# SEMANTIC-BASED QUESTION ANSWERING FRAMEWORK FOR FUZZY FACTOID ANSWER FROM THAI TEXTS

**AUTHAPON KONGWAN**



**DOCTOR OF PHILOSOPHY**

**UNIVERSITI UTARA MALAYSIA**

**2024**

Awang Had Salleh
Graduate School
of Arts And Sciences

Universiti Utara Malaysia

## PERAKUAN KERJA TESIS / DISERTASI
*(Certification of thesis / dissertation)*

Kami, yang bertandatangan, memperakukan bahawa
*(We, the undersigned, certify that)*

**AUTHAPON KONGWAN**

calon untuk Ijazah
*(candidate for the degree of)* **PhD**

telah mengemukakan tesis / disertasi yang bertajuk:
*(has presented his/her thesis / dissertation of the following title):*

**"SEMANTIC-BASED QUESTION ANSWERING FRAMEWORK FOR FUZZY FACTOID ANSWER FROM THAI TEXTS"**

seperti yang tercatat di muka surat tajuk dan kulit tesis / disertasi.
*(as it appears on the title page and front cover of the thesis / dissertation).*

Bahawa tesis/disertasi tersebut boleh diterima dari segi bentuk serta kandungan dan meliputi bidang ilmu dengan memuaskan, sebagaimana yang ditunjukkan oleh calon dalam ujian lisan yang diadakan pada : ***06 November 2023.***
*That the said thesis/dissertation is acceptable in form and content and displays a satisfactory knowledge of the field of study as demonstrated by the candidate through an oral examination held on:*
***06 November 2023.***

| | | |
|---|---|---|
| Pengerusi Viva: *(Chairman for VIVA)* | **Prof. Dr. Huda Haji Ibrahim** | Tandatangan *(Signature)* |
| Pemeriksa Luar: *(External Examiner)* | **Assoc. Prof. Dr. Rabiah Abdul Kadir** | Tandatangan *(Signature)* |
| Pemeriksa Dalam: *(Internal Examiner)* | **Ts. Dr. Juhaida Abu Bakar** | Tandatangan *(Signature)* |
| Nama Penyelia/Penyelia-penyelia: *(Name of Supervisor/Supervisors)* | **Assoc. Prof. Dr. Siti Sakira Kamaruddin** | Tandatangan *(Signature)* |
| Nama Penyelia/Penyelia-penyelia: *(Name of Supervisor/Supervisors)* | **Dr. Farzana Kabir Ahmad** | Tandatangan *(Signature)* |

Tarikh:
*(Date)* ***06 November 2023***

## Permission to Use

In presenting this thesis in fulfilment of the requirements for a postgraduate degree from Universiti Utara Malaysia, I agree that the Universiti Library may make it freely available for inspection. I further agree that permission for the copying of this thesis in any manner, in whole or in part, for scholarly purpose may be granted by my supervisor(s) or, in their absence, by the Dean of Awang Had Salleh Graduate School of Arts and Sciences. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to Universiti Utara Malaysia for any scholarly use which may be made of any material from my thesis.

Requests for permission to copy or to make other use of materials in this thesis, in whole or in part, should be addressed to:



Dean of Awang Had Salleh Graduate School of Arts and Sciences

UUM College of Arts and Sciences

Universiti Utara Malaysia

06010 UUM Sintok

# Abstrak

Teks adalah sumber pengetahuan manusia yang penting. Sistem menjawab soalan boleh mendapatkan semula fakta daripada sumber pengetahuan dan memberikan jawapan kepada pengguna. Menterjemah teks ke pangkalan pengetahuan adalah tugas yang sangat mencabar dan proses yang rumit. Teks Thai merupakan satu bentuk aliran aksara yang ditulis secara berterusan tanpa sebarang tanda baca atau penanda untuk memisahkan setiap perkataan dan setiap ayat dalam perenggan. Penyelidikan ini bertujuan untuk membangunkan rangka kerja menjawab soalan berasaskan semantik yang boleh mengendalikan fakta kabur dan menyasarkan sumber pengetahuan kepada teks Thai. Dalam membina sistem menjawab soalan Thai, analisis morfologi Thai merupakan komponen penting untuk memproses teks Thai. Peleraian elipsis dan anafora dalam teks Thai juga merupakan proses yang diperlukan untuk membina fakta lengkap daripada teks Thai. Penghurai semantik Thai ialah komponen teras untuk membina pangkalan pengetahuan dengan mengekstrak fakta daripada teks Thai ke dalam struktur bingkai semantik. Metodologi kajian ini terbahagi kepada 4 langkah. Pertama ialah membina analisis morfologi Thai yang tepat: segmentasi perkataan Thai dan segmentasi EDU Thai. Kedua adalah untuk membangunkan resolusi elipsis dan anafora untuk teks Thai untuk mencapai matlamat mencipta fakta lengkap dalam segmentasi EDU Thai. Ketiga ialah membangunkan penghurai semantik untuk membina pangkalan pengetahuan yang mengubah teks Thai menjadi perwakilan bingkai semantik. Keempat, membangunkan pengekstrakan jawapan untuk sistem menjawab soalan dengan padanan kabur untuk mengendalikan factoid kabur. Daripada kesemua jujukan proses, sistem menjawab soalan berasaskan semantik mencapai kejituan dan dapatan semula yang tinggi iaitu 0.9892 dan 0.9484. Kesimpulannya, resolusi anafora dan elipsis adalah penting untuk mencapai pembinaan semantik yang tepat, manakala padanan kabur dengan ketara meningkatkan ingatan pengekstrakan jawapan. Bersama-sama, komponen ini penting untuk membina sistem menjawab soalan "Apa" dan "Berapa" yang mantap.

**Kata Kunci:** Semantik, Sistem menjawab soalan, Resolusi anafora, Peruasan kata, Kabur.

# Abstract

Text is an important human knowledge source. The question-answering system can retrieve the fact from the source of knowledge and provide the answer to the user. Translating the text to the knowledge base is a very challenge task and complicated process. Thai text can be a form of character stream written continuously without any punctuation or marker to separate each word and each sentence in a paragraph. This research is aim to develop a semantic base question-answering framework that can handle the fuzzy factoid and target the knowledge source to Thai text. In building a Thai question-answering system, Thai morphological analysis is an important component to process Thai text. Ellipsis and anaphora resolution in Thai text is also the needed process for constructing the complete fact from Thai text. Thai semantic parser is the core component to construct the knowledge base by extracting the fact from Thai text into the semantic frame structure. The methodology of this research is divided into 4 steps. First is building the accurate Thai morphological analysis: Thai word segmentation and Thai EDU segmentation. The second is to develop the ellipsis and anaphora resolution for Thai text to achieve the goal that is creating the complete fact in Thai EDU segmentation. The third is to develop the semantic parser to build the knowledge base that transforms the Thai text into a semantic frame representation. Forth is developed the answer extraction for the question answering system with fuzzy matching to handle the fuzzy factoid. From the pipeline of the processes, the semantic-based question answering system performs high precision and recall to 0.9892 and 0.9484. In conclusion, anaphora and ellipsis resolution are crucial for achieving precise semantic construction, while fuzzy matching significantly enhances answer extraction recall. Together, these components are essential for building robust "What" and "How many" question answering systems.





**Keywords:** Semantic, Question answering system, Anaphora resolution, Word segmentation, Fuzzy.

# Acknowledgement

I would like to express my deepest appreciation to my supervisor Assoc. Prof. Dr. Siti Sakira Binti Kamaruddin and co-supervisor Dr. Farzana Binti Kabir Ahmad for giving me helpful guidance and kindness throughout this research. Their valuable guidance, feedback, motivation, and kindness gave me more inspiration and encouragement to finish my work.

I am very much thankful to my father, my friends, and my colleagues at the Engineering Faculty, Rajamangala University of Technology for their love and support to help me complete the research.

Finally, I would like to thank all the people who have supported me directly or indirectly to make my work successful.

