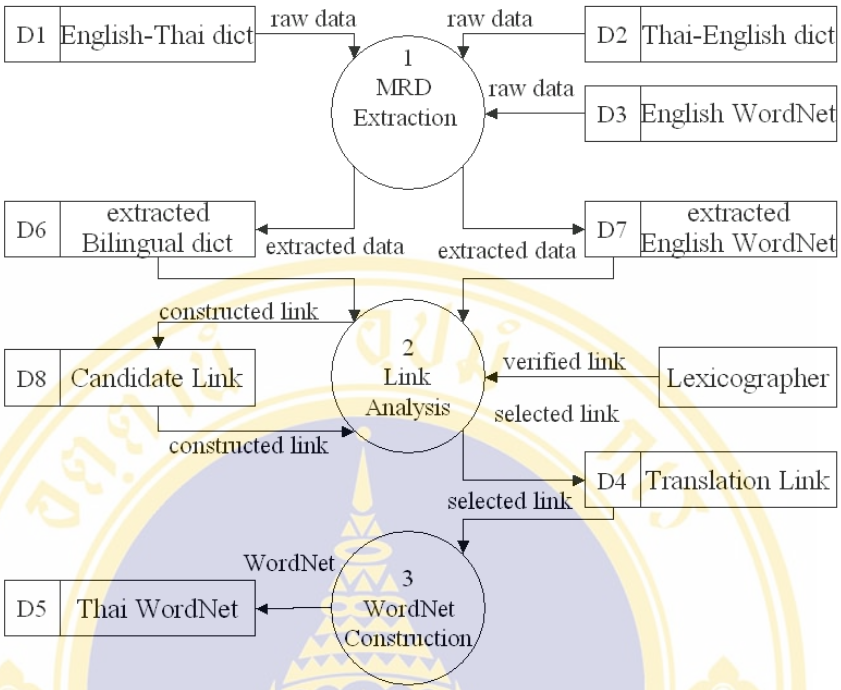


Figure 4.3 Data Flow Diagram Level 1: WordNet Builder.

With respect to the Data Flow Diagram in Figure 4.3, the following is the process description of WordNet Builder.

Module Name	:	0. WordNet Builder
Description	:	The construction of Thai WordNet lexical database from existing machine readable dictionaries (MRDs).
Input	:	English-Thai dictionary, Thai-English dictionary, and English WordNet
Output	:	Thai WordNet and Translation link between Thai word and English concept

Module Name	:	1. MRD Extraction
Description	:	This module is used to convert, clean, align, and select the raw data ready for constructing translation links.
Input	:	English-Thai dictionary, Thai-English dictionary, and English WordNet
Output	:	Extracted Bilingual dictionary and Extracted Monolingual dictionary

Copyright by Mahidol University

Module Name : 2. Link Analysis
Description : This module is used to construct the candidates of translation link (link between Thai word and English concept), evaluate and select them using statistical methods.
Input : Extracted Bilingual dictionary and Extracted Monolingual dictionary.
Output : Translation Links.

Module Name 3. WordNet Construction
Description : This module is used to construct Thai WordNet and fulfill Thai WordNet with Thai glossary.
Input : Translation Links and Extracted Monolingual dictionary.
Output : Thai WordNet.

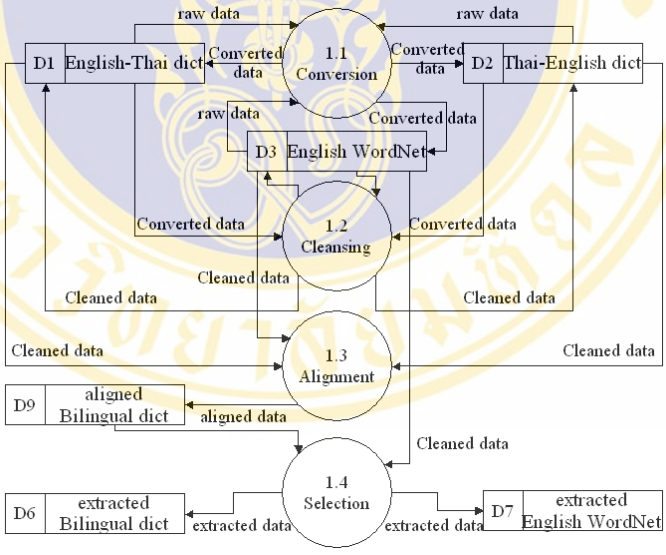


Figure 4.4 Data Flow Diagram Level 2: MRD Extraction

Figure 4.4 shows the process of MRD Extraction via the Data Flow Diagram level 2. The process description is described as follows:

Copyright by Mahidol University

Module Name	:	1.1 Conversion
Description	:	This module is used to converted resources to the same format (Relational Database, RDB). - English-Thai dictionary and Thai-English dictionary are converted from XML to Relational Database. - English WordNet is converted from text file (SQL script) to Relational Database.
Input	:	English-Thai dictionary, Thai-English dictionary and English WordNet.
Output	:	English-Thai dictionary, Thai-English dictionary and English WordNet in form of RDB

Module Name	:	1.2 Cleansing
Description	:	This module is used to clean the data
Input	:	English-Thai dictionary, Thai-English dictionary and English WordNet in form of RDB
Output	:	Cleansed English-Thai dictionary, Cleansed Thai-English dictionary and Cleansed English WordNet.

Module Name	:	1.3 Alignment
Description	:	This module is used to align cleansed English-Thai dictionary and cleansed Thai-English dictionary together for the conveniently use and remove redundant data.
Input	:	Cleansed English-Thai dictionary and Cleansed Thai-English dictionary
Output	:	Aligned bilingual dictionary

Module Name	:	1.4 Selection
Description	:	This module is used to select only the relevant data, the nominal word and relevant data.
Input	:	Aligned bilingual dictionary and cleansed English WordNet.
Output	:	Selected bilingual dictionary and selected English WordNet.

Copyright by Mahidol University

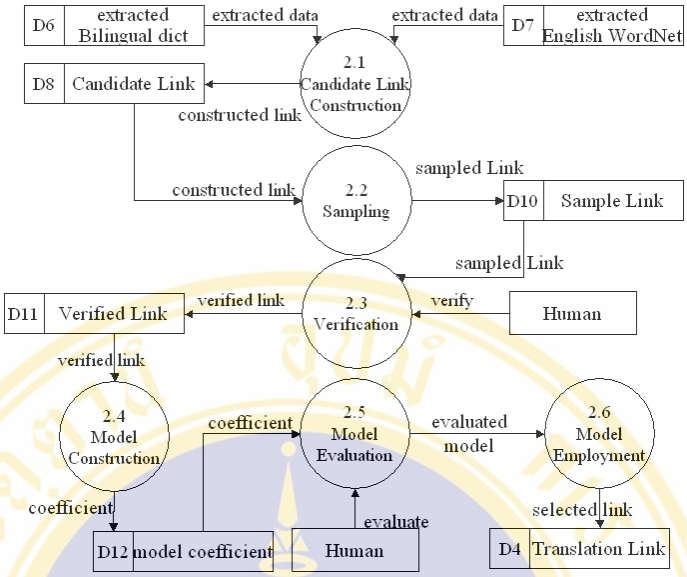


Figure 4.5 Data Flow Diagram Level 2: Link Analysis

Figure 4.5 shows the process of Link Analysis via the Data Flow Diagram level 2. The process description is described as follows:

Module Name : 2.1 Candidate Link Construction
Description : This module perform translation link construction. Using 13 criteria in construction, to construct candidate of translation links (link between Thai word and English concept). And find existence of each link in all criteria.
Input : Selected bilingual dictionary and selected English WordNet.
Output : Candidate links, the matrix of existence of translation link in all criteria.

Module Name : 2.2 Sampling
Description : This module is used to sample the candidate links for verify. The stratified sampling is used.
Input : Candidate links
Output : Sample of candidate links

Module Name : 2.3 Verification
Description : This module is used to verify the sample of candidate link manually, in the sense of linguistic.
Input : Sample of candidate links
Output : Verified sample

Copyright by Mahidol University

Module Name : 2.4 Model Construction
Description : This module is used to construct logistic regression model for predicted the remaining candidate link
Input : Verified sample
Output : The model (coefficient of each factor)

Module Name : 2.5 Model Evaluation
Description : This module is used to evaluate the logistic regression model manually. To evaluate that the model are fitting for the data. And If the model is not fit, factor removal will be performed.
Input : The model (coefficient of each factor)
Output : Evaluated model (coefficient of each factor)

Module Name : 2.6 Model Employment
Description : This module is used to employ the model to the remaining candidate links.
Input : Evaluated model (coefficient of each factor) and remaining candidate links
Output : Selected translation link

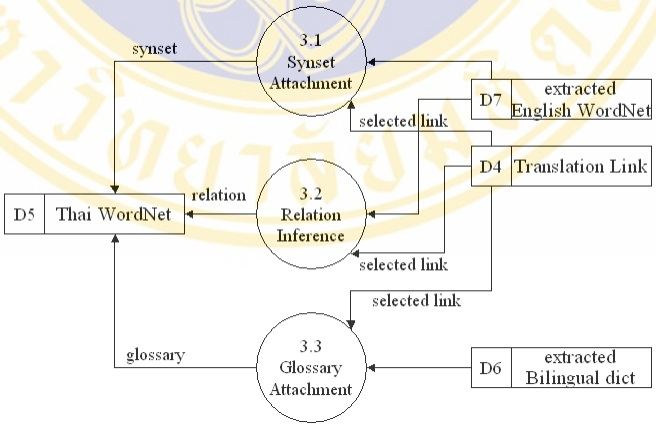


Figure 4.6 Data Flow Diagram Level 2: WordNet Construction

Figure 4.6 shows the process of WordNet Construction via the Data Flow Diagram level 2. The process description is described as follows:

Copyright by Mahidol University

Module Name : 3.1 SynSet Attachment
Description : This module is used to attach Thai word to English synset, to construct Thai WordNet.
Input : Selected translation link and selected English WordNet.
Output : Thai Synset (Thai WordNet)

Module Name : 3.2 Relation Inference
Description : This module is used to infer the relation of English structure
Input : Selected translation link and selected English WordNet.
Output : Relation between Thai synset (Thai WordNet)

Module Name : 3.3 Glossary Attachment
Description : This module is used to save all testing result and their error into the database.
Input : Selected bilingual dictionary and Thai WordNet
Output : Thai WordNet

4.3 Low-Level Data Flow Diagrams and Module Specifications

The module specification represents the programming logic of each program module in WordNet Builder including MRDs Extraction and Link Analysis. The following is the Data Flow Diagram level 3 of MRDs Extraction.

Copyright by Mahidol University

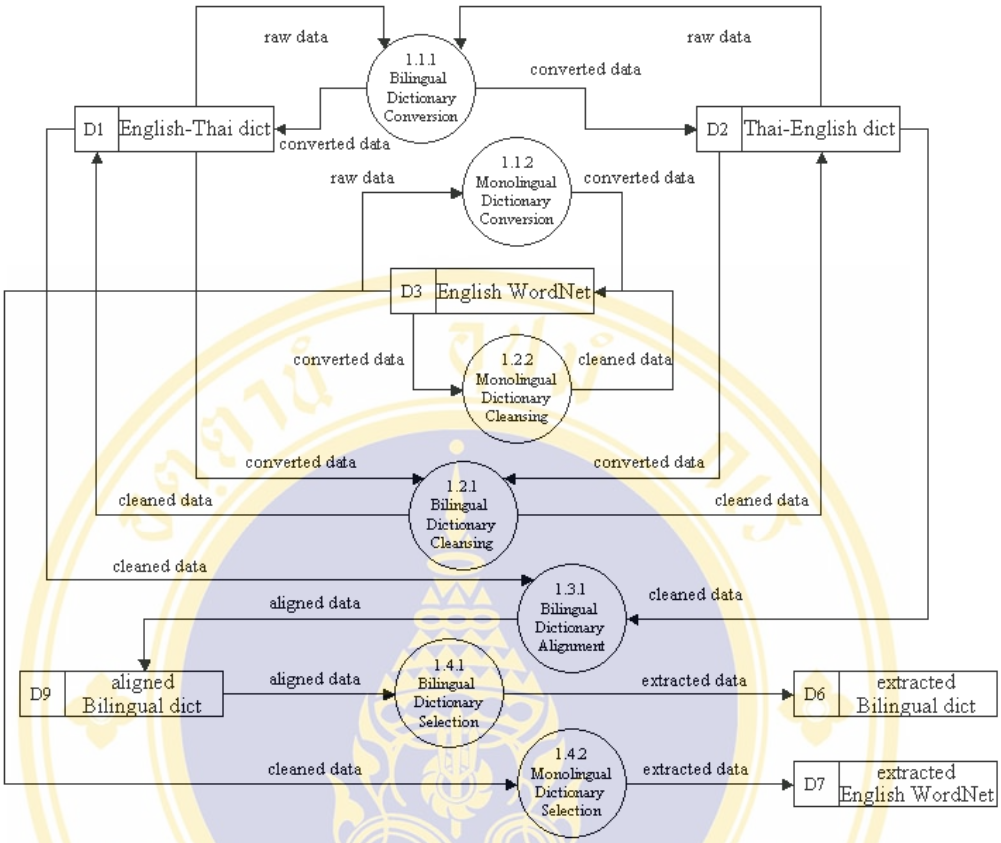


Figure 4.7 Data Flow Diagram Level 3: MRDs Extractor

Figure 4.7 shows the process of MRDs Extraction via the Data Flow Diagram level 3. The process description is described as follows:

Module Name : 1.1.1 Bilingual Dictionary Conversion
Description : This module converts bilingual dictionaries which are in XML-like format into relational database format.
Input : Bilingual dictionaries which are non-relational database format.
Output : Bilingual dictionaries which are relational database format.
Logic Summary :

1. Analyze and design database model for bilingual dictionaries (LEXiTRON).
2. Create tables for Thai-English and English-Thai dictionaries.
3. Load XML data into relational database tables.

Copyright by Mahidol University

Module Name : 1.1.2 Monolingual Dictionary Conversion
Description : This module converts monolingual dictionary which is in SQL script into relational database format.
Input : Monolingual dictionary which are non-relational database format.
Output : Monolingual dictionary which are relational database format.
Logic Summary :

1. Create database for monolingual dictionary (WordNet).
2. Load SQL script for monolingual dictionary.

Module Name : 1.2.1 Bilingual Dictionary Cleansing
Description : This module cleans bilingual dictionaries.
Input : Bilingual dictionaries which are relational database format.
Output : Cleaned bilingual dictionaries.
Logic Summary :

1. Detect anomalous data.
2. Filling missing data.
3. Removing wrong spelling.
4. Removing duplicate record.

Module Name : 1.2.2 Monolingual Dictionary Cleansing
Description : This module cleans monolingual dictionary.
Input : Monolingual dictionary which are relational database format.
Output : Cleaned monolingual dictionary.
Logic Summary :

1. Detect anomalous data.
2. Filling missing data.

Module Name : 1.3.1 Bilingual Dictionary Alignment
Description : This module merges bilingual dictionaries which includes Thai-English and English-Thai dictionaries.
Input : Cleaned Thai-English and English-Thai dictionaries.
Output : Aligned bilingual dictionary.
Logic Summary :

1. Design and create new table for both Thai-English and English-Thai dictionaries.
2. Insert Thai-English dictionary into new table.
3. Insert English-Thai dictionary which translation link between English word and Thai word are not exist the new table.
4. Detect anomalous data.
5. Removing duplicate record.

Copyright by Mahidol University

Module Name : 1.4.1 Bilingual Dictionary Selection
Description : This module selects nominal word from bilingual dictionary.
Input : Aligned bilingual dictionary.
Output : Selected bilingual dictionary.
Logic Summary :

1. Analyze and select noun related table in bilingual dictionary.
2. SELECT noun word FROM bilingual dictionary.

Module Name : 1.4.2 Monolingual Dictionary Selection
Description : This module selects nominal word from monolingual dictionary.
Input : Cleaned monolingual dictionary dictionary.
Output : Selected monolingual (English WordNet) dictionary.
Logic Summary :

1. Analyze and select noun related table in monolingual dictionary.
2. SELECT noun word FROM English WordNet.

The following is the Data Flow Diagram level 3 of Candidate Link Construction which is a process in Link Analysis.

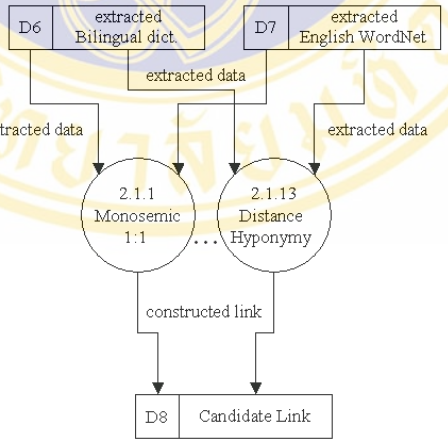


Figure 4.8 Data Flow Diagram Level 3: Candidate Link Construction

Figure 4.8 shows the process of Candidate Link Construction via the Data Flow Diagram level 3. The process description is described as follows:

Copyright by Mahidol University

Module Name : 2.1.1 Monosemic1:1
Description : This module perform translation link construction :
A Thai Word (TW) has only one English translation (EW); symmetrically, EW has TW as its unique translation.
Input : Selected bilingual dictionary and selected English WordNet.
Output : Candidate links
Logic Summary :

1. SELECT monosemic noun FROM Selected English WordNet
2. JOIN WordNet and bilingual dictionary ON English word
3. SELECT FROM 2 WHERE COUNT (English word)=1
4. SELECT FROM 2 WHERE COUNT(Thai word)=1
5. SELECT FROM 2 WHERE English word IN 3 AND Thai word IN 4

Module Name : 2.1.2 Monosemic1:N
Description : This module perform translation link construction :
A TW has more than one translation; each EW has TW as its unique translation.
Input : Selected bilingual dictionary and selected English WordNet.
Output : Candidate links
Logic Summary :

1. SELECT monosemic noun FROM Selected English WordNet
2. JOIN WordNet and bilingual dictionary ON English word
3. SELECT FROM 2 WHERE COUNT (English word)>1
4. SELECT FROM 2 WHERE COUNT(Thai word)=1
5. SELECT FROM 2 WHERE English word IN 3 AND Thai word IN 4

Module Name : 2.1.3 MonosemicM:1
Description : This module perform translation link construction :
Several TWs have the same translation; EW has several translations to Thai.
Input : Selected bilingual dictionary and selected English WordNet.
Output : Candidate links
Logic Summary :

1. SELECT monosemic noun FROM Selected English WordNet
2. JOIN WordNet and bilingual dictionary ON English word
3. SELECT FROM 2 WHERE COUNT (English word)=1
4. SELECT FROM 2 WHERE COUNT(Thai word)>1
5. SELECT FROM 2 WHERE English word IN 3 AND Thai word IN 4

Copyright by Mahidol University

Module Name : 2.1.4 MonosemicM:N
Description : This module perform translation link construction : Several TWs have different translations; EWs also have several translations.
Input : Selected bilingual dictionary and selected English WordNet.
Output : Candidate links
Logic Summary :

1. SELECT monosemic noun FROM Selected English WordNet
2. JOIN WordNet and bilingual dictionary ON English word
3. SELECT FROM 2 WHERE COUNT (English word)>1
4. SELECT FROM 2 WHERE COUNT(Thai word)>1
5. SELECT FROM 2 WHERE English word IN 3 AND Thai word IN 4

Module Name : 2.1.5 Polysemic 1:1
Description : This module perform translation link construction : A Thai Word (TW) has only one English translation (EW); symmetrically, EW has TW as its unique translation.
Input : Selected bilingual dictionary and selected English WordNet.
Output : Candidate links
Logic Summary :

1. SELECT polysemic noun FROM Selected English WordNet
2. JOIN WordNet and bilingual dictionary ON English word
3. SELECT FROM 2 WHERE COUNT (English word)=1
4. SELECT FROM 2 WHERE COUNT(Thai word)=1
5. SELECT FROM 2 WHERE English word IN 3 AND Thai word IN 4

Module Name : 2.1.6 Polysemic 1:N
Description : This module perform translation link construction : A TW has more than one translation; each EW has TW as its unique translation.
Input : Selected bilingual dictionary and selected English WordNet.
Output : Candidate links
Logic Summary :

1. SELECT polysemic noun FROM Selected English WordNet
2. JOIN WordNet and bilingual dictionary ON English word
3. SELECT FROM 2 WHERE COUNT (English word)>1
4. SELECT FROM 2 WHERE COUNT(Thai word)=1
5. SELECT FROM 2 WHERE English word IN 3 AND Thai word IN 4

Copyright by Mahidol University

Module Name : 2.1.7 Polysemic M:1
Description : This module perform translation link construction :
Several TWs have the same translation; EW has several translations to Thai.
Input : Selected bilingual dictionary and selected English WordNet.
Output : Candidate links
Logic Summary :

1. SELECT polysemic noun FROM Selected English WordNet
2. JOIN WordNet and bilingual dictionary ON English word
3. SELECT FROM 2 WHERE COUNT (English word)=1
4. SELECT FROM 2 WHERE COUNT(Thai word)>1
5. SELECT FROM 2 WHERE English word IN 3 AND Thai word IN 4

Module Name : 2.1.8 Polysemic M:N
Description : This module perform translation link construction :
Several TWs have different translations; EWs also have several translations.
Input : Selected bilingual dictionary and selected English WordNet.
Output : Candidate links
Logic Summary :

1. SELECT polysemic noun FROM Selected English WordNet
2. JOIN WordNet and bilingual dictionary ON English word
3. SELECT FROM 2 WHERE COUNT (English word)>1
4. SELECT FROM 2 WHERE COUNT(Thai word)>1
5. SELECT FROM 2 WHERE English word IN 3 AND Thai word IN 4

Module Name : 2.1.9 Variant
Description : This module perform translation link construction :
For a WordNet synset which contains a set of variants EWs1, if it is the case that two or more of the variants EWi have only one translation to the same Thai word TW, a link is produced for TW into the WordNet.7 synset.
Input : Selected bilingual dictionary and selected English WordNet.
Output : Candidate links
Logic Summary :

1. SELECT FROM bilingual dictionary WHERE COUNT (English word)=1
2. SELECT FROM WordNet WHERE English word IN 1
3. SELECT FROM 2 WHERE COUNT (Thai word)>1
4. IF EXIST SELCT 3 GROUP BY synset
5. THEN link Thai word to that synset

Copyright by Mahidol University

Module Name : 2.1.10 Intersection
Description : This module perform translation link construction :
All EWs share at least one common synset in WordNet Link TW is linked to all common synsets of its translations.
Input : Selected bilingual dictionary and selected English WordNet.
Output : Candidate links
Logic Summary :

1. SELECT FROM bilingual dictionary
2. SELECT FROM WordNet WHERE English word IN 1
3. SELECT FROM 1 WHERE COUNT (Thai word)>1
4. IF EXIST SELCT 2 GROUP BY synset
5. THEN link Thai word to that synset

Module Name : 2.1.11 Parent
Description : This module perform translation link construction :
A synset of an EW is direct parent of synsets corresponding to the rest of EWs Link. The TW is linked to all hyponym synsets.
Input : Selected bilingual dictionary and selected English WordNet.
Output : Candidate links
Logic Summary :

1. SELECT FROM bilingual dictionary
2. SELECT FROM WordNet WHERE English word IN 1
3. SELECT FROM 1 WHERE COUNT (Thai word)>1
4. SELECT synset parent of English word FROM 2
5. IF EXIST SELECT Thai word IN 3
6. THEN link Thai word to synset child

Module Name : 2.1.12 Brother
Description : This module perform translation link construction :
All EWs have synsets which are brothers respecting to a common parent Link. The TW is linked to all co-hyponym synsets.
Input : Selected bilingual dictionary and selected English WordNet.
Output : Candidate links
Logic Summary :

1. SELECT FROM bilingual dictionary
2. SELECT FROM WordNet WHERE English word IN 1
3. SELECT FROM 1 WHERE COUNT (Thai word)>1
4. SELECT synset parent of English word FROM 2
5. SELECT synset child of English word FROM 3
6. IF EXIST SELECT Thai word IN 4&3
7. THEN link Thai word to synset child that Thai word exist.

Copyright by Mahidol University

Module Name : 2.1.13 Distant hyponymy
Description : This module perform translation link construction :
A synset of an EW is a distant hypernym of synsets of the rest of the EWs Link. The Thai Word is linked to the lower level (hyponym) synsets.
Input : Selected bilingual dictionary and selected English WordNet.
Output : Candidate links
Logic Summary :

1. SELECT FROM bilingual dictionary
2. SELECT FROM WordNet WHERE English word IN 1
3. SELECT FROM 1 WHERE COUNT (Thai word)>1
4. SELECT synset parent of English word FROM 2
5. SELECT synset parent of English word FROM 3
6. SELECT synset child of English word FROM 4
7. SELECT synset child of English word FROM 5
8. IF EXIST SELECT Thai word IN 3&6
9. THEN link Thai word to synset child that Thai word exist.

4.4 File Structure

This part indicates the detail of each file (database table) in the research, which is divided into three parts as follows.

4.4.1 Source Table

The source table consists of two parts wordnet2sql and LEXiTRON. The description of each part are described following.

wordnet2sql

The following (Figure 4.9) are the relational model for wordnet2sql (there are some tables irrelative with nominal word, then we discard those tables). The grey rectangular represent the explanatory tables.

Copyright by Mahidol University

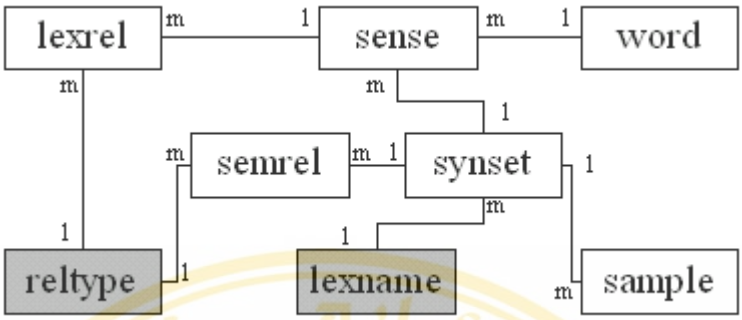


Figure 4.9 The relational model for wordnet2sql.

Table Name: WORD

Description: A 'word' is any sequence of characters that we associate a meaning with.

No.	Field Name	Description	Type	Size
1	WordNo	Primary key	Integer	-
2	Lemma	The string that represents the word	String	50

Table Name: SYNSET

Description: A synset can be roughly viewed as an entity of semantic information. A synset can have several words. All these words are then synonyms, they 'mean the same', but a word can also be present in more than one synset, because a word can have several meanings.

No.	Field Name	Description	Type	Size
1	SynSetNo	Primary key	Integer	-
2	Definition	The definition according to WordNet glossary	String	50
3	LexNo	Foreign key	Integer	-

Table Name: LEXREL

Description: 'lexrel' carries information about lexical relations. lexical relations are relations that exist between 'meaning's, that is between certain usages of certain words. This is saved in the direction from (word1,synset1) to (word2,synset2). So the association (word1,synset1) -> (word2,synset2) has the type 'reltype', but the

Copyright by Mahidol University

association (word2,synset2) -> (word1,synset2) does NOT NECCESSARILY also have the type 'reltype'.

No.	Field Name	Description	Type	Size
1	Word1	Word number	Integer	-
2	SynSet1	SynSet number	Integer	-
3	Word2	Word number	Integer	-
4	SynSet2	SynSet number	Integer	-
5	RelType	Relation Type	Integer	-

Table Name: SAMPLE

Description: WordNet's glossary contains a definition and/or one or more samples, illustrating the use of the synset. These are the samples from the WordNet -glossary.

No.	Field Name	Description	Type	Size
1	SynSetNo	SynSet number	Integer	-
2	SampleNo	Primary key	Integer	-
3	Sample	Sample	Text	-

Table Name: SEMREL

Description: 'semrel' carries information about semantic relations. Semantic relations hold between synsets. This is saved in the direction from synset1 to synset2. So association synset1 -> synset2 is of type 'reltype', but the association synset2 -> synset1 is NOT NECCESSARILY of the same type.

No.	Field Name	Description	Type	Size
1	SynSet1	SynSet number	Integer	-
2	SynSet2	Primary key	Integer	-
3	RelTypeNo	Relation Type	Integer	-

Table Name: SENSE

Description: A 'sense' in this context is the association between the syntactic entity 'word' and the semantic information it carries ('synset').

Copyright by Mahidol University

No.	Field Name	Description	Type	Size
1	WordNo	Primary key	Integer	-
2	SynSetNo	SynSet number	Integer	-
3	TagCnt		Integer	-

Table Name: LEXNAME

Description: Lexical name

No.	Field Name	Description	Type	Size
1	LexNo	Primary key	Integer	-
2	LexName	Word Type (POS/classification)	String	30
3	Description	Description	String	80

Table Name: RELTYPE

Description: Relation type

No.	Field Name	Description	Type	Size
1	RelTypeNo	Primary key	Integer	-
2	Type	POS (n=noun, v=verb, a=adjective,..)	Char	1
3	Description	Description	String	50

LEXiTRON

For LEXiTRON only 2 tables are exists: English to Thai dictionary (E-T) and Thai to English dictionary (T-E)

Table Name: E-T

Description: English to Thai dictionary

No.	Field Name	Description	Type	Size
1	Esearch	English Word	String	250
2	Eentry	English Word	String	250
3	Tentry	Thai Word	String	250
4	ecat	Category	String	10
5	ethai	Related Word	String	250
6	esyn	Synonym	String	250
7	eant	Antonym	String	250
8	id	Primary Key	String	10

Copyright by Mahidol University

Table Name: T-E

Description: Thai to English dictionary

No.	Field Name	Description	Type	Size
1	Tsearch	Thai Word	String	250
2	Tentry	Thai Word	String	250
3	Eentry	English Word	String	250
4	tcat	Category	String	10
5	tenglish	Related Word	String	250
6	tsyn	Synonym	String	250
7	tant	Antonym	String	250
8	tsample	Sample	String	250
9	tdef	Definition	String	250
10	tnum	Numerative Noun	String	250
11	Note	Note	String	250
12	id	Primary Key	String	10

4.4.2 Intermediate Table

The Intermediate table consists of three parts, which are MergeAll (E⇔T bilingual dictionary), MONO11, MONO1N, MONOM1, MONOMN, POLY11, POLY1N, POLYM1, POLYMN, VARIANT, INTERSECTION, PARENT, BROTHER, DISTANCE. (translation link from each Criteria) and Matrix. The description of each part are described following.

Intermediate Table

Table Name: MERGEALL

Description: English to Thai dictionary and Thai to English dictionary are merged together to remove duplicate data and for easily use

No.	Field Name	Description	Type	Size
1	Eentry	English Word	String	250
2	Tentry	Thai Word	String	250
3	Cat	Category	String	10
4	SeeAlso	Related Word	String	250
5	Syn	Synonym	String	250
6	Ant	Antonym	String	250
7	Sample	Sample	String	250
8	Def	Definition	String	250

Copyright by Mahidol University

No.	Field Name	Description	Type	Size
9	Num	Numerative Noun	String	250
10	Note	Note	String	250
11	id	Primary Key	String	10

Table Name: MONO11, MONO1N, MONOM1, MONOMN, POLY11, POLY1N, POLYM1, POLYMN, VARIANT, INTERSECTION, PARENT, BROTHER, DISTANCE.

Description: Translation link between Thai word, English word and English synset from each criteria.

No.	Field Name	Description	Type	Size
1	TWordNo	Thai Word number	Integer	-
2	EWordNo	English Word number	Integer	-
3	ESynSet	SynSet number	Integer	-

Table Name: MATRIX

Description: The result of merging all translation links in each criteria to summarize the data for using in logistic regression model.

No.	Field Name	Description	Type	Size
1	TWordNo	Thai Word number	Integer	-
2	EWordNo	English Word number	Integer	-
3	ESynSet	SynSet number	Integer	-
4	C01	The existence of link in criteria 1 (1=exist, 0=not exist)	Integer	-
5	C02	The existence of link in criteria 2	Integer	-
6	C03	The existence of link in criteria 3	Integer	-
7	C04	The existence of link in criteria 4	Integer	-
8	C05	The existence of link in criteria 5	Integer	-
9	C06	The existence of link in criteria 6	Integer	-
10	C07	The existence of link in criteria 7	Integer	-
11	C08	The existence of link in criteria 8	Integer	-
12	C09	The existence of link in criteria 9	Integer	-
13	C10	The existence of link in criteria 10	Integer	-
14	C11	The existence of link in criteria 11	Integer	-
15	C12	The existence of link in criteria 12	Integer	-
16	C13	The existence of link in criteria 13	Integer	-
17	OK	Result of manual evaluation (OK=1, KO=0)	Integer	-

Copyright by Mahidol University

4.4.3 Target Table

The description of Target table are described following.

Thai WordNet

The following (Figure 4.10) are the relational model for Thai WordNet. The left area are the tables of Thai WordNet, the right area are the tables of English WordNet and in the middle translation link between Thai and English WordNet

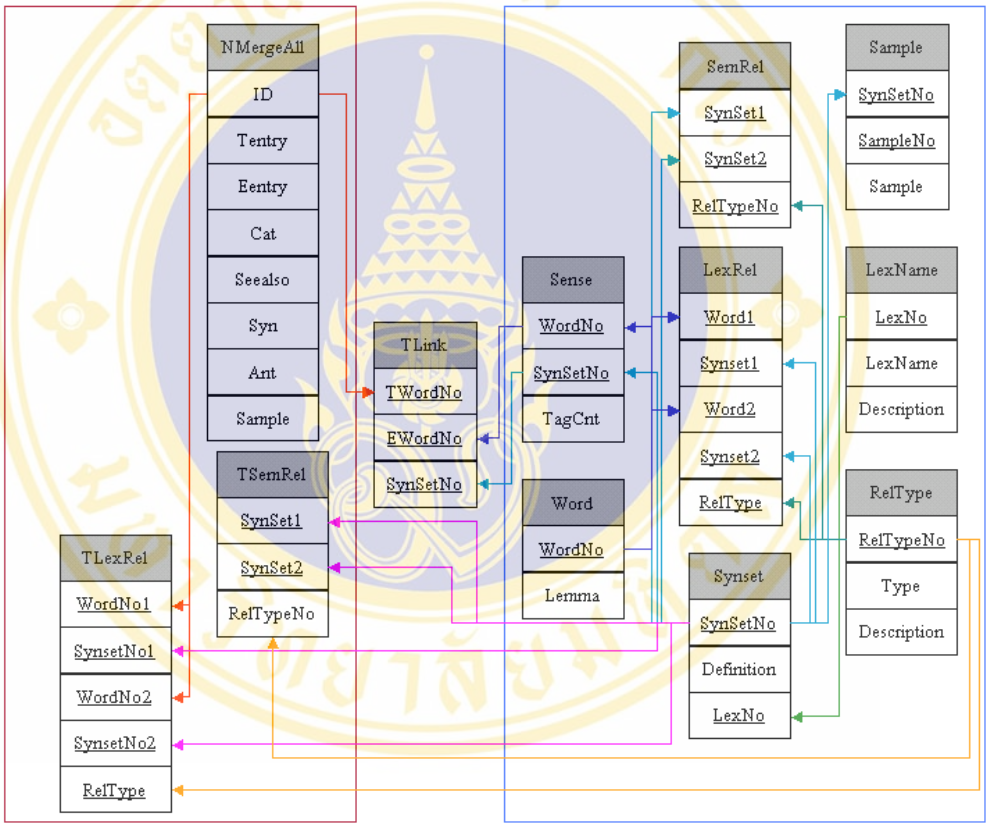


Figure 4.10 The relational model for Thai WordNet.

Table Name: TSemRel

Description: Thai semantic relation

NO	Field Name	Description	Type	Size
1	SynsetNo1	SynSet 1	Integer	
2	SynsetNo2	SynSet 2	Integer	
3	RelTypeNo	Relation Type	Integer	

Copyright by Mahidol University

Table Name: TLexRel

Description: Information about Thai lexical relations

NO	Field Name	Description	Type	Size
1	WordNo1	Word 1	Integer	
2	SynSetNo1	SynSet 1	Integer	
3	WordNo2	Word 2	Integer	
4	SynSetNo2	SynSet 2	Integer	
5	RelType	Relation Type	Integer	

Table Name: TLink

Description: Link between Thai word, English word, and English synset

NO	Field Name	Description	Type	Size
1	TWordNo	Thai Word	Integer	
2	EWordNo	English Word	Integer	
3	ESynSet	English Definition (meaning)	Integer	

4.5 Input-Output Design

In this section, screens created for Link Analyzer Module from the research are described as follows:

- A screen, Figure 4.11 Candidate Link Construction (13 criteria)enables the user to construct candidate link of translation link using 13 Criteria.
- A screen, Figure 4.12 Verify candidate link use for manually verify sample of candidate link.
- screen, Figure 4.13 Count Verified link use for find the existence of the candidate link in all criteria.
- A screen, Figure 4.14 Create matrix use for create the matrix of the existence sum by pattern and export to Microsoft Excel for use in Model Construction, which MiniTab are used.
- A screen, Figure 4.15 Input Coefficients use for input all coefficient of the evaluated model.

Copyright by Mahidol University

- A screen, Figure 4.16 Model Employment use for select the candidate of translation link.

- A screen, Figure 4.17 Thai WordNet (definitions and relations), Figure 4.18 Thai WordNet (hypernym relation), Figure 4.19 Thai WordNet (part meronym relation) use for query Thai word and relation of Thai word (Thai WordNet)

Figure 4.11 Candidate Link Construction (13 criteria)

Copyright by Mahidol University

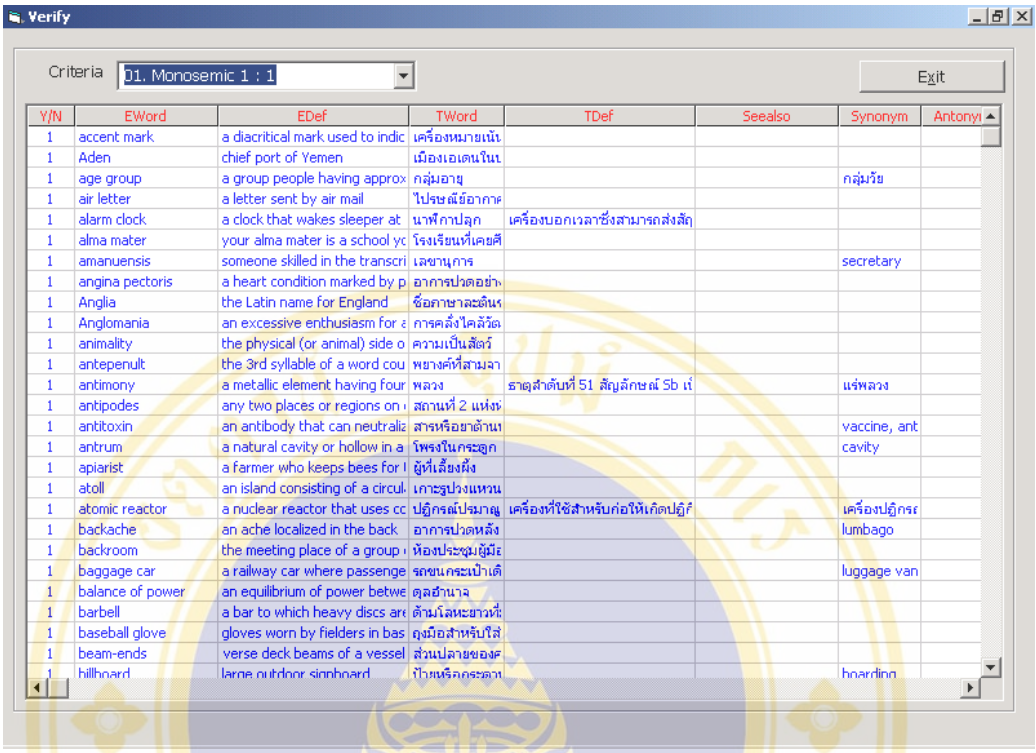


Figure 4.12 Verify candidate link

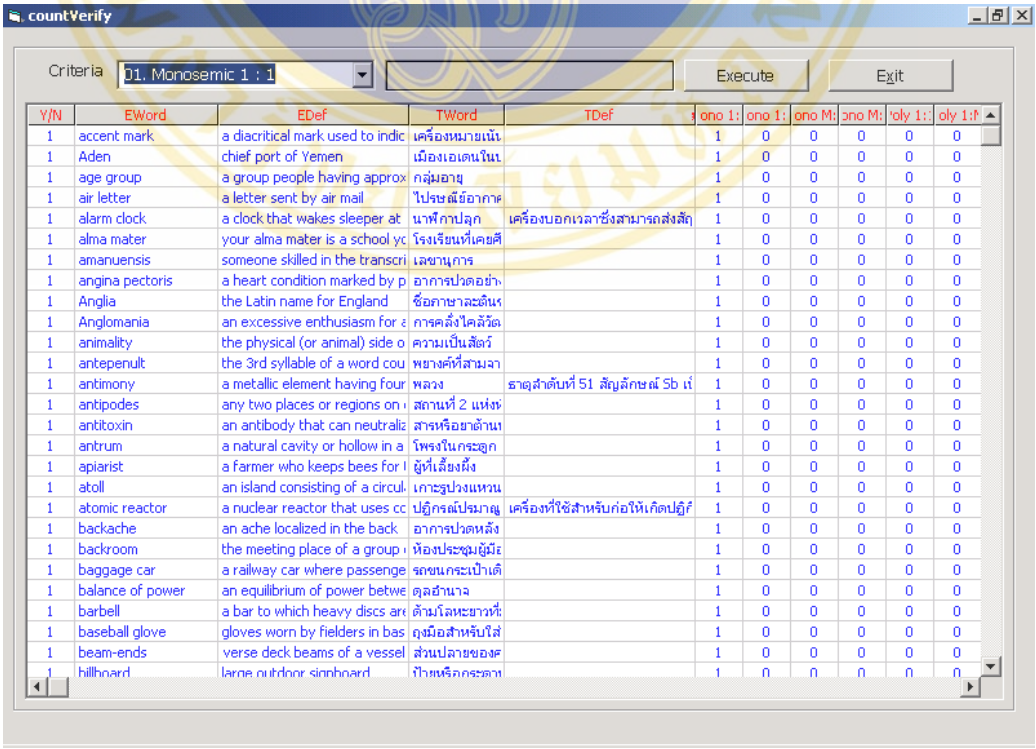
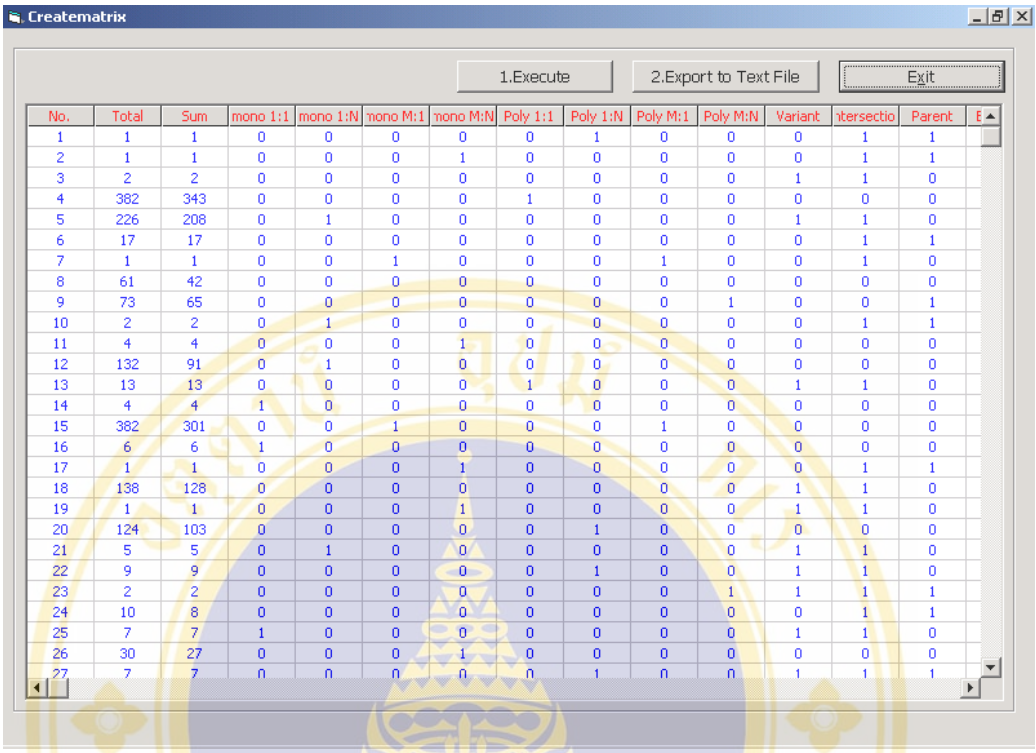