# Table of Contents

Permission to Use ..... ii

Abstrak ..... iii

Abstract ..... iv

Acknowledgement ..... v

Table of Contents ..... vi

List of Tables ..... xi

List of Figures ..... xiii

List of Abbreviations ..... xvii

List of Appendices ..... xviii

CHAPTER ONE INTRODUCTION ..... 1

1.1 Overview ..... 1

1.2 Problem Statement ..... 4

1.3 Research Questions ..... 9

1.4 Research Objectives ..... 10

1.5 Scope of Research ..... 10

1.6 Significance of Research ..... 11

CHAPTER TWO LITERATURE REVIEWS ..... 12

2.1 Overview ..... 12

2.2 Morphological Analysis in Thai text ..... 16

2.2.1 Corpus Resources ..... 17

2.2.2 Word Segmentation ..... 19

2.2.3 Named Entity Extraction ..... 23

2.3 Syntactic Analysis ..... 26
2.3.1 Dependency Parser ..... 26
2.3.2 Context-Free Grammar Parser ..... 28
2.3.3 Syntactic Analysis in Thai text ..... 32
2.4 Ellipsis and Anaphora Resolution ..... 34
2.5 Semantic Analysis ..... 37
2.5.1 Syntactic-Driven Approach ..... 37
2.5.2 Semantic-Driven Approach ..... 38
2.6 Knowledge Representation ..... 38
2.6.1 Resource Description Framework (RDF) ..... 39
2.6.2 Universal Networking Language (UNL) ..... 43
2.6.3 Natural Language Annotation ..... 47
2.6.4 Semantic Frame Representation ..... 48
2.7 Question Answering System ..... 50
2.7.1 Content Based ..... 57
2.7.2 Data Source Based ..... 58
2.7.3 Language Paradigm Based ..... 59
2.7.4 Question Type Based ..... 60
2.7.5 Approach Based on Question Analysis ..... 61
2.7.6 Technique Based ..... 62
2.8 Question Classification ..... 66
2.9 Fuzzy Factoid ..... 70
2.10 Summary ..... 71
CHAPTER THREERESEARCH METHODOLOGY ..... 73

3.1 Stage of Development ..... 73
3.2 Corpus Preparation for Morphological Process ..... 75
3.3 Morphological Analysis ..... 80
3.3.1 Thai Word Segmentation and POS Tagging ..... 81
3.3.2 Named Entities Identification ..... 89
3.3.3 Thai EDU Segmentation ..... 91
3.3.3.1 Issues in Thai EDU Segmentation ..... 91
3.3.3.2 Definition of Thai EDU ..... 93
3.3.3.3 EDU Segmentation by Clue Markers ..... 94
3.3.3.4 Shallow Parser ..... 94
3.3.3.5 EDU Segmentation by Syntactic Pattern ..... 97
3.3.3.6 EDU Reconstruction by Rule-Based ..... 98
3.4 Ellipsis and Anaphora Resolution ..... 99
3.4.1 Anaphora in Thai Texts ..... 100
3.4.1.1 Anaphora Types ..... 100
3.4.1.2 Referential and Non-Referential Anaphora ..... 105
3.4.2 The Resolution ..... 107
3.4.2.1 Corpus Preparation for Anaphora Resolution ..... 109
3.4.2.2 Anaphora Determiner Algorithm ..... 109
3.4.2.3 Resolution for Non-Referential Anaphora ..... 111
3.4.2.4 Resolution for Referential Anaphora ..... 113
3.5 Semantic Parser ..... 118
3.5.1 Word Sense Disambiguation ..... 119
3.5.1.1 Corpus Tagging ..... 119

3.5.1.2 Semantic Disambiguation ..... 121
3.5.2 Semantic Frame Construction ..... 124
3.5.2.1 Frame-based Knowledge Representation ..... 124
3.5.2.2 Semantic Frame Construction Rules ..... 131
3.5.2.3 Knowledge Base ..... 135
3.6 Answer Extraction ..... 136
3.6.1 Frame Matching ..... 137
3.6.2 Fuzzy Matching ..... 140
3.6.2.1 Membership Function in Measurement Value ..... 140
3.6.2.2 Membership Function in Semantic ID ..... 141
3.6.2.3 Confidence Value ..... 142
3.7 Evaluation ..... 142
3.8 Summary ..... 143
CHAPTER FOUR EXPERIMENTAL RESULTS ..... 144
4.1 Word Segmentation and POS Tagging ..... 144
4.2 EDU Segmentation ..... 146
4.3 Ellipsis and Anaphora Resolution ..... 148
4.4 Word Sense Disambiguation ..... 150
4.5 Answer Extraction ..... 151
4.6 Explanation with The Other Thai QAS ..... 156
4.7 Summary ..... 158
CHAPTER FIVE CONCLUSION ..... 159
5.1 Research Summary ..... 159
5.1.1 Thai EDU Segmentation ..... 159

5.1.2 Ellipsis and Anaphora Resolution ..... 160
5.1.3 Semantic Parser ..... 161
5.1.4 Answer Extraction ..... 162
5.2 Limitations ..... 162
5.3 Research Contribution ..... 163
5.4 Future Work ..... 164
REFERENCES ..... 165