Figure 4.13 Count Verified link

Copyright by Mahidol University



No.	Total	Sum	mono 1:1	mono 1:N	mono M:1	mono M:N	Poly 1:1	Poly 1:N	Poly M:1	Poly M:N	Variant	ntersectio	Parent	E
1	1	1	0	0	0	0	0	1	0	0	0	1	1	
2	1	1	0	0	0	1	0	0	0	0	0	1	1	
3	2	2	0	0	0	0	0	0	0	0	1	1	0	
4	382	343	0	0	0	0	1	0	0	0	0	0	0	
5	226	208	0	1	0	0	0	0	0	0	1	1	0	
6	17	17	0	0	0	0	0	0	0	0	0	1	1	
7	1	1	0	0	1	0	0	0	1	0	0	1	0	
8	61	42	0	0	0	0	0	0	0	0	0	0	0	
9	73	65	0	0	0	0	0	0	0	1	0	0	1	
10	2	2	0	1	0	0	0	0	0	0	0	1	1	
11	4	4	0	0	0	1	0	0	0	0	0	0	0	
12	132	91	0	1	0	0	0	0	0	0	0	0	0	
13	13	13	0	0	0	0	1	0	0	0	1	1	0	
14	4	4	1	0	0	0	0	0	0	0	0	0	0	
15	382	301	0	0	1	0	0	0	1	0	0	0	0	
16	6	6	1	0	0	0	0	0	0	0	0	0	0	
17	1	1	0	0	0	1	0	0	0	0	0	1	1	
18	138	128	0	0	0	0	0	0	0	0	1	1	0	
19	1	1	0	0	0	1	0	0	0	0	1	1	0	
20	124	103	0	0	0	0	0	1	0	0	0	0	0	
21	5	5	0	1	0	0	0	0	0	0	1	1	0	
22	9	9	0	0	0	0	0	1	0	0	1	1	0	
23	2	2	0	0	0	0	0	0	0	1	1	1	1	
24	10	8	0	0	0	0	0	0	0	0	0	1	1	
25	7	7	1	0	0	0	0	0	0	0	1	1	0	
26	30	27	0	0	0	1	0	0	0	0	0	0	0	
27	7	7	0	0	0	0	0	1	0	0	1	1	1	

Figure 4.14 Create matrix

Figure 4.15 Input Coefficients

Copyright by Mahidol University

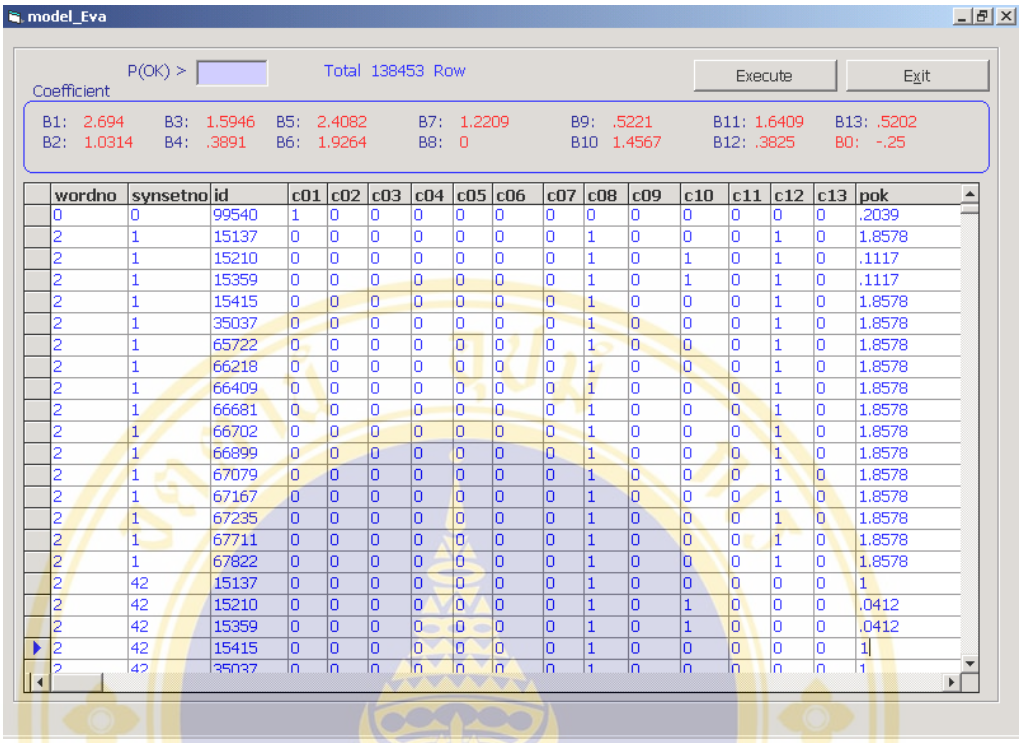


Figure 4.16 Model Employment

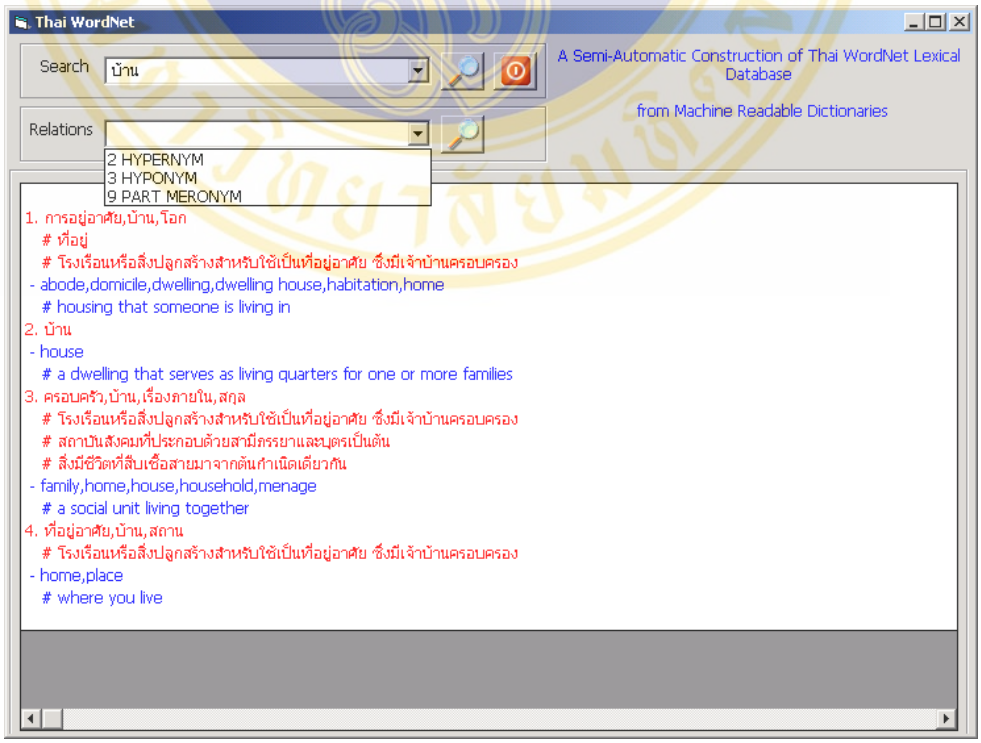


Figure 4.17 Thai WordNet (definitions and relations)

Copyright by Mahidol University

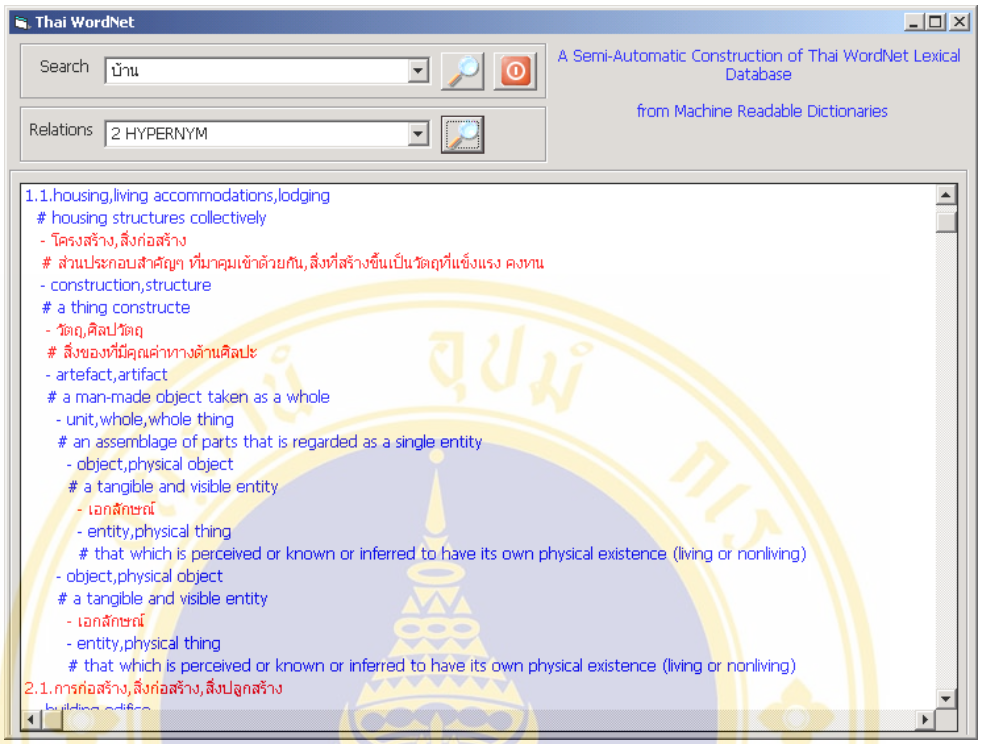


Figure 4.18 Thai WordNet (hypernym relation)

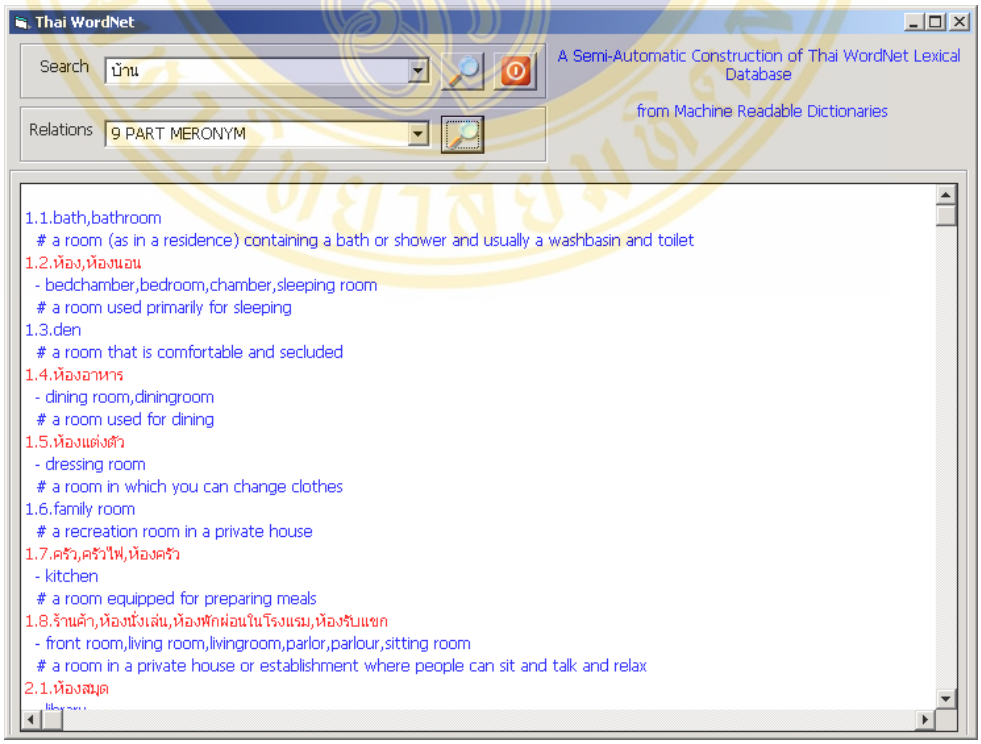


Figure 4.19 Thai WordNet (part meronym relation)

Copyright by Mahidol University

CHAPTER 5

SYSTEM IMPLEMENTATION

In this chapter, we describe the implementation details of WordNet Builder (WNB). Based on the framework discussed in Chapter 3, WNB consists of three main components: MRD Extractor, Link Analyzer and WordNet Constructor.

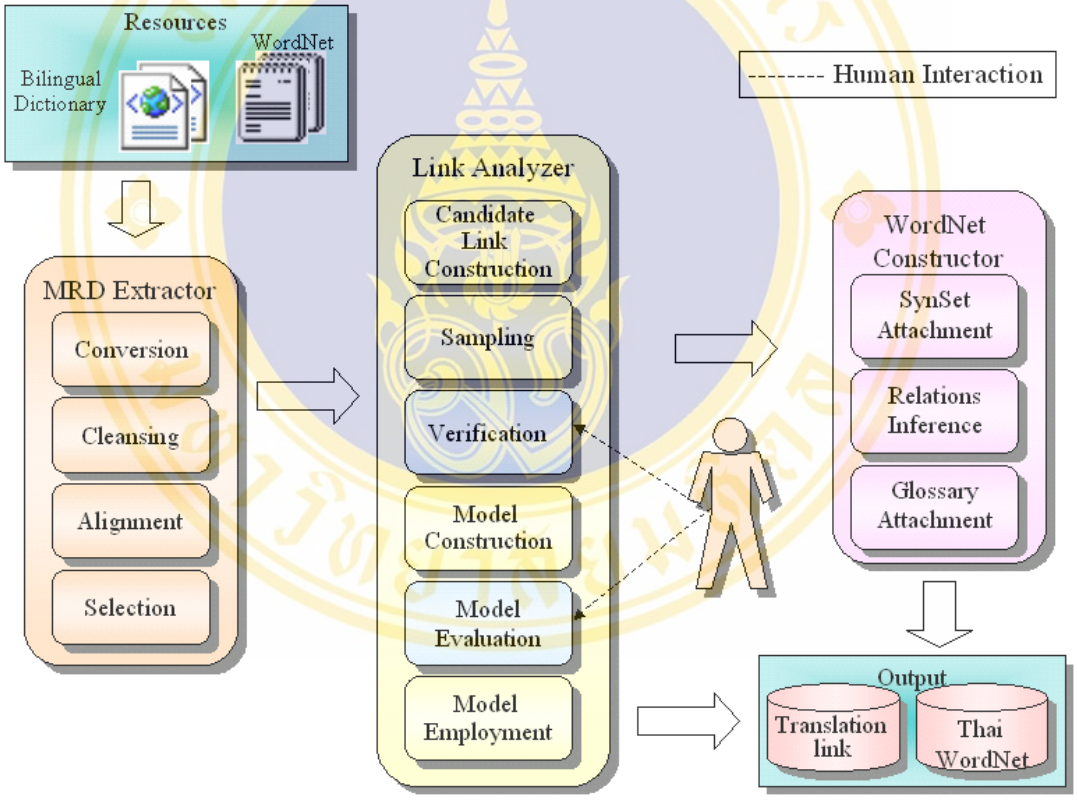


Figure 5.1 Overview of WNB

A diagram of the overview of WNB is shown in Figure 5.1. In WNB, we use bilingual dictionary (LEXiTRON) and monolingual dictionary (wordnet2sql) as resources. Both resources are available for free, but both are in different encoding. Due to the sources of resources are divergent, the resources will be arranged for readily use in next component by MRD Extractor. After the arrangement of resources,

Copyright by Mahidol University

the Link Analyzer will produce the intermediate result (Translation link, link between Thai word and English concept) and the WordNet construction will produce final result (Thai WordNet).

System Tools and Resources

We develop our system based on the following specifications.

Operating System	: Windows 2000 Professional
RAM	: 256 MB
Compiler	: Microsoft Visual Basic 6.0
Database	: Microsoft SQL Server 2000
CPU Speed	: Pentium IV 1.7 GHZ

We use Microsoft Access 2000 in Conversion, which converts XML to relational database. Query Analyzer a services of Microsoft SQL Server 2000 is use in Cleaning, Alignment and Selection process in MRD Extractor module. We use Microsoft Visual Basic 6.0 and Microsoft SQL Server 2000 to manage and store the data for Candidate Link Construction, Sampling, and Model Employment process. Microsoft Excel 2000 is used for splitting the task of verification of candidate links. Data Transform Service, a service of Microsoft SQL Server 2000 is used to transform data from Excel to relational database. And we use MiniTab, a statistical software program, to construct the logistic regression model in Model construction process.

5.1 MRD Extractor

The first module of WordNet Builder (WNB) is MRD Extractor. The purpose of this module is to prepare resources (both monolingual and bilingual dictionary) for next component (Link Analyzer). The main processing tasks are composed of four major components: conversion, cleansing, alignment, and selection.

This module performs the data preparation process. The description of the processing tasks of the system is stated below.

Copyright by Mahidol University

5.1.1 Conversion

The existing resources, LEXiTRON and wordnet2sql, are encoding in divergent format. LEXiTRON is the Corpus-Based Thai⇔English dictionary (bilingual dictionary), which is in XML format. LEXiTRON is the open sources license. And wordnet2sql is English WordNet (monolingual dictionary), which is in the form of SQL script. The wordnet2sql is the GNU General Public License. The Conversion will perform the transformation from XML and SQL script into RDBMS for easily handle.

Monolingual Dictionary Conversion

For the monolingual dictionary conversion, we use wordnet2sql which is the English WordNet in the form of SQL-script for MySQL. The wordnet2sql is the SQL-script converted from WordNet data file.

For this module, we have to modify SQL-script for using with MS SQL Server 2000. Afterward, SQL-script is loaded into MS SQL Server 2000. The following Figure 5.2 and Figure 5.3 are the example of wordnet2sql and converted wordnet2sql (table name:synset), respectively.

Word
(0, 'entity') (1, 'physical thing') (2, 'thing') (3, 'anything') (4, 'something') (5, 'nothing') (6, 'nonentity') (7, 'whole') (8, 'whole thing') (9, 'unit') (10, 'living thing')
Synset
(0, 'that which is perceived or known or inferred to have its own physical existence (living or nonliving)', 3) (1, 'a separate and self-contained entity', 3) (2, 'a thing of any kind', 3) (3, 'a thing of some kind', 3) (4, 'a nonexistent thing', 3)
Sense
(0, 0, 11) (1, 0, 0) (2, 1, 0) (3, 2, 0) (4, 3, 0) (5, 4, 0) (6, 4, 0) (7, 5, 0) (8, 5, 0) (9, 5, 0) (10, 6, 1) (11, 6, 0) (12, 7, 9) (13, 7, 7) (14, 8, 31) (15, 9, 0) (16, 10, 44) (17, 11, 0) (18, 11, 4) (19, 11, 0) (20, 12, 7229)

Figure 5.2 The example data of wordnet2sql

Copyright by Mahidol University

	synsetno	definition	lexno
	0	that which is perceived or known or inferred to have its own physical existence (living or nonliving)	3
	1	a separate and self-contained entity	3
	2	a thing of any kind	3
	3	a thing of some kind	3
	4	a nonexistent thing	3
	5	an assemblage of parts that is regarded as a single entity	3
	6	a living (or once living) entity	3
	7	a living thing that has (or can develop) the ability to act or function independently	3
	8	living things collectively	3
	9	a discrete unit of living matter	3
	10	the basic structural and functional unit of all organisms	3
	11	any entity that causes events to happen	3
	12	a human being	3
	13	a living organism characterized by voluntary movement	3
	14	a living organism lacking the power of locomotion	3
	15	a tangible and visible entity	3
	16	an object occurring naturally	3
	17	that which has mass and occupies space	3
	18	any substance that can be metabolized by an organism to give energy and build tissue	3
	19	a man-made object taken as a whole	3
	20	one of a class of artifacts	3
	21	a feature of the mental life of a living organism	3
	22	a general concept formed by extracting common features from specific examples	3
	23	the psychological result of perception and learning and reasoning	3
	24	the psychological feature that arouses an organism to action toward a desired goal	3
	25	the psychological feature of experiencing affective and emotional states	3
	26	a point or extent in space	3
	27	the spatial arrangement of something as distinct from its substance	3
▶	28	the continuum of experience in which events pass from the future through the present to the past	3

Figure 5.3 The example of converted wordnet2sql (table name: synset)

Bilingual Dictionary Conversion

For the bilingual dictionary conversion, we use LEXiTRON which is the Thai-English and English-Thai dictionaries. LEXiTRON is in the form of XML. For transforming XML to RDBMS, Microsoft Access 2000 is utilized. Transforming XML to RDBMS using Microsoft Access 2000, we have to define the XML Schema Definition (XSD). And due to the enormous number of data and limitation of MS Access 2000, we have to split the XML file into small pieces before transform to MS Access 2000. Then transform MS Access 2000 to MS SQL Server 2000 using Data Transform Services (DTS), a component of MS SQL Server 2000. The following Figure 5.4 and Figure 5.5 are the example of LEXiTRON in XML file format and the example of converted LEXiTRON (RDBMS).