# List of Tables

Table 2.1 Exhaustive list of Thai corpus resources (Arreerard et al., 2022) ..... 18

Table 2.2 Summary of word segmentation research..... 22

Table 2.3 Summary of Thai named entity extraction research ..... 25

Table 2.4 Classification of QAS based on techniques (Mishra & Jain, 2016) .. 64

Table 2.5 The coarse and fine-grained question categories ..... 68

Table 3.1 Examples of symbol tags ..... 79

Table 3.2 Examples of the POS tagging label ..... 82

Table 3.3 The example of dictionary ..... 84

Table 3.4 POS pattern rules ..... 86

Table 3.5 Examples of the POS pattern for word correction ..... 88

Table 3.6 Examples of the named entity matching template ..... 90

Table 3.7 Examples of a dictionary for the shallow parser..... 96

Table 3.8 Shallow parser feature pattern ..... 96

Table 3.9 Examples of the syntactic pattern ..... 97

Table 3.10 Examples of the EDU reconstruction rule ..... 98

Table 3.11 Example of the features of non-referential anaphora in the database..... 112

Table 3.12 Kinds of feature values for non-referential anaphora ..... 113

Table 3.13 Example of the features of referential anaphora in the database ..... 114

Table 3.14 First group of feature values on the anaphora side ..... 115

Table 3.15 Second group of feature values on the reference side ..... 116

Table 3.16 Third group of feature values on both sides of anaphora and

reference ..... 117

Table 3.17 The examples of semantic dictionary ..... 121

Table 3.18 The feature types for the semantic concept training ..... 122

Table 3.19 The examples of the feature that are extracted from the tagged corpus ..... 123

Table 3.20 The question words and the semantic ID ..... 138

Table 4.1 Results of the anaphora resolution ..... 149

Table 4.2 The examples of the predefined question ..... 152

Table 4.3 The example of the question and the answers that extract from answer extraction with and without the anaphora resolution ..... 153

Table 4.4 The example of the question and the answers that extract with and without fuzzy matching ..... 154

Table 4.5 The explanation of the Thai question answering system ..... 157

# List of Figures

Figure 2.1 Overview of QAS in data-intensive approach ..... 13

Figure 2.2 Overview of QAS in knowledge-intensive approach ..... 14

Figure 2.3 Example of dependency structure ..... 27

Figure 2.4 Algorithm for finding maximum spanning trees in directed graphs (McDonald et al., 2005) ..... 28

Figure 2.5 Example of Context-Free Grammar ..... 29

Figure 2.6 Pseudocode of CYK algorithm (Lange & Leiß, 2009) ..... 30

Figure 2.7 Pseudocode of Earley algorithm (Jurafsky & Martin, 2009) ..... 31

Figure 2.8 Grammar flow graph (Pingali & Bilardi, 2015) ..... 32

Figure 2.9 Algorithm to find the best dependency structure (Tongchim et al., 2008) ..... 33

Figure 2.10 Example of zero anaphora in the Thai language ..... 34

Figure 2.11 Example of textual ellipsis in the Thai language ..... 35

Figure 2.12 Ellipsis resolution algorithm (Kongwan & Kawtrakul, 2005) ..... 36

Figure 2.13 Example of RDF graph (Wikipedia, 2007) ..... 39

Figure 2.14 Example of N-Triples syntax (Wikipedia, 2007) ..... 40

Figure 2.15 Example of Turtle syntax (Wikipedia, 2007) ..... 40

Figure 2.16 Example of XML syntax (Wikipedia, 2007) ..... 41

Figure 2.17 System architecture of QAST system (Jitkrittum et al., 2009) ..... 42

Figure 2.18 The proposed system of Thailand tourism information system (Kongthon et al., 2011) ..... 43