Copyright by Mahidol University

<Doc>
<tsearch>ความพึงพอใจ</tsearch>
<tentry>ความพึงพอใจ</tentry>
<eentry>contentment</eentry>
<tcat>N</tcat>
<tenglish>satisfaction</tenglish>
<tsyn>ความพอใจ, ความชอบใจ, ความสมใจ, ความสมหวัง, ความสมปรารถนา</tsyn>
<tant>ความเกลียด</tant>
<tsample>ความสำเร็จในชีวิตของคนปัจจุบันมิได้วัดด้วยความสุข หรือความพึงพอใจจากภาวะภายใน หากเอาปัจจัยภายนอก หรือค่านิยมที่คนทั่วไปยึดถือเป็นเกณฑ์กำหนด</tsample>
<tdef>ความรู้สึกที่สมใจ, ชอบใจ</tdef>
<id>3555</id>
</Doc>

Figure 5.4 The example data of LEXiTRON in XML file format

	esearch	eentry	tentry	ecat	ethai	esyn	eant	id
	assay	assay	สารที่ใช้วิเคราะห์	N				3007
	assemblage	assemblage	การรวบรวม	N		callection; gathering		3008
	assemblage	assemblage	ชุมนุมชน	N	ฝูงชน	assembly		3009
	assembly	assembly	กลุ่มคน	N	คนที่รวมตัวกัน			3013
	assembly	assembly	การประกอบ	N	การทำ, การสร้าง	construction; piecing toge		3014
	assembly	assembly	การรวมกลุ่ม	N	การชุมนุม, การป	assemblage; meeting; ass		3015
	assembly	assembly	ชิ้นส่วน	N	ส่วนประกอบ			3016
	assembly	assembly	สัญญาณรวมพล	N				3017
	assent	assent	การตกลง	N	การยินยอม, การ	concent; agreement		3018
	assertion	assertion	การยืนยัน	N		affirmation; statement		3022
	assessment	assessment	การประเมินค่า	N				3027
	asset	asset	สิ่งที่มีคุณค่า	N				3028
	asset	asset	สินทรัพย์	N	สินทรัพย์, ของมีค			3029
	asshole	asshole	คนโง่ (คำสแลง)	N				3030
	asshole	asshole	ทวารหนัก	N		anus		3031
	assiduity	assiduity	ความเพียร	N	ความขยัน, ความ	diligence		3032
	assignee	assignee	ผู้รับโอน	N	ผู้รับทรัพย์สิน (จ			3040
	assigner	assigner	ผู้โอน	N		assignor		3041
	assignment	assignment	การมอบหมาย	N		designation; authorization		3042
	assignment	assignment	งาน	N	งานที่ได้รับมอบห	task; responsibility; job		3043
	assignor	assignor	ผู้โอน	N		assigner		3044
	assist	assist	การช่วย	N	การช่วยเหลือ			3048
	assistance	assistance	ความช่วยเหลือ	N	การสนับสนุน	aid		3051
	assistant	assistant	ผู้ช่วย	N	ผู้ช่วยเหลือ	helper		3053
	associate	associate	ผู้ที่มีความสัมพันธ์กัน (ท	N				3057

Figure 5.5 The example of converted LEXiTRON (table name: E-T)

After converted and loaded both monolingual and bilingual dictionary into RDBMS, cleansing, aligning, and selecting will be perform for ease of use in Link Analyzer module.

Copyright by Mahidol University

5.1.2 Cleansing

In both monolingual and bilingual dictionary, there are some erroneous data then the manually cleansing is required.

Monolingual Dictionary Cleansing

The wordnet2sql is the SQL script for crate WordNet database, which converted from WordNet file format and there need to be cleaned.

- ***Filling missing data***

In wordnet2sql, there are some missing data need to be filled up. In table reltype (relation type), a record of data is missing. The following tables (Table 5.1) are the missing data and the filled data, respectively.

Table 5.1 The example of filling missing data in monolingual dictionary

word1	synset1	word2	synset2	reltype
2423	108981	2423	92602	21
8364	107964	8364	94746	21
23117	105544	134226	105543	21
26979	110814	26979	98042	21
28262	109940	28262	97369	21

a) Table lexrel has reltype which refers to table reltype.

Realtypeno	Type	Description
18	a	ANTONYM
19	a	SIMILAR TO
20	a	PARTICIPLE OF VERB
22	a	PERTAINYM
23	a	ATTRIBUTE

b) Table reltype which miss some data (reltypeno=21)

Realtypeno	Type	Description
18	A	ANTONYM
19	a	SIMILAR TO
20	a	PARTICIPLE OF VERB
21	a	ATTRIBUTE
22	a	PERTAINYM
23	a	ATTRIBUTE

Copyright by Mahidol University

c) Table reltype with filled data

Bilingual Dictionary Cleansing

Due to the LEXiTRON is a corpus-based dictionary, which is the semi-automatic construction and more missing and erroneous data. The following are the description of cleansing the data in bilingual dictionary.

- ***Filling missing data***

There are some missing data in LEXiTRON (category = 6 records,). Then we need to find and fill the missing information from other dictionaries (e.g. Babylon, So Sethaputra, and etc.) And there are some Thai word or English word or category that cannot be fill the data are removed (4 records). The following tables (Table 5.2) are the example of filling missing data in bilingual dictionary.

Table 5.2 The example of filling missing data in bilingual dictionary

Esearch	eentry	Tentry	ecat	ethai	esyn	eant	Id
arouse from	arouse from	กระตุ้นให้ฟื้นจาก	NULL	NULL	awaken from	NULL	3833
dirty work	dirty work	การทรยศ	NULL	การหักหลัง	NULL	NULL	15783
In my book	in my book	ในความคิดของ	NULL	ตามความเห็นของ	NULL	NULL	30953
In poor taste	in poor taste	หยาบโลน	NULL	NULL	NULL	NULL	30981
medullary	medullary	เกี่ยวกับสมองส่วนใน (medulla oblongata)	NULL	NULL	NULL	NULL	41700
rinky-dink	rinky-dink	ราคาถูก	NULL	ถูก, ราคาต่ำ	NULL	NULL	60974
Sponds	Sponds	เงิน	NULL	NULL	spondulics	NULL	62251

a) missing category.

esearch	eentry	Tentry	ecat	ethai	esyn	eant	Id
arouse from	arouse from	กระตุ้นให้ฟื้นจาก	PHRV	NULL	awaken from	NULL	3833
dirty work	dirty work	การทรยศ	IDM	การหักหลัง	NULL	NULL	15783
In my book	in my book	ในความคิดของ	IDM	ตามความเห็นของ	NULL	NULL	30953
In poor taste	in poor taste	หยาบโลน	ADJ	NULL	NULL	NULL	30981
medullary	medullary	เกี่ยวกับสมองส่วนใน (medulla oblongata)	ADJ	NULL	NULL	NULL	41700
rinky-dink	rinky-dink	ราคาถูก	ADJ	ถูก, ราคาต่ำ	NULL	NULL	60974
sponds	sponds	เงิน	SL	NULL	spondulics	NULL	62251

b) Filled missing category.

Copyright by Mahidol University

tsearch	tentry	eentry	cat	tenglish	tsyn	tant	id
NULL	จนถึงสมัยอยุธยาตอนปลายในรัชกาลพระเจ้าเอกทัศน์	NULL	NULL	NULL	NULL	NULL	40015
NULL	ขอบเขต	NULL	ประธานคณะกรรมการมีกำหนดจะประกาศรายชื่อนักการเมืองที่อยู่ในข่ายต้องสงสัยเรื่องการทุจริต	NULL	NULL	NULL	40021
เป็นกษัตริย์ต่อเนื่องแทน	เพราะไร้ราชโอรสสืบราชสมบัติ	NULL	NULL	NULL	NULL	NULL	40681
NULL	เป็นครั้งแรกหลังจากกระทรวงกลาโหมสหรัฐนำเทปโทรทัศน์มาเผยแพร่	NULL	NULL	NULL	NULL	NULL	40734

c) missing English word or Thai word or category

- ***Removing wrong spelling***

In this process, we consider the visible wrong spelling which occurs in all fields. All visible wrong spelling are removed (103 records). The example of the visible wrong spelling and the result of removing are depicted in

Table 5.3. Table a) is wrong spelling English word or Thai word or category and b) is the removed Wrong spelling in English word

Table 5.3 The example of removing wrong spelling

tsearch	Tentry	Eentry	cat	tenglish	tsyn	tant	tsample	tdef	tnum	note	Id
ถกล	ถกล 2	?be beautiful	V	NULL	งาม, ดำกล	NULL	NULL	NULL	NULL	(เขมร)	23158
ถ่ายทุกข์	ถ่ายทุกข์	?urinate	V	pass urine; void urine	ถ่าย ปัสสาวะ, ฉี่ , เยี่ยว, ปัสสาวะ	NULL	คนขับรถมักแวะถ่ายทุกข์ข้างทาง	ขับถ่ายของเหลวที่เสียออกทางทวารเบา	NULL	NULL	23360
ดีแตะ	ดีแตะ	?น่าจะใช้คำใน Teng ได้	V	to make a wall or fence by wearing strips of split bamboo	ขัดแตะ	NULL	NULL	ขัดแตะโดยเอาไม้ไผ่ผ่าให้เป็นซีกๆ แล้วขัดและตีให้แน่นในการทำรั้วบ้านและฝาเรือน	NULL	NULL	22688

a) Wrong spelling English word or Thai word or category

Copyright by Mahidol University

tsearch	Tentry	Eentry	cat	tenglish	tsyn	tant	Tsample	tdef	tnum	note	id
ถกล	ถกล 2	be beautiful	V	NULL	งาม, ดำกล	NULL	NULL	NULL	NULL	(เขมร)	23158
ถ่ายทุกข์	ถ่ายทุกข์	urinate	V	pass urine; void urine	ถ่าย ปัสสาวะ, ฉี่, เยี่ยว, ปัสสาวะ	NULL	คนขับรถมักแวะถ่ายทุกข์ข้างทาง	ขับถ่ายของเหลวที่เสียออกทางทวารเบา	NULL	NULL	23360
ตีแตะ	ตีแตะ	Teng	V	to make a wall or fence by wearing strips of split bamboo	ขัดแตะ	NULL	NULL	ขัดแตะโดยเอาไม้ไผ่ผ่าให้เป็นซีกๆ แล้วขัดและตีให้แน่นในการทำรั้วบ้านและฝาเรือน	NULL	NULL	22688

b) Removed Wrong spelling in English word

- ***Manipulating words with parenthesis***

This process manipulate both Thai and English words with parenthesis. Due to the LEXiTRON is corpus-based dictionary, there are some word including parenthesis which cannot match with WordNet for creating candidate links between Thai word and English concepts. Then to gain the candidate links, the data with in parenthesis are split into the new filed. The following (Table 5.4) are the example of data with parenthesis and the split data, respectively. There are 17,442 records, which are split.

Table 5.4 The example of manipulating words with parenthesis

tsearch	tentry	eentry	cat	tenglish	tsyn	tant	tsample	Tdef	Id
ราบ	ราบ	(lie) flat	ADV	NULL	เสมอ, เรียบ	NULL	พอได้ยินเสียงปืนเขาจึงพุ่งตัว นอนราบลงกับพื้นทันที	เรียบเสมอพื้นไม่มีลุ่มๆ ดอนๆ	5867
เริ่มแรก	เริ่มแรก	(at) first	ADV	firstly; initially; at the beginning; at the outset; in the first place	เริ่มต้น	NULL	การรับรู้ของเด็กๆ นั้นเริ่มแรกจะมองภาพต่างๆ เป็น 2 มิติ	NULL	6873
ตับ	ตับ 2	(in a) row	N	(in a) line; (in a) rank	แถว	NULL	รถเก๋งงามหรูจอดเป็นตับอยู่ในโรงรถ	เรียกของที่ผูก ตั้ง หรือวางเรียงกันเป็นแถว	21306
ตอนหลัง	ตอนหลัง	(the) rear	N	back part	ข้างหลัง, ด้านหลัง, ตอนปลาย, ท้าย, หาง, ปลาย	NULL	เขาชอบนั่งตอนหลังของรถ	NULL	22526

a) The data with parenthesis

Copyright by Mahidol University

tsearch	trelated	tentry	erelated	eentry	cat	tenglish	tsyn	tant	tsample	tdef	id
ราบ	NULL	ราบ	lie	flat	ADV	NULL	เสมอ, เรียบ	NULL	พอได้ยินเสียงปืน เขาจึงพุ่งตัว นอนราบลงกับพื้นทันที	เรียบเสมอพื้น ไม่มีลุ่มๆ ดอนๆ	5867
เริ่มแรก	NULL	เริ่มแรก	at	first	ADV	firstly; initially; at the beginning; at the outset; in the first place	เริ่มต้น	NULL	การรับรู้ของเด็กๆ นั้นเริ่มแรกจะมองภาพต่างๆ เป็น 2 มิติ	NULL	6873
ตับ	NULL	ตับ 2	in a	row	N	(in a) line; (in a) rank	แถว	NULL	รถเก๋งงามหรูจอดเป็นตับอยู่ในโรงรถ	เรียกของที่ผูกตั้ง หรือวางเรียงกันเป็นแถว	21306
ตอนหลัง	NULL	ตอนหลัง	the	rear	N	back part	ข้างหลัง, ด้านหลัง, ตอน ปลาย, ท้าย, หาง, ปลาย	NULL	เขาชอบนั่งตอนหลังของรถ	NULL	22526

b) The split data

- ***Removing duplicate record***

There are some duplicate records in LEXiTRON, then removal of the duplicate records from the resources must be performed. Tremendous duplication of data are removed manually. The following (Table 5.5) are the example of duplicate data and the remaining data. Some records differ in only white space and some record are subset of others, the manually removal are performed.

Copyright by Mahidol University

Table 5.5 The example of removing duplicate record

tsearch	trelated	tentry	erelated	eentry	cat	tenglish	tsyn	tant	tsample	tdef	tnum	note	id
เป้า	NULL	เป้า 1	NULL	Aim	N	purpose; objective; target; goal; quota	จุดหมาย, จุดประสงค์, วัตถุประสงค์, เป้าหมาย, จุดมุ่งหมาย, ที่หมาย, เป้าประสงค์	NULL	ปีนี้บริษัทตั้งเป้าไว้ว่าจะทำยอดขายให้ได้ร้อยล้านบาท	สิ่งที่กำหนดไว้เป็นจุดหมาย	NULL	NULL	15901
เป้า	NULL	เป้า 1	NULL	Aim	N	purpose; objective; target; goal	จุดหมาย, จุดประสงค์, วัตถุประสงค์, เป้าหมาย, จุดมุ่งหมาย, ที่หมาย	NULL	ปีนี้บริษัทตั้งเป้าไว้ว่าจะทำยอดขายให้ได้ร้อยล้านบาท	สิ่งที่กำหนดไว้เป็นจุดหมาย	NULL	NULL	16499
เปลือกไม้	NULL	เปลือกไม้	NULL	Bark	N	husk; rind	เปลือกต้นไม้, เปลือก, กาบ	NULL	เปลือกไม้บางชนิดมียางจึงเป็นเชื้อเพลิงได้ดี	สิ่งที่หุ้มส่วนนอกของต้นไม้	ชิ้น,อัน	NULL	16473
เปลือกไม้	NULL	เปลือกไม้	NULL	bark	N	NULL	เปลือก, กาบ	NULL	เปลือกไม้บางชนิดมียางจึงเป็นเชื้อเพลิงได้ดี	สิ่งที่หุ้มส่วนนอกของต้นไม้	ชิ้น,อัน	NULL	15881

a) duplicate data.

tsearch	trelated	tentry	erelated	eentry	cat	Tenglish	tsyn	tant	tsample	tdef	tnum	note	id
เป้า	NULL	เป้า 1	NULL	Aim	N	purpose; objective; target; goal; quota	จุดหมาย, จุดประสงค์, วัตถุประสงค์, เป้าหมาย, จุดมุ่งหมาย, ที่หมาย, เป้าประสงค์	NULL	ปีนี้บริษัทตั้งเป้าไว้ว่าจะทำยอดขายให้ได้ร้อยล้านบาท	สิ่งที่กำหนดไว้เป็นจุดหมาย	NULL	NULL	15901
เปลือกไม้	NULL	เปลือกไม้	NULL	Bark	N	husk; rind	เปลือกต้นไม้, เปลือก, กาบ	NULL	เปลือกไม้บางชนิดมียางจึงเป็นเชื้อเพลิงได้ดี	สิ่งที่หุ้มส่วนนอกของต้นไม้	ชิ้น,อัน	NULL	16473

b) removal of the subset of others.

5.1.3 Alignment

Once the cleansing is performed, the resources have to be aligned. Because of the LEXiTRON consists of 2 tables (English-Thai and Thai-English dictionaries), then the alignment of both dictionaries are performed to maximize the number of links and to facilitate links analyzer module. Merging English-Thai and

Copyright by Mahidol University

Thai-English dictionaries, we choose all data in Thai-English dictionary and add English-Thai dictionary, which is not exist in Thai-English dictionary to gain the number of links. The illustration of Alignment module process is shown in Figure 5.6. There are 83,030 records for English-Thai dictionary and 39,752 for Thai-English dictionary and after merge both dictionaries together, the number of data is 119,483 records.

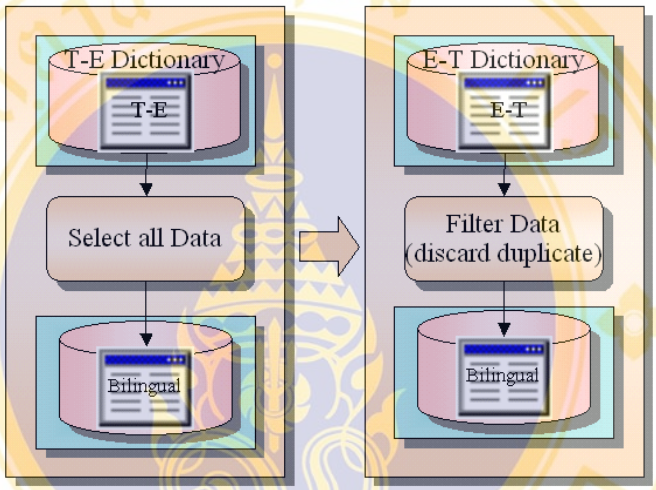


Figure 5.6 Alignment module

5.1.4 Selection

This is the last process that is used to prepare the data before creating the candidate links. Since the methodology to crate candidate link are evidenced with only the nominal words and the concept of nominal word in every language is similarity compared to other categories (verb, adjective, adverb, and etc.), the nominal words are selected for the link analyser module. Both monolingual and bilingual dictionaries are selected only noun category. There are 119,231 nominal words for monolingual dictionary and 51,646 nominal words for bilingual dictionary. As the result of emphasizing on the nominal words, only the related information is selected.

Copyright by Mahidol University

5.1.5 Result of MRD Extractor

As aforementioned, the purpose of this module is to prepare resources (both monolingual and bilingual dictionary) for next component (Link Analyzer). The main processing tasks are composed of 4 major component: conversion, cleansing, alignment, and selection. The resources for this process are LEXiTRON – bilingual dictionary and Wordnet2sql – monolingual dictionary. The number of data result from each process are shown in Table 5.6 And these data are used in Link Analyzer to create the candidate of translation link.

Table 5.6 Number of data from each process in MRD Extractor

	Conversion	Cleansing	Alignment	Selection
E-T	83,232	79,711	-	-
T-E	40,850	39,772	-	-
Mergeall	-	-	119,231	51,646
Wordnet2sql(table sense)	195,817	195,817	195,817	134,716

5.2 Link Analyzer

All data prepared by the MRD Extractor are used in Link Analyzer to generate the candidate link of Thai Word with English concepts by automatically constructing the link by using the criteria that lexicographer use in create the dictionary and sampling the candidate links for manually verify, then using Logistic regression (statistical method) to construct and choose the candidate links to generate Thai WordNet.

This module performs the Thai word and English concept link generation. The description of the processing tasks of the module is stated below.

5.2.1 Candidate Link Construction

To construct the candidate link of Thai Word and English concepts, 13 criteria are utilized for disambiguate the link.

Copyright by Mahidol University

We group all criteria into two criteria groups: criteria 1-8 are in the Class criteria group and criteria 9-13 are in Structural criteria group. The description of the criteria is stated below.

Class criteria group

For Class criteria group, all candidate links are analyzed and classified by considering on the English word and Thai word. This class criteria group can be divided into two subgroups: Monosemic criteria and Polysemic criteria. In Class criteria group, a translation link should exist in only one criteria. The illustration of monosemic criteria and polysemic criteria are shown in Figure 5.7 and Figure 5.8, respectively.

Monosemic criteria will consider only the English word, which has only one concept (meaning) and Polysemic criteria will consider English word, which has more than one concepts. Both Monosemic and Polysemic criteria consist of 4 criteria:

1:1 – A Thai Word (TW) has only one English translation (EW); symmetrically, EW has TW as its unique translation.

1:N – A TW has more than one translation; each EW has TW as its unique translation.

M:1 – Several TWs have the same translation; EW has several translations to Thai.

M:N – Several TWs have different translations; EWs also have several translations.

Copyright by Mahidol University

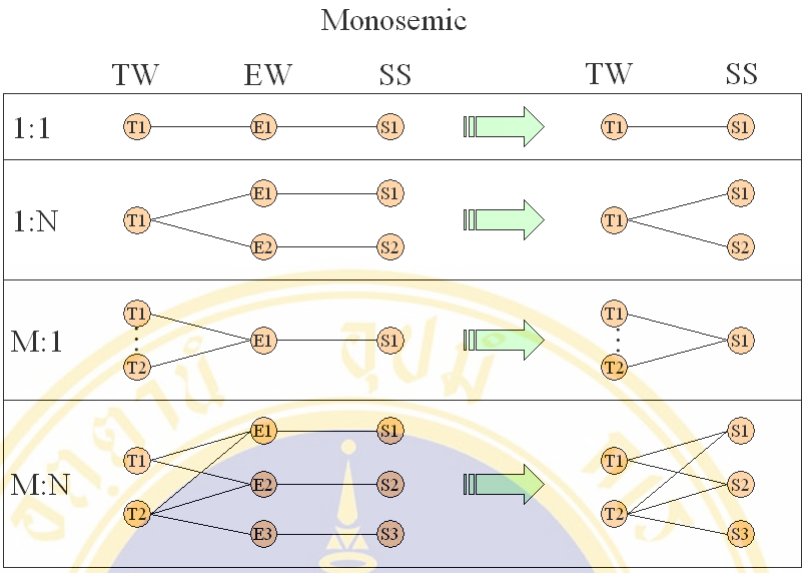


Figure 5.7 Monosemic Criteria

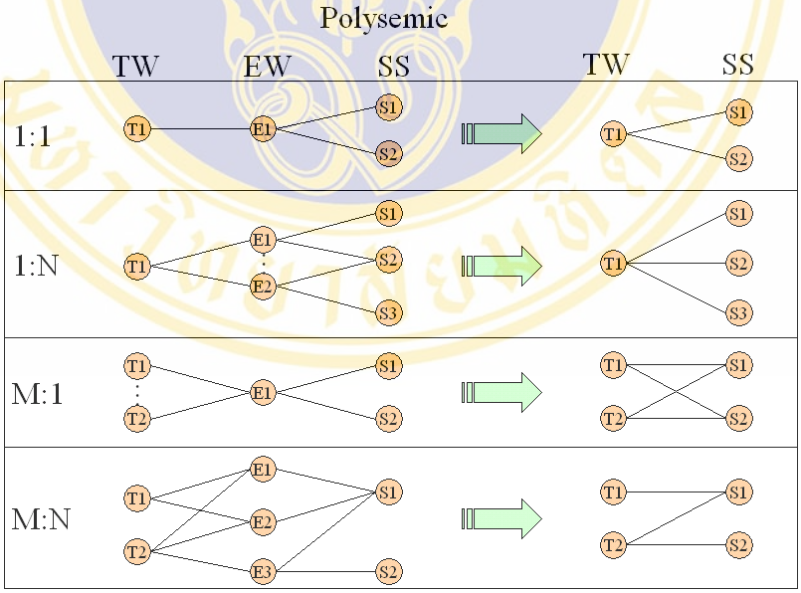


Figure 5.8 Polysemic Criteria

Copyright by Mahidol University

Structural criteria group

For Structural criteria group, the translation link will be constructed from implicit link, which exist in the English WordNet structure. Structural criteria group consists of 5 criteria. The illustration of Structural criteria is shown in Figure 5.9

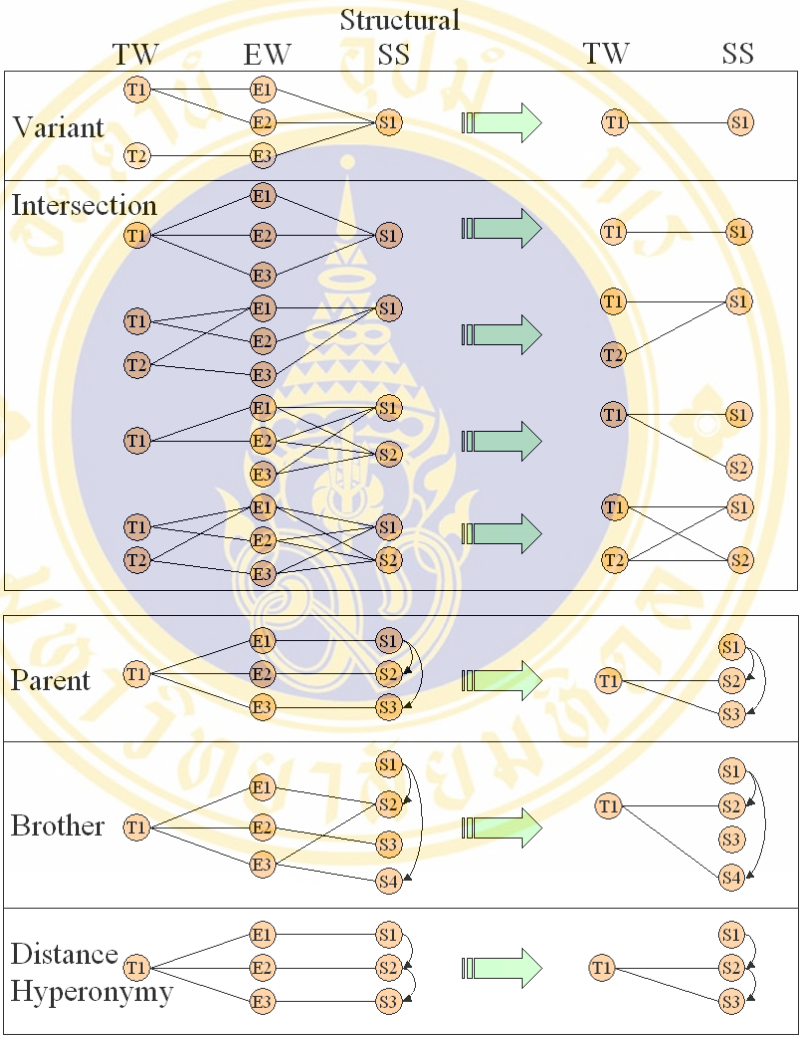


Figure 5.9 Structural Criteria

Variant criteria – For a WordNet synset which contains a set of variants English Words (EWs), if it is the case that two or more of the variants EWi have only one translation to the same Thai word Thai word (TW), a link is produced for TW into the WordNet synset.

Copyright by Mahidol University

Intersection criteria – All EWs share at least one common synset in WordNet Link TW is linked to all common synsets of its translations

Parent criteria – A synset of an EW is direct parent of synsets corresponding to the rest of EWs Link. The TW is linked to all hyponym synsets

Brother criteria – All EWs have synsets which are brothers respecting to a common parent Link. The TW is linked to all co-hyponym synsets.

Distant hyperonymy criteria – A synset of an EW is a distant hyperonymy of synsets of the rest of the EWs Link. The Thai Word is linked to the lower level (hyponym) synsets.

The result of candidate link construction are sets of translation link of each criteria. The following are the examples of candidate links created from monosemic 1:1 criteria (Table 5.7).

Table 5.7 The examples of candidate links created from monosemic 1:1.

wordno	synsetno	Id	Eword	E Definition	Tword	def	seealso	Syn	ant	sample
36388	23823	58786	walkway	a path set aside for walking	ทางเดินเท้า	NULL	ทางเท้า	footpath, pavement, sidewalk	NULL	NULL
67245	48998	3337	cooperator	an associate who works with others toward a common goal	ผู้ร่วมมือ	ผู้ที่พร้อมใจช่วยกันทำสิ่งใดสิ่งหนึ่ง	cooperant	NULL	NULL	NULL
97497	65310	79607	sales	income (at invoice values) received for goods and services over some given period of time	ยอดการขาย	NULL	รายได้จากการขาย	receipts	NULL	NULL
101521	69118	76480	self-sufficiency	personal independence	ความพึ่งพาตนเองได้	NULL	NULL	NULL	NULL	NULL

Copyright by Mahidol University

The result of this module is shown in Table 5.8. From the result The minimum number of links constructed by criteria 6 Polysemic 1:N and the maximum number of links constructed by criteria 8 Poly M:N. Next, the result have to be checked the existence of the same candidate link in other criteria. Then transforms to the matrix of existence of translation link in each criteria for using in Model Construction module.

Table 5.8 Number of link created in each criteria

No	Criteria	# of Links
1.	Monosemic 1:1	7,784
2.	Monosemic 1:N	1,688
3.	Monosemic M:1	5,463
4.	Monosemic M:N	1,802
5.	Polysemic 1:1	856
6.	Polysemic 1:N	332
7.	Polysemic M:1	1,230
8.	Polysemic M:N	88,958
9.	Variant	5,067
10.	Intersection	26,929
11.	Parent	6,130
12.	Brother	21,277
13.	Distance Hyponymy	489

The result have to be checked the existence of the same candidate link in other criteria. Then transforms to the matrix of existence of translation link in each criteria. The illustration of the matrix of existence of translation link in all criteria is shown in the following (Table 5.9).

Copyright by Mahidol University

Table 5.9 The matrix of existence of translation link in all criteria.

Eword	Edefinition	Tword	c01	c02	c03	c04	c05	c06	c07	c08	c09	c10	c11	c12	c13
girl	a friendly informal reference to a grown woman	เด็กหญิง	0	0	0	0	0	0	0	1	0	0	0	1	0
abandonment	the act of giving something up	การสละ	0	0	0	0	0	0	0	1	0	0	0	1	0
Thai	a branch of the Tai languages	ภาษาไทย	0	0	0	0	0	0	0	1	0	1	0	0	0
desertion	withdrawing support or help despite allegiance or responsibility	การทอดทิ้ง	0	0	0	0	0	0	0	0	0	1	0	0	0

c01 = monosemic1:1
c02 = monosemic1:N
c03 = monosemicM:1
c04 = monosemicM:N
c05 = polysemic1:1
c06 = polysemic1:N
c07 = polysemicM:1
c08 = polysemicM:N
c09 = Variant
c10 = Intersection
c11 = Parent
c12 = Brother
c13 = Distance hyperonymy

5.2.2 Sampling

There are tremendous amount of candidate links from Candidate Link Constructor. If all candidate links are verified, this will consume a lot of time. Therefore, a statistical approach is considered to accelerate the process of verification. It starts with sampling candidate links. Then, the logistic regression is applied to construct a model that will be used for verifying all candidate links.

In our work, the stratified sampling is used. In the startified sampling, the population is divided into stratum and then random sampling is applied within each stratum. For our system, the candidate links are already classified in to several startum (criteria). Then the candidate links are randomly sampled with the number of sampling of each stratum at around 400 links. In the appendix B, Determining the Random Sample Size from a Determined Population.(with 95 % confidence level, the most common confidence level for research purposes). In the appendix B the maximum number of sample are not more than 400, the we also choose 400 for our data.

Copyright by Mahidol University

Startified sampling is perfomed with the condition that all samples in each startum should not exist in other startum, to avoid inconsistency of verification and to haste the verification process.

5.2.3 Verification

The verification of each candidate link in the sample has to be performed manually. The translation links will be manually evaluated by using existing Thaisoft So Sethaputra, LEXiTRON, and WordNet 2.0. The specification of evaluator is the expert in both Thai and English language, but in our research the specification of evaluator will be compromised to non-expert but using standard well-known bilingual dictionary in evaluation. The evaluation of candidate link considering mainly on Thai word which matches English definition (English concept).

The candidate links are randomly sampled with the number of sampling of each stratum around 400 links with the condition that all samples in each startum should not exist in other startum, to avoid inconsistency of verification and to haste the verification process. Then the samples are verified manually. The evaluation of candidate links considering mainly on Thai word which matches English definition (English concept). The result of manually the sample are shown in Figure 5.10 and Table 5.10. From this point the matrix of boolean variables of each criteria is create from the result of verification.

Copyright by Mahidol University

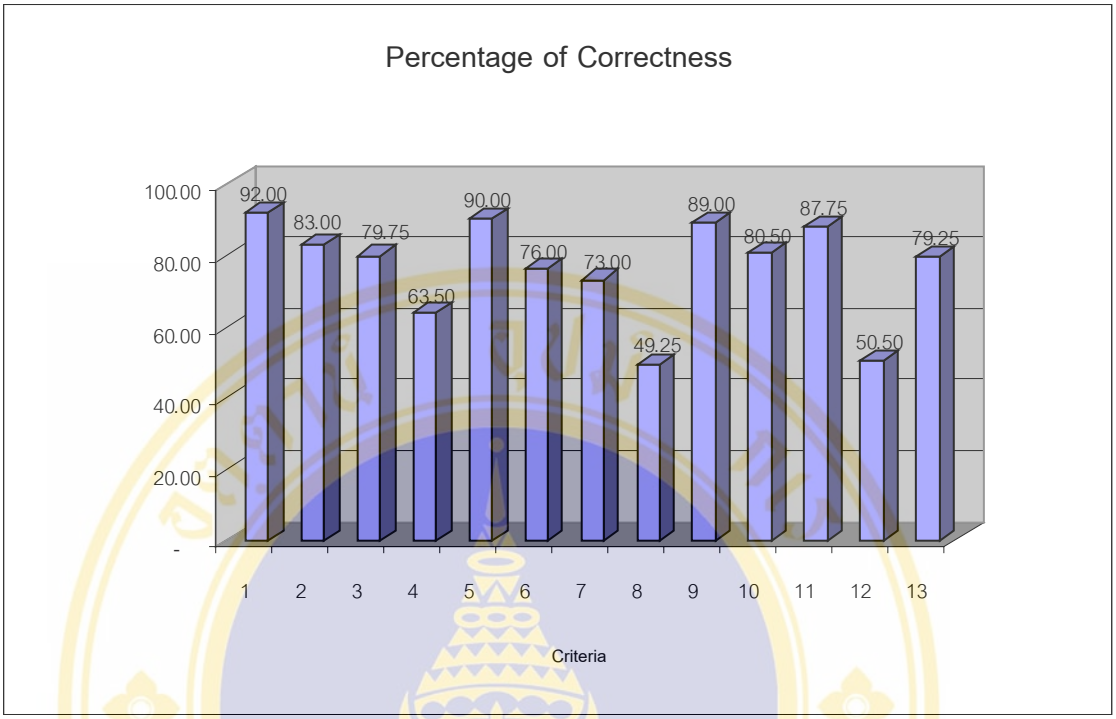


Figure 5.10 Correctness of each criteria (sample)

Table 5.10 Correctness of each criteria (sample size=400)

Criteria	No of OK	%
1	368	92.00
2	332	83.00
3	319	79.75
4	254	63.50
5	360	90.00
6	304	76.00
7	292	73.00
8	197	49.25
9	356	89.00
10	322	80.50
11	351	87.75
12	202	50.50
13	317	79.25

From the result of verification, we notice that criteria 1 Monosemic1:1 has the highest accuracy, which is 92% and the lowest is from criteria 8 Polysemic M:N, which is 49.25%.

Copyright by Mahidol University

The result of the evaluation are transformed to the matrix of boolean variables of each criteria, which are the sum of distinct pattern of as illustrated in Table 5.11 (where NOK = Number of correct link, NTOT = total number of link verified)

Table 5.11 The matrix of boolean variables of each criteria

NTOT	NOK	c01	c02	c03	C04	c05	c06	c07	c08	c09	c10	c11	c12	c13
345	345	1	0	0	0	0	0	0	0	0	0	0	0	0
301	301	0	0	1	0	0	0	0	0	0	0	0	0	0
281	281	0	0	0	0	0	0	1	0	0	0	0	0	0
258	258	0	0	0	0	0	0	0	0	1	1	0	0	0
1	0	0	0	0	1	0	0	0	0	0	0	1	0	0
14	0	0	0	0	1	0	0	0	0	0	1	0	0	0
41	0	0	1	0	0	0	0	0	0	0	0	0	0	0

5.2.4 Model Construction

Logistic regression, which is used to predict a discrete outcome from a set of binary variables, will be used for finding the model for approximating P(OK) (probability of correctness of each candidate link) and finding significant of the criteria. The result of evaluation matrix is used as a training set to construct model. The detail of Logistic regression is shown in Appendix A. The probability that a link would be correct can be estimated by P(OK)=NOK/NTOT, where OK is the Correct link. And the least squares criterion will be used for estimating every value of β_i (unknown parameters) to fit the formula $\log(P(nok/ntot)) = \beta_0+\beta_1 C_1 +...+\beta_{13} C_{13}$. The result of this process is the regression model, which includes coefficient of each factor.

In our work, MiniTab, a statistical software that provides a wide range of data analysis, is used to find the Logistic regression model. Where the criteria 01-13 are factors of the model (independent variable) and NOK and NTOT are used to calculate the P(OK) (dependent variable) to calculate coefficients of the model. And after the model is constructed, the model evaluation must be performed to assure the fitting of the model.

Copyright by Mahidol University

The matrix of boolean variables of each criteria is used in constructing the model using Logistic Regression. For the model construction and evaluation, we first found that criteria 8 Polysemic M:N is a factor (criteria) that give the insinificant value. P-value indicate the sinificant of the factor effect model. P-value of criteria 8 (polysemic M:N) is more than 0.05 at the first time. It means that criteria 8 are not significant to the model. Then we try to remove criteria 8 from the model and construct model again. The second time, no other P-value of other criteria are more than 0.05 that means all coefficients are significant in the model. The result of Model Construction (include c08) is shown in Table 5.12 , the result of Model Construction (exclude c08) is shown in Table 5.13 and detail of both result are shown in Appendix C.

Table 5.12 Coefficient of each factor in the first model

Odds 95% CI		
Predictor	Coef	P
Constant	-0.2500	0.032
C01	2.6940	0.000
C02	1.0314	0.000
C03	1.5946	0.000
C04	0.3891	0.008
C05	2.4082	0.000
C06	1.9264	0.000
C07	1.2209	0.000
C08	-0.0255	0.817
C09	0.5221	0.002
C10	1.4567	0.000
C11	1.6409	0.000
C12	0.3825	0.000
C13	0.5202	0.000

Method	Chi-Square	DF	P
Pearson	154.663	94	0.000

Copyright by Mahidol University

Table 5.13 Coefficient of each factor in the second model

Odds 95% CI		
Predictor	Coef	P
Constant	-0.3877	0.003
c01	2.8314	0.000
c02	1.0943	0.000
c03	1.7286	0.000
c04	0.4997	0.001
c05	2.5447	0.000
c06	2.0337	0.000
c07	1.3575	0.000
c09	0.5440	0.001
c10	1.5421	0.000
c11	1.7096	0.000
c12	0.4664	0.000
c13	0.5774	0.000

Method	Chi-Square	DF	P
Pearson	162.911	77	0.000

5.2.5 Model Evaluation

To evaluate the statistical model, the P-value of each factor (criteria) are the indicator of the appropriate model. P-value is less than the required 0.05 (the odd 95%) and as such there is statistical evidence that at least one factor contributes in the proposed model. If the P-value of factors are more than 0.05, it means that the factors are not significant in the model. For backward method, these factors are remove from model until the P-value of all factors are less than 0.05.

And other methods to evaluate the model is the P-value of Pearson goodness of fit test, if it equals to 0 means the model is fit to the data. Both the first time and the second time, P-value of Pearson goodness of fit test equal to 0 that means the model is fit to the data. . The result of Model Construction (include c08) is shown in Table 5.12 , the result of Model Construction (exclude c08) is shown in Table 5.13 and detail of both result are shown in Appendix C.

Copyright by Mahidol University

5.2.6 Model Employment

The evaluated model is employed to all remain translation links, by applying all coefficients into the formula E(P(OK)) = $\text{Exp}(\beta_0+\beta_1C_1 +...+\beta_{13}C_{13}) / (1 + \text{Exp}(\beta_0+\beta_1C_1 +...+\beta_{13}C_{13}))$ to calculate the expected probability of each translation link. The example of P(OK) of each link are shown in Table 5.14

Table 5.14 The example of E(P(OK)) of each link

Eword	Definition	Tword	c01	c02	c03	C04	c05	c06	c07	c08	c09	c10	c11	c12	c13	E(P(OK))
match	a burning piece of wood or cardboard	คู่ที่เหมือนกัน	0	0	0	0	0	0	0	1	0	0	0	1	0	0.5099
suffering	a state of acute pain	ความปวดร้าว	0	0	0	0	0	0	0	0	0	0	0	1	0	0.5099
centiliter	a metric unit of volume equal to one hundredth of a liter	เซนติลิตร	0	1	0	0	0	0	0	0	1	1	0	0	0	0.9906
mandioc	cassava with long tuberous edible roots and soft brittle stems	มันสำปะหลัง	0	0	0	0	0	0	0	0	0	0	1	0	0	0.7544

The result of applying the first and second model to all candidate links are ploted as line graph of the number of translation links V.S. E(P(OK)) as shown in Figure 5.11 and Figure 5.12 respectively. The graph shown that the coverage and E(P(OK)) are inverse of each other. The graph shown that including criteria 8 in model has the expected probability lower than excluding criteria 8, which are harmonious with the result of model evaluation. Since the second model is more significant, then we focus only the Coverage V.S. P(OK) (After remove C08). From the graph, we cut point at accuracy equal to 76% and coverage equal to 44,844. The E(P(OK)) of the candidate links at coverage (number of translation links) nearly 50% are accepted. Then we accept Thai WordNet at 76% of E(P(OK)). The selected link will be used in next step (Wordnet Constructor).