Figure 2.19 Example of UNL (Wikipedia, 2005) ..... 44

Figure 2.20 Layout of MT module (Boguslavsky et al., 2000) ..... 45
Figure 2.21 Step of tools to convert from native language to UNL graphical representation (Ripon et al., 2014)..... 46
Figure 2.22 The example of T-expression (Katz, 1997)..... 47
Figure 2.23 The example of frame semantics and its heritance (Lowe et al., 1997) ..... 48
Figure 2.24 The related to the CAUSE_TO_MAKE_NOISE frame and MAKE_NOISE frame (Das et al., 2010) ..... 49
Figure 2.25 CubeQA pipeline (Konrad et al., 2016)..... 54
Figure 2.26 High-level components diagram of the vocabulary independent query approach and distributional inverted index structure (Freitas & Curry, 2014)..... 55
Figure 2.27 QuASE framework (Sun et al., 2015) ..... 56
Figure 2.28 WabiQA system architecture (Noraset et al., 2021)..... 56
Figure 2.29 Classification of QAS based on knowledge and data intensive View..... 65
Figure 2.30 The hierarchical classifier (Li & Roth, 2006) ..... 67
Figure 2.31 The proposed architecture that integrated genetic algorithm and machine learning approach for question classification in English-Chinese cross-language question answering (Day et al., 2007) ..... 69
Figure 3.1 Research design stages ..... 74
Figure 3.2 Example of the Thai Wikipedia page ..... 76
Figure 3.3 Example of HTML source..... 77

Figure 3.4 Overview of corpus preparation process ..... 78

Figure 3.5 Example of a clean corpus..... 79

Figure 3.6 Overview of morphological analysis..... 81

Figure 3.7 Example of a tagged corpus ..... 83

Figure 3.8 Processes of Thai word segmentation and POS tagging ..... 89

Figure 3.9 Example of the corpus with a manual named entity tagging..... 90

Figure 3.10 Example of a corpus with a phrase tagged ..... 95

Figure 3.11 Zero anaphora..... 100

Figure 3.12 Embedded relative clause ..... 101

Figure 3.13 Zero anaphora on the embedded relative clause after EDU segmentation ..... 102

Figure 3.14 Pronominal anaphora..... 102

Figure 3.15 Nominal anaphora ..... 103

Figure 3.16 Nominal anaphora on the same head word..... 104

Figure 3.17 Ellipsis of the owner..... 104

Figure 3.18 Non-referential in zero anaphora..... 106

Figure 3.19 Non-referential in pronominal anaphora ..... 106

Figure 3.20 Non-referential in nominal anaphora..... 107

Figure 3.21 Overview of the anaphora resolution processes ..... 108

Figure 3.22 Anaphora tagging in the corpus..... 109

Figure 3.23 Anaphora determiner algorithm ..... 110

Figure 3.24 The example of transforming from syntactic phrase structure with semantic ID to semantic frame structure ..... 118

Figure 3.25 The example of tagged corpus by the semantic ID ..... 120

Figure 3.26 The structure of the action frame..... 126
Figure 3.27 The structure of the object frame..... 127
Figure 3.28 The structure of the modifier frame..... 128
Figure 3.29 The structure of the attribute frame ..... 129
Figure 3.30 The structure of the time frame ..... 130
Figure 3.31 The structure of the determiner frame ..... 130
Figure 3.32 The Example of semantic frame construction rule..... 135
Figure 3.33 The example of knowledge base ..... 136
Figure 3.34 The process of answer extraction from question to answer ..... 137
Figure 3.35 The example of frame matching for what question..... 139
Figure 3.36 The example of frame matching for how many question..... 139
Figure 4.1 Bar chart of the results of word segmentation..... 146
Figure 4.2 Bar chart of the results of EDU segmentation..... 147
Figure 4.3 Bar chart of the results of the answer extraction ..... 155

# List of Abbreviations

| | |
|---|---|
| **QAS** | Question Answering System |
| **HCI** | Human-Computer Interface |
| **WWW** | World Wide Web |
| **HTML** | HyperText Markup Language |
| **AI** | Artificial Intelligence |
| **RDF** | Resource Description Framework |
| **NLP** | Natural Language Processing |
| **CRF** | Conditional Random Field |
| **XML** | eXtensible Markup Language |
| **UNL** | Universal Networking Language |
| **SQL** | Structured Query Language |
| **SVM** | Support Vector Machine |
| **GA** | Genetic Algorithm |
| **EDU** | Elementary Discourse Units |
| **POS** | Part of Speech |
| **CFG** | Context-Free Grammar |
| **CNF** | Chomsky Normal Form |
| **CCG** | Combinatory Categorial Grammar |

# List of Appendices

Appendix A Thai Part of Speech ..... 190

Appendix B Question Answer Pair Validation Form ..... 198

Appendix C List of Expert Validator ..... 206

# CHAPTER ONE

# INTRODUCTION

This chapter describes the brief information of this research that includes an overview, problem statement, research questions, research objectives, the scope of research, the significance of the research, and organization of the thesis.