Copyright by Mahidol University

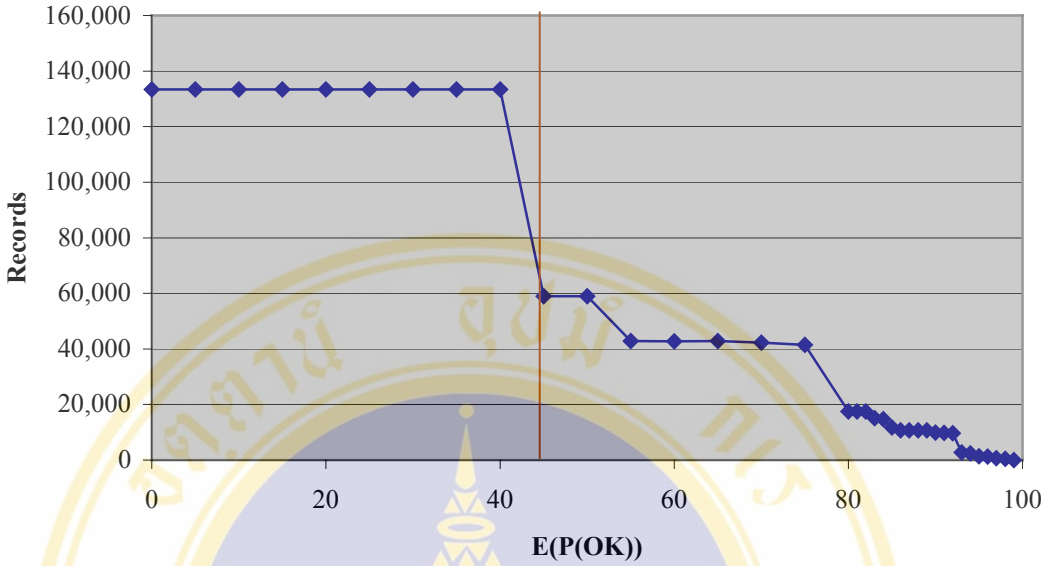


Figure 5.11 Coverage V.S. E(P(OK)) (Before remove C08)

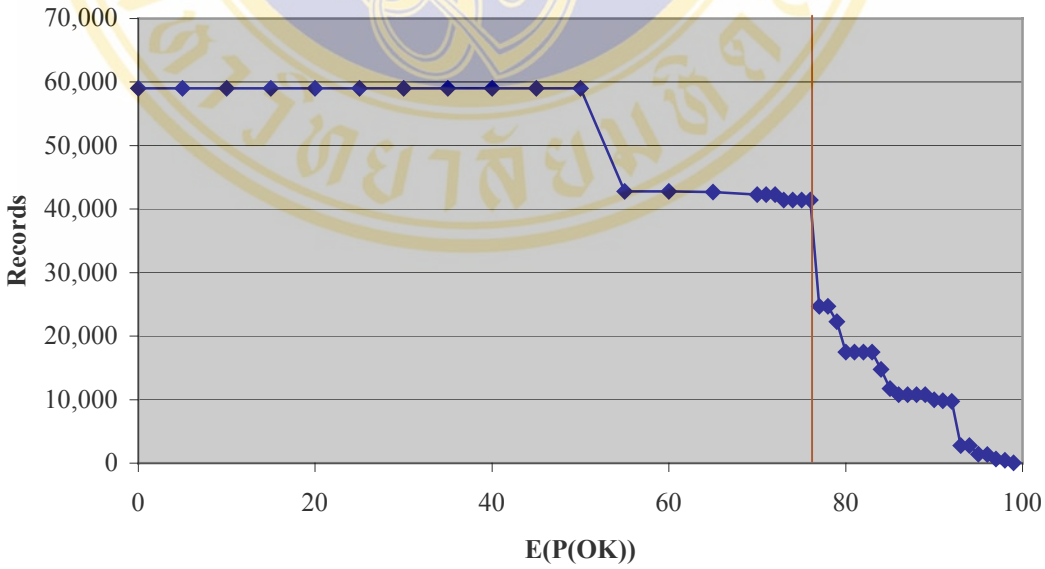


Figure 5.12 Coverage V.S. E(P(OK)) (After remove C08)

Copyright by Mahidol University

5.3 WordNet Constructor

The last module, constructing Thai WordNet from the selected Translation link and English WordNet are performed by attached Thai words to English WordNet.

5.3.1 Synset Attachment

Utilize English WordNet as a skeleton of Thai WordNet, translates each English word into Thai word using accepted translation link, then all Thai word will be attached to English synset. All Thai words (19,582 Thai words) attached to 13,730 English synsets.

5.3.2 Relation Inference

The internal relations of English WordNet (for example, Hypernym, Hyponym, Meronym, and Holonym) are inferred to Thai WordNet. The theory of Transitive relation (in basic discrete mathematics) will be adapted for inference the relations of English WordNet. The 57,494 Thai semantic relations and 650 Thai lexical relations are inferred from nominal English WordNet.

5.3.3 Glossary Attachment

Thai glossary, for example sample, definition, and etc., are attached to Thai WordNet to fulfill Thai WordNet. There are few Thai words with glossary, we attached all glossary we have to Thai WordNet.

5.4 Summary

From the result of our experiment, we accept the accuracy at coverage 76%, 44,844 links (word sense), 19,582 Thai words, and 13,730 synsets. The accuracy and coverage of our Thai WordNet are acceptable, in addition our work use less time comparing to the conventional method of manually creating the model from scratch.

Copyright by Mahidol University

From our experiment, difficulties, performance and limitations of the methodology will be discussed below.

5.4.1 Difficulties

The difficulties are found throughout the developmental phase. First of all, insomuch as we using the existing resources, which are unfit, then data preparation process, which is a time-consuming process, must be performed. And since the verification of candidate of translation link process in our approach has the human involving, then it depends on subjective decisions. In most of NLP system, this problem is usually faced, including in WordNet construction of other languages. Finally, as a result of using the English WordNet as the skeleton of Thai in this approach, there still have some missing structures of Thai language. The structure of English and European language are close and has similar network of relations but for Thai, quite divergence. There are some structural discrepancies on divergences between the languages.

5.4.2 Performance

The performance of the system is measured in terms of accuracy and coverage.

Accuracy

The aim of this research is to construct Thai WordNet and the most considerable is the accuracy of the result. From our result, we accept the accuracy at 76% precision. Even comparing with Spanish we has less accuracy than Spanish WordNet (85.8%), but greater than Italian WordNet (70%) and Chinese WordNet (69.23% using in IR). Our result is moderate and make sense for the non-European language, in which English, Spanish and Italian language are.

Copyright by Mahidol University

Coverage

For coverage of Thai WordNet, we compare with Spanish WordNet (noun category only), Italian WordNet and Chinese WordNet (all categories). In contrast with accuracy, less coverage but more accuracy for our approach. The following is the comparison of the coverage among the WordNet. Our coverage of Thai WordNet, comparing with others, is definitely reasonable.

Table 5.15 Coverage comparison

	Chinese	Italian	Spanish	Thai
Word senses	-	37,235	47,831	**44,844**
Words	62,995	26,463	11,141	**19,582**
Synsets	-	20,571	16,477	**13,730**

5.4.3 Limitations

The limitations of the methodology are described below.

- The structure of each language are not exactly the same, then there are some structural discrepancies
- Thai taxonomy is inconsiderable in this methodology, but there are some criteria, which considering both Thai and English taxonomy, could be added in candidate link construction process.
- This methodology of WordNet construction is merely testified with nominal word.
- The coverage of Thai WordNet is limited to the resources and the accepted accuracy.

Copyright by Mahidol University

CHAPTER 6

CONCLUSION AND FUTURE WORK

In natural language processing applications, wordnet lexical databases have become one of the most important information resources. Construction of wordnet lexical databases, particularly for Thai language, has been a major, ongoing concern. Manual construction is cumbersome, labor-intensive, expensive and time-consuming. The intent of this thesis is to determine the semi-automatic approach to accelerate the construction of Thai WordNet lexical database.

6.1 Conclusion

In this thesis, we have presented the semi-automatic method to construct Thai WordNet. Our methodology takes advantage of the statistical model to reduce human intervention. The model is built upon classification of links which represent relationships between lexical entries. We classify links into thirteen groups in which stratified sampling is applied. Then, the sample links are verified by human. Our model has been evaluated and deployed to construct the Thai WordNet lexical database with 76% accuracy of semantic links which describe relation among words. Our ThaiWordNet lexical database consists of 44,844 semantic links connecting 13,730 words and 19,582 synonym sets.

Our semi-automatic approach is couched within the WordNet Builder system. The system takes into account two existing machine-readable dictionaries: English WordNet and LEXiTRON. The English WordNet provides semantic relations among lexical entries which represents English words. The LEXiTRON provides translation links which represents relationships between English and Thai words. Both dictionaries are converted into relational tables which RDBMS can manipulate conveniently. For each relational table, missing, duplicate, noisy, and erroneous data is

Copyright by Mahidol University

handled. Then, the data from multiple tables are integrated and only relevant data is selected according to the link classification.

Our semi-automatic approach is couched within the WordNet Builder system. The system takes into account two existing machine-readable dictionaries: Princeton WordNet and LEXiTRON. The Princeton WordNet provides semantic relations among lexical entries which represents English words. The LEXiTRON provides translation links which represents relationships between English and Thai words.

Like other wordnets, Thai WordNet organizes words around their semantics instead of alphabetical order. Therefore, it can be considered as a thesaurus which can be used to search for synonyms, antonyms and related words. In Thai WordNet, we have attached meaning description associated with each word sense. The description are represented in both English and Thai; hence people can use the Thai WordNet as a bilingual electronic dictionary.

6.2 Suggestion for future work

Our semi-automatic construction of Thai WordNet lexical database from machine-readable dictionaries allows further technical and methodological improvements. The suggestions for further extension are as follows:

- To expand Thai WordNet, other lexical resources could be added to increase the volume of Thai WordNet.

- To merge Thai WordNet with existing Thai lexical-conceptual taxonomy (such as Thai concept type hierarchy), Thai taxonomy could be added for automatically linking Thai words to English taxonomy.

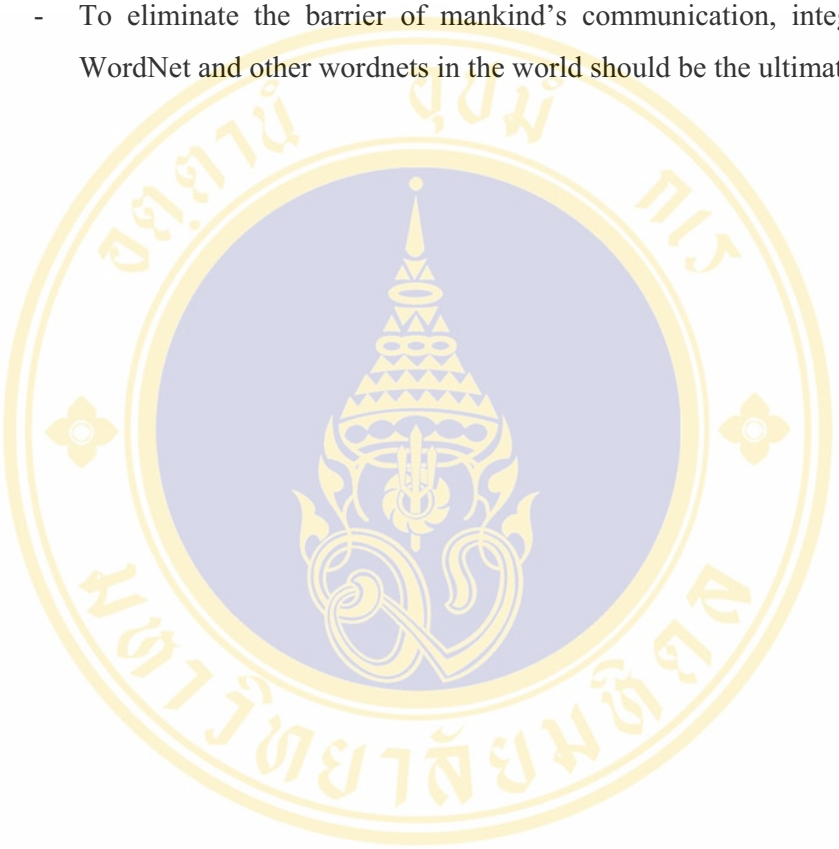
- To fulfill Thai WordNet, Thai word verbs, adjectives and adverbs can be added.

- To improve the methodology of constructing Thai WordNet, other techniques in information retrieval, information extraction, and conceptual distances could be applied in a stepwise refinement.

Copyright by Mahidol University

- To increase the correctness of Thai WordNet, the greater degree of lexicographer's involving in verifying the link would be considered. However, trading-off between the economical issue and the correctness of the result should be considered.

- To eliminate the barrier of mankind's communication, integrating Thai WordNet and other wordnets in the world should be the ultimate goal.



Copyright by Mahidol University

REFERENCES

1. Atserias J., Climent S., Farreres X., Rigau G. and Rodríguez H., Combining Multiple Methods for the automatic Construction of Multilingual WordNets, in proceedings of of International Conference "Recent Advances in Natural Language Processing" (RANLP'97).Tzigov Chark, Bulgaria, 1997
2. Babylon's dictionaries http://www.babylon.com
3. Boguraev B. and Briscoe T., Computational Lexicography for Natural Language Processing. Longman, Cambridge, England. 1989.
4. Chen, H .H., Lin, C. C., and Lin, W. C. 2000, Construction of a Chinese-English WordNet and its application to CLIR, in Proceedings of the Fifth International Workshop on Information Retrieval with Asian Languages (Hong Kong, Sept.-Oct. 2000). ACM Press, New York, NY, 189-196.
5. Chen, H .H., Lin, C. C., and Lin, W. C.2002, Building a Chinese-English WordNet for Translingual Applications, ACM Transactions on Asian Language Information Processing (TALIP) Volume 1 , Issue 2 (June 2002) Pages: 103 – 122
6. Church K., and Hanks P., Word association norms, mutual information, and lexicography, Computational Linguistics, vol. 16, ns. 1, 22-29. 1990. Also in proceedings of the 27th Annual Meeting of the Association for Computational Linguistics (ACL'89). Pittsburg, Pennsylvania, 1989.
7. Collins Cobuild English Language Dictionary http://www.cobuild.collins.co.uk/
8. Dolan W., Vanderwende L. and Richard son S., Automatically deriving structured knowledge bases from on-line dictionaries, in proceedings of the first Conference of the Pacific Association for Computational Linguistics (Pacling'93), April 21-24, Simon Fraser University, Vancouver, Canada. 1993.
9. Farreres J., Rodríguez H., Gibert K., Semiautomatic creation of taxonomies In Proceedings of COLING'02 Workshop "SemaNet: Building and Using Semantic Networks". Taipei, Taiwan. 2002

Copyright by Mahidol University

10. Farreres X., Rigau G., and Rodríguez H., 1998, Using wordnet for building wordnets, in Proceedings of COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems.
11. Gomez F., Hull R. and Segami C., Acquiring Knowledge From Encyclopedic Texts, in proceedings of the 4th Conference on Applied Natural Language Processing, (ANLP'94), pages 84-90, Stuttgart, Germany, October, 1994.
12. Grefenstette G., Evaluation Techniques for Automatic Semantic Extraction: Comparing Syntactic and window Based Approaches, in proceedings of SIGLEX Workshop on Acquisition of Lexical Knowledge from Text. Columbus, OH. 1993.
13. Gruber, T. R., A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition, 5(2), 199-220, 1993
14. Kristin Javaras and Wiesner Vos., Introduction to Generalized Linear Models, Lecture notes in Introduction to Statistical Modeling, Trinity 2002. Department of Statistics Oxford University.
15. Krejcie, R. V. and Morgan, D. W. (1970). Determining sample size for research activities. Educational and Psychological Measurement, 30, 607-610.
16. Lenat D. and Guha R., Building Large Knowledge-based Systems: Representation and Inference in the Cyc Project. Addison Wesley. 1990.
17. LEXiTRON http://lexitron.nectec.or.th
18. Longman Dictionary of Contemporary English http://www.longman.com/ldoce/
19. Miller G., Beckwith R., Fellbaum C., Gross D., and Miller K., Introduction to WordNet: An On-line Lexical Database (Revised August 1993) International Journal of Lexicography 3 (4), 1990, Revised August 1993
20. Miller G., Five papers on WordNet, Special Issue of International Journal of Lexicography 3(4). 1990.
21. Mutsuko Tomokiyo, Mathieu Mangeot-Lerebours and Emmanuel Planas., Papillon : A Project of Lexical Database for English, French and Japanese, using Interlingual Links, Journées Science et Technologie de l'ambassade de France au Japon, 13 novembre 2000, Tokyo, Japon, 3 p.
22. Neff M., Blaser B., Lange J-M., Lehmann H. and Dominguez., Get it where you can: Acquiring and Maintaining Bilingual Lexicons for Machine Translation,

Copyright by Mahidol University

Paper presented at the AAAI Spring Simposium on Building Lexicons for Machine Translation, Stanford University. 1993.

23. Oxford Advanced Lerner's Dictionary http://www.oup.com/elt/oald/
24. Palingoon P., Chantanapraiwan P., Theerawattanasuk S., Charoenporn T., Sornlertlumvanich V., Qualitative and Quantitative Approaches in Bilingual Corpus-Based Dictionary, PROCEEDINGS OF SNLP-Oriental COCOSDA 2002.
25. Pianta E., Bentivogli L., Christian Girardi, MultiWordNet Developing an aligned multilingual database, ITC-irst, Centro per la Ricerca Scientifica e Tecnologica
26. Richard Bergmair, wordnet2sql http://wordnet2sql.infocity.cjb.net
27. Rigau G.,. Automatic Acquisition of Lexical Knowledge from MRDs. Ph.D. thesis, Polytechnic University of Catalonia, 1998.
28. Royal Institute Dictionary http://www.royin.go.th
29. Sager N., Natural Language Processing, Addison-Wesley, Reading, Mass. 1981.
30. Saikam http://saikam.nii.ac.jp
31. Sornlertlamvanich V., Boriboon M., Technical Report - Thai Concept Classification. Linguistics and Knowledge Science, 1995.
32. Uchida H., Electronic Dictionary, in proceedings of the International Workshop on Electronic Dictionaries, OISO, Kanagawa, Japan, pp. 23-43. 1990.
33. Vossen P., 1999, EuroWordNet General Document Version 3 Final University of Amsterdam EuroWordNet LE2-4003, LE4-8328. 1999
34. Vossen P., Bloksma L., Boersman P., The Dutch WordNet Version 2, Final, University of Amsterdam
35. Wikipedia encyclopedia http://en.wikipedia.org
36. Wilks Y., Slator B. and Guthrie L., Electric Words: Dictionaries, Computers, and Meanings. The MIT Press, Cambridge, MA. 1996.
37. Yarowsky D., Word-Sense Disambiguation Using Statistical Models of Rogets Categories Traiend on Large Corpora, in proceedings of the 14th International Conference on Computational Linguistics (COLING'92), pages 454-460, Nantes, France, 1992.

Copyright by Mahidol University

38. Yokoi T., The Impact of the EDR Electronic Dictionary on Very Large Knowledge Bases, Towards Very Large Knowledge Bases, Mars N. (Ed.). IOS Press. 1995.

Copyright by Mahidol University

APPENDIX

Copyright by Mahidol University

APPENDIX A

LOGISTIC REGRESSION

A.1 Introduction

Regression methods have become an integral component of any data analysis concerned with describing the relationship between a response variable and one or more explanatory variables. The Logistic Regression [14] is the standard method of analysis in the case that outcome variable is discrete, taking on two or more possible values (binary or dichotomous variable). Logistic Regression generates the coefficients of a prediction formula (and standard errors of estimate and significance levels), and odds ratios (with their 95% confidence intervals). The goal of an analysis using logistic regression are : to find the best fitting and most efficient and describe the relationship between an outcome (dependent or response variable) and a set of independent variable (predictor or explanatory).

The situation where the response, Y, can only take one of two possible values. In practice, these responses may be alive or dead, or present or absent. This situation also arises frequently in medical trials, where at the end of the trial period, the patient has either recovered or has not. It is convenient to denote the two levels by 0 and 1 and to refer to the categories as a "failure" or a "success". The main objective is to investigate the relationship between the response probability $\pi = \pi(\underline{x})$ and the explanatory variables $x_1, x_2, \ldots, x_n$. A binary random variable can be defined as

$$Z = \begin{cases} 1 \text{ if the outcome is a success} \\ 0 \text{ if the outcome is a failure} \end{cases}$$

with probabilities $P(Z=1) = \pi$ and $P(Z=0) = 1-\pi$. Typically, there are a number of explanatory variables or covariates $x_1, x_2, \ldots, x_n$ associated with each observation in the study. For example, in a designed experiment the covariates may

Copyright by Mahidol University

consist of indicator variables associated with blocking or related to treatment factors, together with other quantitative variables relating to the response. In an observational study the covariates may mainly consist of measured variables relating to the probability of a positive response.