## 1.1 Overview

Question answering system (QAS) is the research field that concerns with enabling the machine to give precise answers to the users who pose questions in the form of natural language. The purpose of QAS is to provide a natural interface (human language) for human-computer interaction (HCI) that helps humans to solve the problem in terms of the question and then gives the most satisfactory answer. Question answering system research attempts to deal with various kinds of questions including: fact, list, definition, reason, procedure, etc.



Early research in question answering systems is based on a closed corpus and closed domain approach. for example, BASEBALL (Green et al., 1961); the database-oriented question answering system is developed to answer the questions about results, locations, and dates of a baseball game. However, this system works on the question that presents in a known domain and corpus and is hard to scale. Developing the open domain question answering system is a challenge in this area. Until today, the open domain question answering system is a high research area and some systems are proposed for example: MIT START (Katz, 1997; Katz & Lin, 2002; Katz et al., 2006),

IBM Watson (Ferrucci et al., 2010), Dr.QA (Chen et al., 2017), BERTserini (Yang et al., 2019) and WabiQA (Noraset et al., 2021). These systems are good to respond to the explicit factoid answer, however, there are the drawbacks in answer that is obscure and inferred.

The question classification is the first phrase for a user to start working with the question answering system. This phase is responsible to classify the kind of expected answer and construct information for the answer extraction process (Li & Roth, 2002). To improve question classification to be more efficient, knowledge resources and language processing are needed to construct the class hierarchy (Bakhtyar et al., 2011) as background knowledge for the answer extraction process.

Nevertheless, the source of the answer is also an important part of the question answering system. Besides text, World Wide Web (WWW) is an essential source of question answering system for extracting the answer (Kwok et al., 2001). QAST (Jitkrittum et al., 2009) is an example of QAS systems that have been proposed using Wikipedia as a source of the answer. However, WWW is an unstructured text that needs huge preprocessing to prune out unused information before the knowledge extraction process.

For language-specific question answering systems in the particular Thai language, the morphological analysis is addressing the different problems in the English language. In Thai text, there are no whitespace or some clues between the word which is a word boundary problem. Word segmentation (Sutantayawalee & Supnithi, 2016) is needed

to separate each word into sentences before the knowledge extraction process. Moreover, anaphora resolution and name entity extraction are also crucial issues in Thai morphological analysis.

Semantic-based analysis (Kongthon et al., 2011; Narayanan & Harabagiu, 2004; Shen & Lapata, 2007) is used to develop and improve question answering systems to get a more reliable answer. Question answering system with Semantic enrichment (Sun et al., 2015) is good for open domain question answering system. Knowledge-intensive question answering systems (Bao et al., 2014; Fader et al., 2014; Yih et al., 2015) tend to the direction of question answering system research for open domain question answering systems. The knowledge base is constructed for finding the answer with good accuracy and performance. Explicit factoid answers can be extracted directly from a knowledge source, however, some factoid answers can be taken by inference from existing knowledge by using semantic-based.



Fuzziness is the new interesting topic in Natural Language Processing (NLP) research (Urrutià et al., 2017). The factoid can be formed in linguistics that is fuzzy, such as nouns, and adjectives (Novák, 2017b). Fuzzy logic will be an important tool to represent and process the semantics in some applications. Moreover, fuzzy logic is used in some processes of question answering systems such as answer ranking (Pota et al., 2017) and ontology similarity (Rani et al., 2014).

This study aims to develop the semantic based question answering system that can deal with fuzzy factoid answer and to develop the knowledge extraction module that can extract knowledge from Thai textual sources automatically.

## 1.2 Problem Statement

A question answering system (QAS) is a useful tool for a human to find out the resolution to certain problems. An intensive knowledge question answering system consists of two parts: the knowledge extraction module and the answering module (Pereira et al., 2022). The knowledge extraction module is a crucial part to construct knowledge from various sources into knowledge representation. There are three issues that could be identified in knowledge extraction: morphological analysis, knowledge representation, and semantic analysis (Antoniou & Bassiliades, 2022). The answering module is the part for answering the user's question from knowledge-based that is constructed by the knowledge extraction module. The answer can be the attribute of the object that can be the measured value and word value. The value can be in the form of a fuzzy word or range number. To gain more recall, the fuzzy value could be handled. Then, the interesting issue in the answering module is how to deal with the fuzzy factoid answer.

Morphological analysis is the group of processes to analyze the structure of language. For Thai textual data, there are some issues that have to be solved: word boundary, sentence boundary, and named entity extraction (Tapsai et al., 2019). Morphological analysis is important to construct precise knowledge. The precision of morphological analysis will directly affect the precision of knowledge.

In languages such as Thai, Chinese, and Lao, there is no space or clue word used for separating each word. Word segmentation is the first process for finding the boundary of words in every sentence from textual data. The accuracy of word segmentation is an important factor obtain precise knowledge. However, there is some successful Thai word segmentation research (Kongyoung et al., 2015; Kruengkrai et al., 2006; Sudprasert & Kawtrakul, 2003) that can be considered in the morphological analysis. The Condition Random Fields (Lafferty et al., 2001) together with the dictionary base is performed with good results in Thai word segmentation. However, it still needs to improve its performance to produce the accurate answer to the question answering system.

Sentence boundary is also an important part of the Thai morphological process. The discourse clue and syntactic information (Chareonsuk et al., 2005) is used to find the boundary of discourse unit in Thai sentence. However, improvement is needed for practical development in the question answering system.

Named entities are essential for question answering systems to give the factoid answer that involved with object: person, animal, organization, location, etc. The accuracy of the named entity extraction process affects the accuracy of the question answering system directly. The wrong answer, that is a named entity, can be given as a result of using an inaccurate name entities module. Anyway, there are some researches (Chanlekha & Kawtrakul, 2004; Charoenpornsawat et al., 1998; Sutheebanjard & Premchaiswadi, 2009) that are successful in the broad domain that can be considered.

Some machine learning such as Winnow and the Maximum Entropy Model is applied to identify the boundary of the Thai named entity with good results. However, the less complex algorithm such as the template matching technique can be a good option for Thai named entity identification.

Anaphora & ellipsis are interesting phenomena in language that are an essential part of the discourse process. The factoid answer may be in form of anaphora or ellipsis which is able to identify the answer's object. In Thai textual sources, there are three major anaphora types that are concerned: Nominal, Pronominal, and Zero anaphora. The anaphora and ellipsis resolution is needed to identify the object in a sentence with precision to build precise knowledge. There is the resolution for zero anaphora resolution (Aroonmanakun, 2000) in this research area. A centering theory (Grosz et al., 1995) is applied to resolve the zero anaphora in Thai text. The result is acceptable in their experiments. There is still no experiment on the other type of anaphora. The ellipsis of the owner in Thai text also needs to resolve to complete the information in the sentence. Then, the completion of anaphora and ellipsis is needed to complete the knowledge extraction process.

Some Thai question answering system is developed and handled in some morphological analysis. QAST (Jitkrittum et al., 2009) is the Thai Question Answering System that uses the knowledge source from Thai Wikipedia. The system constructed the knowledge from Wikipedia's infoboxes to Resource Description Framework (RDF) (Lassila & Swick, 1999) and also used the textual search index as a subsystem. This system not used complex NLP techniques to identify the answer from the knowledge

and then the accuracy is the issue. The Ontology-based question answering system in tourism (Kongthon et al., 2011) is proposed by using the predefined ontology in the tourism domain. It utilized pattern matching to analyze the question and generate the query to find the answer from ontology. This system has not constructed the ontology automatically which is hard to scale and it is domain specific.

To achieve the high-performance result in the Thai morphological process, the development of the Thai corpus is a crucial part of Thai NLP research. Thai corpus contains the necessary information depends on its purpose and domain. Orchid corpus (Sornlertlamvanich et al., 1999) is the very first corpus that is developed to experiment the process of tokenization and word segmentation. This corpus contains the information that is only needed for the word segmentation process. The parallel corpus such as TALPCo (Nomoto, 2019) consists of the Japanese language and its translation is developed for its own purpose. To accomplish the objective in the study, the corpus development is needed for serve the purpose in each process of study.

Semantic-based analysis (Kaisser & Webber, 2007; Narayanan & Harabagiu, 2004; Shen & Lapata, 2007) is used to improve the precision of the question answering system by utilizing synonym and thematic roles. Each sentence is processed to construct the semantic structure in form of predicate logic or frame to represent the semantic concepts and semantic relations. Explicit factoid answer is possible to identify by the semantic-based question answering system. However, factoid answer that is not explicit and obscure is difficult to identify without inference and fuzzy theory. To develop the question answering system, that gives the precise answer, complex NLP technique,

such as discourse analysis and semantic analysis, is needed to capture the semantics from the textual source and also is able to scale the knowledge-based automatically. The knowledge can be extracted from various textual sources. The renewed knowledge could be reconstructed by merging the existing knowledge with new knowledge. The definition answer is the example that needs to reconstruct new semantics each time that new knowledge emerges. Moreover, some knowledge could be inferred from the existing knowledge. The problem is how to produce precise knowledge from the existing knowledge with new knowledge.