It is important to distinguish between grouped and ungrouped data when we are dealing with binary data. Suppose that we have N independent random variables, $Y_1, Y_2, \ldots, Y_N$, corresponding to the numbers of successes in N different subgroups. It is assumed that each random variable is distributed $Y_i \sim binomial(n_i, \pi_i)$. Thus, we are dealing with N independent binomial distributions. In other words, of the $n_{total} = n_1 + n_2 + \cdots + n_N$ observations in the study, n_i share the covariate values $x_{i1}, x_{i2}, \ldots, x_{ip}$. These observations are said to form a covariate class. This situation is illustrated below:

	Subgroups			
	1	2	…	N
Successes	Y_1	Y_2	…	Y_N
Failures	$n_1 - Y_1$	$n_2 - Y_2$	…	$n_N - Y_N$
Totals	n_1	n_2	…	n_N

When binary data are grouped by covariate class, the responses have the form $y_1/n_1, \ldots, y_N/n_N$, where $0 \le y_i \le n_i$ is the number of successes out of the n_i subjects in the ith covariate class. Ungrouped data, or data listed by individual subjects, can be considered as a special case for which $n_1 = \ldots = n_N$.

Copyright by Mahidol University

Our goal is to describe the proportion of success, $P_i = Y_i / n_i$, in each subgroup in terms of factor levels and other explanatory variables. From the fact that $Y_i \sim binomial(n_i, \pi_i)$, it follows that $E(Y_i) = n_i \pi_i$ and $E(P_i) = \pi_i$. So within the generalized linear model set-up we can model the probabilities π_i as

$$g(\pi_i) = \beta_{i0} + \beta_{i1} x_{i1} + \ldots + \beta_{ip} x_{ip}$$

where $x_1, \ldots, x_p$ are the observed explanatory variables, $\beta_{i1}, \ldots, \beta_{ip}$ are the parameters and g is a link function. The explanatory variables can be dummy variables for factors and measured values for covariates. A simple method to model the probability of 'success' would be the linear model

$$\pi = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p .$$

The disadvantage of this approach is that π is a probability and is therefore restricted to the interval $[0,1]$, while the fitted values are not restricted to this interval. The next subsection will introduce the general logistic regression model that uses link function, g, that is different from the identity link and which overcomes the above problem.

Logistic regression model

The general linear logistic regression model is defined as

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \text{logit}(\pi_i) = \beta_{i0} + \beta_{i1} x_{i1} + \cdots + \beta_{ip} x_{ip}$$

where $x_{i1}, \ldots, x_{ip}$ are continuous measurements corresponding covariates and/or dummy variables corresponding to factor levels and $\beta_{i1}, \ldots, \beta_{ip}$ are the parameters. This model is very widely used for analysing data involving binary or binomial responses and several explanatory variables. Estimates for the parameters and response probabilities are typically obtained by the method of maximum likelihood. These estimates will be computed and returned by your software package.

Copyright by Mahidol University

Note that the general linear logistic regression model can be rewritten in terms of the probability of a positive response

$$\pi_i = \frac{\exp\left(\beta_{i0} + \beta_{i1}x_{i1} + \ldots + \beta_{ip}x_{ip}\right)}{1 + \exp\left(\beta_{i0} + \beta_{i1}x_{i1} + \ldots + \beta_{ip}x_{ip}\right)}.$$

A.2 Goodness of fit

To evaluate the goodness of fit we can calculate the deviance. There are no nuisance parameters (like σ^2 for the Normal distribution), so goodness of fit can be assessed and the hypotheses can be tested directly using the approximation

$$D \sim \chi^2(N - p)$$

under the hypothesis that the model is correct, where p is the number of parameters estimated and N the number of covariate classes. It must be noted that the above approximation can be poor if the number of observations in the respective covariate classes are small.

Another goodness of fit statistic that is used is the Pearson chi-square statistic that can also be tested using the approximation

$$X^2 \sim \chi^2(N - p).$$

X^2 is asymptotically equivalent to the deviance. The choice between D and X^2 depends on the adequacy of the approximation to the $\chi^2(N - p)$ distribution.

A.3 Model Diagnostics

There are two main forms of residuals used to evaluate the model adequacy: the Pearson residual and the deviance residual. There are also standardized versions of each of these types of residuals. Sometimes there may be very few distinct values in each group or covariate class that may lead to uninformative plots. In cases like these we may have to rely on the deviance, Pearson chi-square or other statistics.

Copyright by Mahidol University

A.4 Over-dispersion

Over-dispersion is another important issue that must be given consideration when assessing the adequacy of models for binary or binomial data. Over-dispersion refers to the situation where the observations Y_i have a variance larger than that expected under the binomial model, i.e. $n_i\pi_i(1-\pi_i)$. This is phenomenon is quite common in practice. It can occur for various regions, such as an inadequate model specification (e.g., relevant explanatory variables have been omitted or the link function is incorrect). One approach to deal with over-dispersion in this case is to include an extra parameter ϕ in the model so that $\text{var}(Y_i) = n_i\pi_i(1-\pi_i)\phi$. The implementation of this depends on the statistical software you use. Another reason for over-dispersion may be that the Y_i are correlated. In this case we must use modelling approaches designed for correlated data. A more detailed discussion of over-dispersion can be found in Nelder and McCullagh (1989).

A.5 Example

Weisberg (1985) reports data on an experiment carried out on cows[1]. The effect of small electrical currents on farm animals is of interest, with the eventual goal of understanding the effects of high-voltage powerlines on livestock. This experiment was carried out with 7 cows, and 6 shock intensities 0, 1, 2, 3, 4 and 5 milliamps. It is reported that shocks in the order of 15 milliamps are painful for any human. Each cow was given 30 shocks, five at each intensity level, in random order. The entire experiment was then repeated, so each unfortunate cow received a total of 60 shocks! For each shock, the response, mouth movement, was either present or absent. The data in Table 1 give the total number of responses, out of the 70 resulting trials for each shock level. We ignore differences between cows and blocks in this experiment.

---

[1] The data were taken from Weisberg, S. (1985). *Applied linear regression (2nd ed.)*. Toronto: Wiley.

Copyright by Mahidol University

We use statistical software program to fit the following logistic regression model to the data:

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_i$$

Binary logistic regressions models can be fitted using either the Logistic Regression procedure or the Multinomial Logistic Regression procedure. For example, the fitted values and residuals are only available in the Logistic regression procedure, while the results from the deviance and Pearson goodness of fit tests are only available in the Multinomial Logistic Regression procedure. Each procedure has options not available in the other. The result summaries from the fit of the above model in statistical software program are presented below Table 1.

Cow shock data

Current (Mill amperes)	Number of Trials	Number of Responses	Proportion of Responses
0	70	0	0.000
1	70	9	0.129
2	70	21	0.300
3	70	47	0.671
4	70	60	0.857
5	70	63	0.900

Likelihood Ratio Test

Model	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	271.227			
Final	30.093	241.134	1	.000

Copyright by Mahidol University

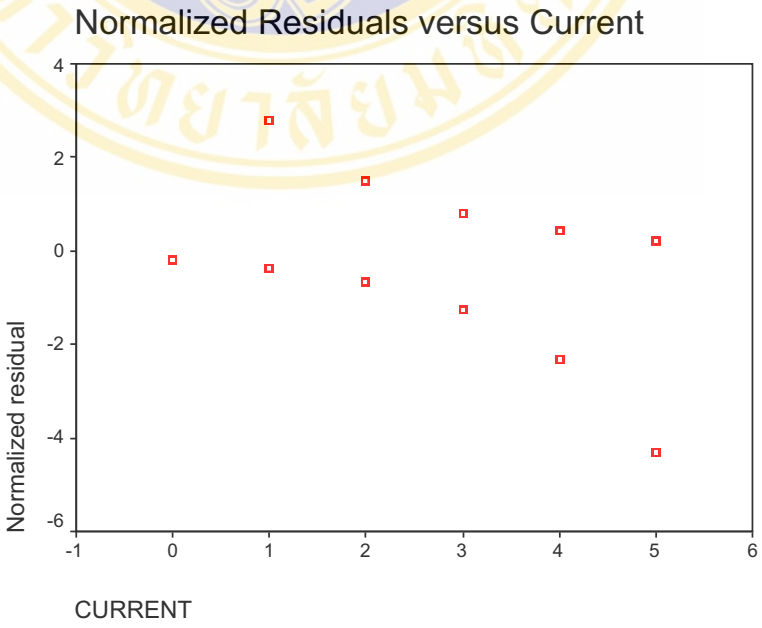
Goodness-of-Fit

	Chi-Square	df	Sig.
Pearson	7.583	4	.108
Deviance	9.353	4	.053

Parameter Estimates

	B	S.E.	Wald	df	Sig.	Exp(B)
b1	1.246	.112	123.918	1	.000	3.476
Constant	-3.301	.324	103.928	1	.000	.037

We see that the estimated parameters for the above model are $\beta_0 = -3.301$ and $\beta_1 = 1.246$. The Likelihood ratio test compares the above model to the reduced model with just a constant term β_0. There is clear evidence from this test that the response rate increases with intensity, since the test shows that the model including the β_1 term provides a much better fit. This is confirmed the Wald test.



Standardized residuals versus the dose for the cow shock example.

Copyright by Mahidol University

The plot of the standardized residuals against the current shows some systematic pattern, which suggests that we may need alternative terms in the model. The two curve like patterns represent the residuals of a positive (top curve) and negative (bottom curve) response, respectively. It seems as though the residuals for a negative response decrease as the current increases and the other way around for the residuals relating to a positive response. This lack of fit is also reflected by the relatively small p-value relating to the deviance (0.053).

Copyright by Mahidol University

APPENDIX B

SAMPLE SIZE

The size of the population and amount of error determines the minimum size of a randomly selected sample. This table helps the researcher determine (with 95 percent certainty) what the results would have been if the entire population had been surveyed.

Table 1 Determining the Random Sample Size from a Determined Population [15]

Population	Random Sample size	Population	Random Sample size	Population	Random Sample size
10	10	220	140	1,200	291
15	14	230	144	1,300	297
20	19	240	148	1,400	302
25	24	250	152	1,500	308
30	28	260	155	1,600	310
35	32	270	159	1,700	313
40	36	280	162	1,800	317
50	44	300	169	2,000	322
55	48	320	175	2,200	327
60	52	340	181	2,100	331
65	56	360	186	2,600	335
70	59	380	191	2,800	338
75	63	400	196	3,000	341
80	66	420	201	3,500	346
85	70	440	205	4,000	351
90	73	460	210	4,500	354
95	76	480	214	5,000	357
100	80	500	217	6,000	361
110	86	550	226	7,000	364
120	92	600	231	8,000	367
130	97	650	242	9,000	368
140	103	700	248	10,000	370
150	108	750	254	15,000	375
160	113	800	260	20,000	377
170	118	850	365	30,000	379
180	123	900	269	40,000	380
190	127	950	274	50,000	381
200	132	1,000	278	75,000	382
210	136	11,000	285	100,000	384

Copyright by Mahidol University

APPENDIX C

RESULT FROM MODEL CONSTRUCTION

The statistical software, MiniTAB is used for constructing Logistic Regression model to assist in translation link verification. The following is the sample of using MiniTAB to construct Logistic Regression model, when c01-c013 are the independent variables and dependent variable are NOK and Total.

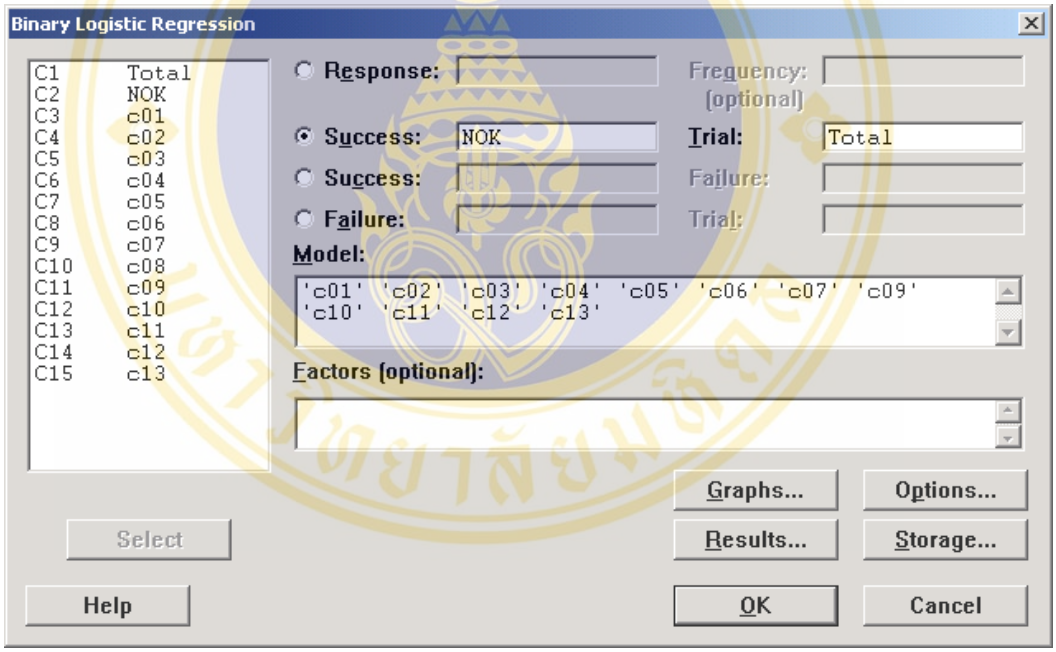


Figure. 1 Binary Logistic Regression screen.

The result of model construction (include c08) and the result of model construction (exclude c08) are shown in Table 1 and Table 2 respectively.

Copyright by Mahidol University

Table 1 The result of Model Construction (include c08)Binary Logistic Regression: NOK, Total versus C01, C02, ...

Link Function: Logit

Response Information

Variable	Value	Count
NOK	Success	3974
	Failure	1153
Total	Total	5127

Logistic Regression Table

					Odds	95% CI	
Predictor	Coef	SE Coef	Z	P	Ratio	Lower	Upper
Constant	-0.2500	0.1168	-2.14	0.032			
C01	2.6940	0.2195	12.28	0.000	14.79	9.62	22.74
C02	1.0314	0.1630	6.33	0.000	2.81	2.04	3.86
C03	1.5946	0.1697	9.40	0.000	4.93	3.53	6.87
C04	0.3891	0.1467	2.65	0.008	1.48	1.11	1.97
C05	2.4082	0.2033	11.85	0.000	11.11	7.46	16.55
C06	1.9264	0.2247	8.57	0.000	6.86	4.42	10.66
C07	1.2209	0.1619	7.54	0.000	3.39	2.47	4.66
C08	-0.0255	0.1104	-0.23	0.817	0.97	0.79	1.21
C09	0.5221	0.1713	3.05	0.002	1.69	1.20	2.36
C10	1.4567	0.1232	11.82	0.000	4.29	3.37	5.46
C11	1.6409	0.1333	12.31	0.000	5.16	3.97	6.70

Copyright by Mahidol University

C12	0.3825	0.1035	3.70	0.000	1.47	1.20	1.80
C13	0.5202	0.1468	3.54	0.000	1.68	1.26	2.24

Log-Likelihood = -2358.615

Test that all slopes are zero: G = 748.409, DF = 13, P-Value = 0.000

Goodness-of-Fit Tests

Method	Chi-Square	DF	P
Pearson	154.663	94	0.000
Deviance	116.718	94	0.056
Hosmer-Lemeshow	10.719	7	0.151

Table of Observed and Expected Frequencies:

(See Hosmer-Lemeshow Test for the Pearson Chi-Square Statistic)

	Group									
Value	1	2	3	4	5	6	7	8	9	Total
Success										
Obs	256	294	435	568	447	795	515	491	173	3974
Exp	252.6	306.8	431.9	568.7	442.8	775.0	527.7	496.2	172.3	
Failure										
Obs	284	259	176	160	91	90	61	30	2	1153
Exp	287.4	246.2	179.1	159.3	95.2	110.0	48.3	24.8	2.7	
Total	540	553	611	728	538	885	576	521	175	5127

Copyright by Mahidol University

Measures of Association:

(Between the Response Variable and Predicted Probabilities)

Pairs	Number	Percent	Summary Measures	
Concordant	3339187	72.9%	Somers' D	0.50
Discordant	1040089	22.7%	Goodman-Kruskal Gamma	0.52
Ties	202746	4.4%	Kendall's Tau-a	0.17
Total	4582022	100.0%		

Table 2 The result of Model Construction (exclude c08)

Results for: Matrix 04-05-04.xls

Binary Logistic Regression: NOK, Total versus c01, c02, ...

Link Function: Logit

Response Information

Variable	Value	Count
NOK	Success	3797
	Failure	969
Total	Total	4766

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI Lower	95% CI Upper
Constant	-0.3877	0.1287	-3.01	0.003			

Copyright by Mahidol University

c01	2.8314	0.2254	12.56	0.000	16.97	10.91	26.40
c02	1.0943	0.1631	6.71	0.000	2.99	2.17	4.11
c03	1.7286	0.1775	9.74	0.000	5.63	3.98	7.98
c04	0.4997	0.1494	3.35	0.001	1.65	1.23	2.21
c05	2.5447	0.2102	12.11	0.000	12.74	8.44	19.23
c06	2.0337	0.2265	8.98	0.000	7.64	4.90	11.91
c07	1.3575	0.1704	7.97	0.000	3.89	2.78	5.43
c09	0.5440	0.1713	3.18	0.001	1.72	1.23	2.41
c10	1.5421	0.1351	11.42	0.000	4.67	3.59	6.09
c11	1.7096	0.1441	11.87	0.000	5.53	4.17	7.33
c12	0.4664	0.1242	3.76	0.000	1.59	1.25	2.03
c13	0.5774	0.1518	3.80	0.000	1.78	1.32	2.40

Log-Likelihood = -2120.899

Test that all slopes are zero: G = 571.518, DF = 12, P-Value = 0.000

Goodness-of-Fit Tests

Method	Chi-Square	DF	P
Pearson	162.911	77	0.000
Deviance	106.243	77	0.015
Hosmer-Lemeshow	21.859	7	0.003

Table of Observed and Expected Frequencies:

Copyright by Mahidol University

(See Hosmer-Lemeshow Test for the Pearson Chi-Square Statistic)

	Group									
Value	1	2	3	4	5	6	7	8	9	Total
Success										
Obs	303	528	427	461	444	440	530	484	180	3797
Exp	331.0	506.7	418.0	463.5	427.5	433.6	546.2	491.1	179.4	
Failure										
Obs	330	206	115	113	61	45	66	31	2	969
Exp	302.0	227.3	124.0	110.5	77.5	51.4	49.8	23.9	2.6	
Total	633	734	542	574	505	485	596	515	182	4766

Measures of Association:

(Between the Response Variable and Predicted Probabilities)

Pairs	Number	Percent	Summary Measures	
Concordant	2609916	70.9%	Somers' D	0.47
Discordant	879251	23.9%	Goodman-Kruskal Gamma	0.50
Ties	190126	5.2%	Kendall's Tau-a	0.15

Copyright by Mahidol University

APPENDIX D

PATTERN OF EXISTENCE OF LINKS IN EACH CRITERIA

Table 1 The matrix of boolean variables of each criteria (with C08)

No	Total	NOK	c01	c02	c03	c04	c05	c06	c07	c08	c09	c10	c11	c12	c13
1.	345	345	1	0	0	0	0	0	0	0	0	0	0	0	0
2.	343	343	0	0	0	0	1	0	0	0	0	0	0	0	0
3.	301	301	0	0	1	0	0	0	0	0	0	0	0	0	0
4.	281	281	0	0	0	0	0	0	1	0	0	0	0	0	0
5.	224	224	0	0	0	0	0	0	0	0	1	1	0	0	0
6.	172	172	0	0	0	0	0	0	0	0	0	1	0	0	0
7.	166	166	0	1	0	0	0	0	0	0	1	1	0	0	0
8.	129	0	0	0	0	1	0	0	0	0	0	0	0	0	0
9.	114	114	0	0	0	1	0	0	0	0	0	0	0	0	0
10.	108	0	0	0	0	0	0	0	1	0	0	0	0	0	0
11.	103	0	0	0	0	0	0	0	0	1	0	0	0	1	0
12.	103	103	0	0	0	0	0	1	0	0	0	0	0	0	0
13.	103	103	0	0	0	0	0	0	0	1	0	0	0	1	0
14.	98	0	0	0	0	0	0	0	0	0	0	0	0	1	0
15.	96	96	0	0	0	0	0	0	0	0	0	0	1	0	0
16.	91	91	0	1	0	0	0	0	0	0	0	0	0	0	0
17.	86	86	0	0	0	0	0	0	0	0	0	0	0	1	0
18.	85	85	0	0	0	1	0	0	0	0	0	1	0	0	0
19.	81	0	0	0	1	0	0	0	0	0	0	0	0	0	0
20.	65	65	0	0	0	0	0	0	0	1	0	0	1	0	0
21.	63	63	0	0	0	0	0	0	0	0	0	0	0	1	1
22.	62	62	0	0	0	0	0	0	0	1	0	1	0	0	0
23.	61	61	0	0	0	0	0	1	0	0	1	1	0	0	0
24.	55	0	0	0	0	0	0	0	0	0	0	1	0	0	0
25.	49	49	0	0	0	0	0	0	0	0	0	0	1	1	0
26.	47	47	0	0	0	0	0	0	0	1	0	0	0	1	1
27.	46	46	0	0	0	0	0	0	0	1	0	0	1	1	0
28.	44	44	0	0	0	0	0	1	0	0	0	1	0	0	0
29.	41	0	0	1	0	0	0	0	0	0	0	0	0	0	0
30.	39	0	0	0	0	0	1	0	0	0	0	0	0	0	0
31.	37	37	0	1	0	0	0	0	0	0	0	1	0	0	0
32.	37	37	0	0	0	0	0	0	0	0	0	1	0	1	0
33.	36	36	0	0	0	0	0	0	0	1	0	0	1	1	1
34.	35	35	0	0	0	0	0	0	0	0	0	1	1	0	0
35.	34	34	0	0	0	0	0	0	0	1	1	1	0	0	0
36.	33	33	0	1	0	0	0	0	0	0	0	0	1	0	0
37.	31	0	1	0	0	0	0	0	0	0	0	0	0	0	0
38.	31	31	0	0	0	0	0	0	0	0	0	0	1	1	1
39.	29	0	0	0	0	0	0	0	0	0	1	1	0	0	0
40.	28	28	0	0	0	0	0	1	0	0	0	0	0	1	0

Copyright by Mahidol University

No	Total	NOK	c01	c02	c03	c04	c05	c06	c07	c08	c09	c10	c11	c12	c13
41.	27	27	0	0	0	0	0	0	0	0	1	1	0	1	0
42.	27	27	0	0	0	1	0	0	0	0	0	0	0	1	0
43.	26	26	0	0	0	0	0	1	0	0	0	0	1	0	0
44.	25	0	0	0	0	0	0	0	0	0	0	0	0	1	1
45.	25	25	0	0	0	0	0	0	0	0	0	1	0	1	1
46.	24	24	0	1	0	0	0	0	0	0	0	0	0	1	0
47.	24	24	0	0	0	0	0	0	0	0	0	1	1	1	1
48.	23	23	0	0	0	0	0	0	0	0	0	1	1	1	0
49.	22	0	0	0	0	0	0	0	0	1	0	1	0	0	0
50.	21	0	0	0	0	0	0	1	0	0	0	0	0	0	0
51.	19	19	1	0	0	0	0	0	0	0	0	0	1	0	0
52.	18	0	0	0	0	0	0	0	0	1	0	0	1	0	0
53.	18	18	0	0	0	0	0	0	0	1	0	1	0	1	0
54.	18	18	0	0	0	1	0	0	0	0	0	0	1	0	0
55.	18	0	0	0	0	0	0	0	0	0	0	0	1	0	0
56.	17	0	0	0	0	0	0	0	0	1	0	0	0	1	1
57.	16	0	0	1	0	0	0	0	0	0	1	1	0	0	0
58.	16	16	0	0	0	0	0	0	0	0	1	1	1	1	1
59.	15	15	0	1	0	0	0	0	0	0	0	0	1	1	0
60.	14	0	0	0	0	1	0	0	0	0	0	1	0	0	0
61.	13	13	0	0	0	0	1	0	0	0	1	1	0	0	0
62.	13	13	0	0	1	0	0	0	0	0	0	0	1	0	0
63.	12	12	0	0	0	0	0	0	0	1	0	1	1	1	0
64.	12	12	0	0	0	0	0	0	0	0	1	1	1	0	0
65.	12	12	0	0	0	0	0	1	0	0	0	0	1	1	0
66.	12	0	0	0	0	0	0	0	0	1	0	0	1	1	1
67.	11	11	0	0	0	0	0	0	0	0	1	1	1	1	0
68.	11	11	0	0	0	0	0	0	0	0	0	0	0	0	1
69.	11	11	0	0	0	1	0	0	0	0	0	1	1	0	0
70.	10	10	0	1	0	0	0	0	0	0	0	1	0	1	0
71.	10	0	0	0	0	0	0	0	0	0	0	0	1	1	1
72.	10	10	0	0	0	0	0	0	0	1	0	1	1	0	0
73.	10	10	0	0	0	0	0	0	0	1	0	1	0	1	1
74.	9	9	1	0	0	0	0	0	0	0	0	1	0	0	0
75.	9	9	0	0	0	0	0	1	0	0	1	1	0	1	0
76.	9	9	0	0	0	0	0	0	1	0	0	1	0	0	0
77.	9	0	0	0	0	0	0	0	0	1	1	1	0	0	0
78.	8	8	0	1	0	0	0	0	0	0	1	1	1	0	0
79.	8	8	0	0	0	1	0	0	0	0	0	1	0	1	0
80.	8	8	0	0	0	0	0	0	0	1	0	0	0	0	1
81.	7	0	0	0	0	0	0	0	0	0	0	0	0	0	1
82.	7	7	0	0	1	0	0	0	0	0	0	1	0	0	0
83.	7	7	0	0	0	0	0	1	0	0	1	1	1	1	0
84.	7	0	0	0	0	0	0	0	0	0	0	1	0	1	0
85.	7	0	0	0	0	0	0	0	0	0	0	0	1	1	0
86.	7	7	1	0	0	0	0	0	0	0	1	1	0	0	0
87.	7	7	0	0	0	0	0	1	0	0	1	1	1	0	0
88.	7	7	0	0	0	1	0	0	0	0	0	0	1	1	0
89.	6	6	0	0	0	0	0	0	0	1	0	1	1	1	1
90.	6	6	0	0	0	0	0	0	0	1	1	1	0	1	0
91.	6	6	0	1	0	0	0	0	0	0	0	0	0	1	1
92.	6	6	0	1	0	0	0	0	0	0	1	1	1	1	0

Copyright by Mahidol University

No	Total	NOK	c01	c02	c03	c04	c05	c06	c07	c08	c09	c10	c11	c12	c13
93.	6	6	1	0	0	0	0	0	0	0	0	0	0	1	0
94.	5	0	0	0	0	0	0	0	0	1	0	0	0	0	1
95.	5	5	0	1	0	0	0	0	0	0	1	1	0	1	0
96.	5	5	0	0	1	0	0	0	0	0	0	0	0	1	0
97.	4	0	0	0	0	0	0	1	0	0	1	1	0	0	0
98.	4	0	0	0	0	0	0	0	0	0	0	1	0	1	1
99.	4	4	0	1	0	0	0	0	0	0	0	0	0	0	1
100.	4	4	0	0	0	0	0	0	0	0	1	1	0	0	1
101.	4	4	0	0	0	1	0	0	0	0	0	0	0	1	1
102.	4	4	0	1	0	0	0	0	0	0	0	1	1	0	0
103.	4	0	0	0	0	0	0	0	0	1	0	1	0	1	0
104.	3	3	0	0	0	0	0	1	0	0	0	1	0	1	0
105.	3	3	0	0	0	0	0	0	0	1	0	0	1	0	1
106.	3	0	0	0	0	1	0	0	0	0	0	0	0	1	0
107.	3	3	0	0	0	0	0	0	0	0	1	1	0	1	1
108.	3	3	0	1	0	0	0	0	0	0	0	0	1	1	1
109.	3	3	1	0	0	0	0	0	0	0	0	0	0	1	1
110.	3	0	0	1	0	0	0	0	0	0	0	1	0	0	0
111.	3	0	0	0	0	0	0	0	0	1	0	1	0	1	1
112.	3	0	0	1	0	0	0	0	0	0	0	0	1	0	0
113.	2	0	0	0	0	0	0	0	0	0	0	1	1	0	0
114.	2	0	0	0	0	0	0	0	0	1	0	0	1	1	0
115.	2	2	0	0	0	0	0	0	1	0	0	0	0	1	0
116.	2	2	0	0	0	0	0	0	0	1	1	1	0	1	1
117.	2	0	0	0	0	0	0	0	0	0	0	1	1	1	0
118.	2	2	0	0	0	0	1	0	0	0	0	0	1	0	0
119.	2	2	0	0	0	0	0	0	0	0	0	0	1	0	1
120.	2	0	0	1	0	0	0	0	0	0	1	1	1	1	0
121.	2	0	0	1	0	0	0	0	0	0	0	0	1	1	0
122.	2	2	0	1	0	0	0	0	0	0	1	1	0	0	1
123.	2	2	0	1	0	0	0	0	0	0	0	1	1	1	1
124.	2	2	1	0	0	0	0	0	0	0	0	0	1	1	0
125.	2	0	0	1	0	0	0	0	0	0	0	0	0	1	0
126.	2	0	0	0	0	0	0	0	0	0	1	1	0	1	0
127.	2	0	0	0	0	0	0	1	0	0	0	0	0	1	0
128.	2	2	0	1	0	0	0	0	0	0	0	1	1	1	0
129.	1	1	0	0	0	0	1	0	0	0	0	1	0	0	0
130.	1	1	0	0	1	0	0	0	0	0	0	0	0	1	1
131.	1	1	0	0	0	0	0	1	0	0	0	1	1	0	0
132.	1	1	0	0	1	0	0	0	0	0	0	1	1	1	0
133.	1	1	0	0	0	1	0	0	0	0	1	1	1	0	0
134.	1	1	0	1	0	0	0	0	0	0	1	1	1	1	1
135.	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0
136.	1	1	1	0	0	0	0	0	0	0	1	1	1	0	0
137.	1	0	0	0	0	0	0	0	0	1	1	1	0	1	0
138.	1	0	0	1	0	0	0	0	0	0	0	1	0	1	0
139.	1	1	0	0	0	0	1	0	0	0	0	0	0	1	0
140.	1	1	0	0	0	0	0	1	0	0	1	1	1	1	1
141.	1	1	0	0	1	0	0	0	0	0	0	1	0	1	0
142.	1	1	0	0	0	1	0	0	0	0	1	1	0	0	0
143.	1	1	0	0	0	0	0	1	0	0	0	0	1	1	1
144.	1	0	0	0	0	0	0	1	0	0	0	1	0	0	0

Copyright by Mahidol University

No	Total	NOK	c01	c02	c03	c04	c05	c06	c07	c08	c09	c10	c11	c12	c13
145.	1	1	0	0	0	1	0	0	0	0	0	1	1	1	0
146.	1	1	0	0	0	0	1	0	0	0	1	1	0	1	0
147.	1	1	0	0	0	1	0	0	0	0	0	1	1	1	1
148.	1	1	0	0	0	0	1	0	0	0	1	1	1	0	0
149.	1	0	0	1	0	0	0	0	0	0	1	1	0	1	0
150.	1	0	0	0	0	0	1	0	0	0	0	0	0	1	0
151.	1	1	0	0	0	0	0	0	0	1	1	1	1	1	1
152.	1	1	1	0	0	0	0	0	0	0	0	0	1	1	1
153.	1	1	0	0	0	0	0	1	0	0	0	1	1	1	0
154.	1	0	0	0	0	0	0	0	0	1	0	1	1	0	0

Table 2 The matrix of boolean variables of each criteria (with out C08)

No	Total	NOK	c01	c02	c03	c04	c05	c06	c07	c08	c09	c10	c11	c12	c13
1.	345	345	1	0	0	0	0	0	0	0	0	0	0	0	0
2.	343	343	0	0	0	0	1	0	0	0	0	0	0	0	0
3.	301	301	0	0	1	0	0	0	0	0	0	0	0	0	0
4.	281	281	0	0	0	0	0	0	1	0	0	0	0	0	0
5.	258	258	0	0	0	0	0	0	0	0	1	1	0	0	0
6.	234	234	0	0	0	0	0	0	0	0	0	1	0	0	0
7.	201	0	0	0	0	0	0	0	0	0	0	0	0	1	0
8.	189	189	0	0	0	0	0	0	0	0	0	0	0	1	0
9.	166	166	0	1	0	0	0	0	0	0	1	1	0	0	0
10.	161	161	0	0	0	0	0	0	0	0	0	0	1	0	0
11.	129	0	0	0	0	1	0	0	0	0	0	0	0	0	0
12.	114	114	0	0	0	1	0	0	0	0	0	0	0	0	0
13.	110	110	0	0	0	0	0	0	0	0	0	0	0	1	1
14.	108	0	0	0	0	0	0	0	1	0	0	0	0	0	0
15.	103	103	0	0	0	0	0	1	0	0	0	0	0	0	0
16.	95	95	0	0	0	0	0	0	0	0	0	0	1	1	0
17.	91	91	0	1	0	0	0	0	0	0	0	0	0	0	0
18.	85	85	0	0	0	1	0	0	0	0	0	1	0	0	0
19.	81	0	0	0	1	0	0	0	0	0	0	0	0	0	0
20.	77	0	0	0	0	0	0	0	0	0	0	1	0	0	0
21.	67	67	0	0	0	0	0	0	0	0	0	0	1	1	1
22.	61	61	0	0	0	0	0	1	0	0	1	1	0	0	0
23.	55	55	0	0	0	0	0	0	0	0	0	1	0	1	0
24.	45	45	0	0	0	0	0	0	0	0	0	1	1	0	0
25.	44	44	0	0	0	0	0	1	0	0	0	1	0	0	0
26.	42	0	0	0	0	0	0	0	0	0	0	0	0	1	1
27.	41	0	0	1	0	0	0	0	0	0	0	0	0	0	0
28.	39	0	0	0	0	0	1	0	0	0	0	0	0	0	0
29.	38	0	0	0	0	0	0	0	0	0	1	1	0	0	0
30.	37	37	0	1	0	0	0	0	0	0	0	1	0	0	0
31.	36	0	0	0	0	0	0	0	0	0	0	0	1	0	0
32.	35	35	0	0	0	0	0	0	0	0	0	1	0	1	1
33.	35	35	0	0	0	0	0	0	0	0	0	1	1	1	0
34.	33	33	0	0	0	0	0	0	0	0	1	1	0	1	0
35.	33	33	0	1	0	0	0	0	0	0	0	0	1	0	0

Copyright by Mahidol University

No	Total	NOK	c01	c02	c03	c04	c05	c06	c07	c08	c09	c10	c11	c12	c13
36.	31	0	1	0	0	0	0	0	0	0	0	0	0	0	0
37.	30	30	0	0	0	0	0	0	0	0	0	1	1	1	1
38.	28	28	0	0	0	0	0	1	0	0	0	0	0	1	0
39.	27	27	0	0	0	1	0	0	0	0	0	0	0	1	0
40.	26	26	0	0	0	0	0	1	0	0	0	0	1	0	0
41.	24	24	0	1	0	0	0	0	0	0	0	0	0	1	0
42.	22	0	0	0	0	0	0	0	0	0	0	0	1	1	1
43.	21	0	0	0	0	0	0	1	0	0	0	0	0	0	0
44.	19	19	0	0	0	0	0	0	0	0	0	0	0	0	1
45.	19	19	1	0	0	0	0	0	0	0	0	0	1	0	0
46.	18	18	0	0	0	1	0	0	0	0	0	0	1	0	0
47.	17	17	0	0	0	0	0	0	0	0	1	1	1	1	1
48.	16	0	0	1	0	0	0	0	0	0	1	1	0	0	0
49.	15	15	0	1	0	0	0	0	0	0	0	0	1	1	0
50.	14	0	0	0	0	1	0	0	0	0	0	1	0	0	0
51.	13	13	0	0	0	0	1	0	0	0	1	1	0	0	0
52.	13	13	0	0	1	0	0	0	0	0	0	0	1	0	0
53.	12	0	0	0	0	0	0	0	0	0	0	0	0	0	1
54.	12	12	0	0	0	0	0	0	0	0	1	1	1	0	0
55.	12	12	0	0	0	0	0	1	0	0	0	0	1	1	0
56.	11	0	0	0	0	0	0	0	0	0	0	1	0	1	0
57.	11	11	0	0	0	0	0	0	0	0	1	1	1	1	0
58.	11	11	0	0	0	1	0	0	0	0	0	1	1	0	0
59.	10	10	0	1	0	0	0	0	0	0	0	1	0	1	0
60.	9	0	0	0	0	0	0	0	0	0	0	0	1	1	0
61.	9	9	0	0	0	0	0	0	1	0	0	1	0	0	0
62.	9	9	0	0	0	0	0	1	0	0	1	1	0	1	0
63.	9	9	1	0	0	0	0	0	0	0	0	1	0	0	0
64.	8	8	0	0	0	1	0	0	0	0	0	1	0	1	0
65.	8	8	0	1	0	0	0	0	0	0	1	1	1	0	0
66.	7	0	0	0	0	0	0	0	0	0	0	1	0	1	1
67.	7	7	0	0	0	0	0	1	0	0	1	1	1	0	0
68.	7	7	0	0	0	0	0	1	0	0	1	1	1	1	0
69.	7	7	0	0	0	1	0	0	0	0	0	0	1	1	0
70.	7	7	0	0	1	0	0	0	0	0	0	1	0	0	0
71.	7	7	1	0	0	0	0	0	0	0	1	1	0	0	0
72.	6	6	0	1	0	0	0	0	0	0	0	0	0	1	1
73.	6	6	0	1	0	0	0	0	0	0	1	1	1	1	0
74.	6	6	1	0	0	0	0	0	0	0	0	0	0	1	0
75.	5	5	0	0	0	0	0	0	0	0	0	0	1	0	1
76.	5	5	0	0	0	0	0	0	0	0	1	1	0	1	1
77.	5	5	0	0	1	0	0	0	0	0	0	0	0	1	0
78.	5	5	0	1	0	0	0	0	0	0	1	1	0	1	0
79.	4	0	0	0	0	0	0	1	0	0	1	1	0	0	0
80.	4	4	0	0	0	0	0	0	0	0	1	1	0	0	1
81.	4	4	0	0	0	1	0	0	0	0	0	0	0	1	1
82.	4	4	0	1	0	0	0	0	0	0	0	0	0	0	1
83.	4	4	0	1	0	0	0	0	0	0	0	1	1	0	0
84.	3	0	0	0	0	0	0	0	0	0	0	1	1	0	0
85.	3	0	0	0	0	0	0	0	0	0	1	1	0	1	0
86.	3	0	0	0	0	1	0	0	0	0	0	0	0	1	0
87.	3	0	0	1	0	0	0	0	0	0	0	0	1	0	0

Copyright by Mahidol University

No	Total	NOK	c01	c02	c03	c04	c05	c06	c07	c08	c09	c10	c11	c12	c13
88.	3	0	0	1	0	0	0	0	0	0	0	1	0	0	0
89.	3	3	0	0	0	0	0	1	0	0	0	1	0	1	0
90.	3	3	0	1	0	0	0	0	0	0	0	0	1	1	1
91.	3	3	1	0	0	0	0	0	0	0	0	0	0	1	1
92.	2	0	0	0	0	0	0	0	0	0	0	1	1	1	0
93.	2	0	0	0	0	0	0	1	0	0	0	0	0	1	0
94.	2	0	0	1	0	0	0	0	0	0	0	0	0	1	0
95.	2	0	0	1	0	0	0	0	0	0	0	0	1	1	0
96.	2	0	0	1	0	0	0	0	0	0	1	1	1	1	0
97.	2	2	0	0	0	0	0	0	1	0	0	0	0	1	0
98.	2	2	0	0	0	0	1	0	0	0	0	0	1	0	0
99.	2	2	0	1	0	0	0	0	0	0	0	1	1	1	0
100.	2	2	0	1	0	0	0	0	0	0	0	1	1	1	1
101.	2	2	0	1	0	0	0	0	0	0	1	1	0	0	1
102.	2	2	1	0	0	0	0	0	0	0	0	0	1	1	0
103.	1	0	0	0	0	0	0	1	0	0	0	1	0	0	0
104.	1	0	0	0	0	0	1	0	0	0	0	0	0	1	0
105.	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0
106.	1	0	0	1	0	0	0	0	0	0	0	1	0	1	0
107.	1	0	0	1	0	0	0	0	0	0	1	1	0	1	0
108.	1	1	0	0	0	0	0	1	0	0	0	0	1	1	1
109.	1	1	0	0	0	0	0	1	0	0	0	1	1	0	0
110.	1	1	0	0	0	0	0	1	0	0	0	1	1	1	0
111.	1	1	0	0	0	0	0	1	0	0	1	1	1	1	1
112.	1	1	0	0	0	0	1	0	0	0	0	0	0	1	0
113.	1	1	0	0	0	0	1	0	0	0	0	1	0	0	0
114.	1	1	0	0	0	0	1	0	0	0	1	1	0	1	0
115.	1	1	0	0	0	0	1	0	0	0	1	1	1	0	0
116.	1	1	0	0	0	1	0	0	0	0	0	1	1	1	0
117.	1	1	0	0	0	1	0	0	0	0	0	1	1	1	1
118.	1	1	0	0	0	1	0	0	0	0	1	1	0	0	0
119.	1	1	0	0	0	1	0	0	0	0	1	1	1	0	0
120.	1	1	0	0	1	0	0	0	0	0	0	0	0	1	1
121.	1	1	0	0	1	0	0	0	0	0	0	1	0	1	0
122.	1	1	0	0	1	0	0	0	0	0	0	1	1	1	0
123.	1	1	0	1	0	0	0	0	0	0	1	1	1	1	1
124.	1	1	1	0	0	0	0	0	0	0	0	0	1	1	1
125.	1	1	1	0	0	0	0	0	0	0	1	1	1	0	0

Copyright by Mahidol University

BIOGRAPHY

NAME	Miss Patanakul Sathapornrungkij
DATE OF BIRTH	17 July 1975
PLACE OF BIRTH	Bangkok, Thailand
INSTITUTIONS ATTENDED	Assumption University, 1992-1996: Bachelor of Science (Computer Science) Mahidol University, 1999-2004: Master of Science (Computer Science)
POSITION&OFFICE	1999-2003, Merlin's solution International Ltd. Rama IX Rd, HuayKwang Bangkok, Thailand. Position: System Engineer 1996-1999, Foryou Infosys Ltd. Klongchan, Bangkapi, Bangkok, Thailand. Position: System Engineer

Copyright by Mahidol University