Knowledge representation is crucial in the question answering system. In most question answering systems, the answer is explicit in data sources, but there are some data that are obscure values such as numeric range data and linguistic symbol data that are fuzzy data (Novák, 2017b; Urrutià et al., 2017). Thus, the semantics that is extracted from natural language textual source normally can be obscure and contain fuzzy data (Zadeh, 1965, 1988, 1999). The knowledge representation should be able to compute the value with fuzzy to extract the most precise answer to the user. However, fuzzy logic is used for some purposes of question answering systems such as answer ranking (Pota et al., 2017) and ontology similarity (Rani et al., 2014). Semantic networks (Sowa, 2014) are commonly used to represent knowledge in Artificial Intelligent (AI) research. However, the variable that is used in logic is hard to formulate in fuzzy functions. The frame-based representation (Andrade et al., 2014; Tettamanzi, 2003; Zadeh, 1989) is more suitable to represent knowledge that contains fuzzy data. The knowledge that is extracted from natural language always contains obscure data and the knowledge representation could be able to compute the fuzzy data.

In a summary, Thai morphological analysis is highly researched and we can consider repeating the development. The major problem in Thai morphological analysis for the question answering system is that there is still no complete anaphora & ellipsis resolution. To build the precision knowledge extraction process, the ellipsis and all of the anaphora resolutions are needed. The problem to create the precise answer is that the precise answer can not be constructed without using semantic reconstruction, thus, knowledge merging and inference are needed in the knowledge reconstruction process. Lastly, semantics that is constructed from natural language sources can contain some values with words that make the value obscure (Novák, 2017b).



### 1.3 Research Questions

The main question of this research is about how to develop the question answering for the Thai language with automatic analysis and construct knowledge based on Thai textual sources. The main question can be detailed by the following sub-questions:

1. How to resolve the anaphoric reference and textual ellipsis in Thai textual sources for knowledge extraction process.

2. How to analyze and construct the adequate semantic to provide a precise answer from Thai textual sources.

3. How to represent and formulate the obscure value from natural language sources to semantic structure.

4. How to extract the precise answer from semantic structure dealing with fuzzy factoid answer.

## 1.4 Research Objective

The main objective of this research is to develop a semantic-based question answering framework for fuzzy factoid answer from Thai textual sources. To achieve the main objective, the framework has to be developed by certain modules: morphological module, semantic parsing module, and question answering module. Moreover, the main objective can be formulated by the following sub-objectives:

1. To integrate the ellipsis resolution algorithm with the anaphora resolution algorithm for nominal, pronominal and zero anaphora from Thai textual source with accepted result.

2. To develop semantic parser for semantic based question answering framework from Thai textual sources.

3. To formulate frame-based for knowledge representation that handle the fuzzy factoid for semantic based question answering framework from Thai textual sources.



4. To develop semantic based question answering framework from Thai textual sources dealing with fuzzy factoid answer.

5. To evaluate the question answering framework by comparing with the criteria of with/without ellipsis and anaphora resolution and with/without using fuzzy logic.

## 1.5 Scope of Research

This research is aim to develop the semantic-based question answering system for the Thai language in the open domain. The question type in this system focus on factoid (Ranjan & Balabantaray, 2016) question: entity, location, date-time, and property. The

source of knowledge is focused on Thai unstructured textual sources. Thai Wikipedia is used as the data source for development.

### 1.6 Significance of Research

The significance of this research could be anticipated to provide theoretical contributions as below.

- First is the algorithm for ellipsis and anaphora resolution for Thai to enhance the discourse analysis in Thai NLP research and push it forward to a higher level.
- Second is the mechanism of a semantic-based question answering system that deals with fuzzy factoid answers.



The practical contributions are provided below.

- First is the semantic parser for Thai that can be acquired the knowledge and handle the factoid answer with fuzzy from Thai textual sources.
- second is a semantic-based Thai question answering system that deals with fuzzy factoid answers.

In addition, the resources (training model, semantic dictionary, etc.) that are used in this research could be distributed to the public to be used or adapted in other research.

# CHAPTER TWO

# LITERATURE REVIEWS

This chapter describes the issues in the overview of components of question answering system with natural language processing emphasized especially for the Thai language and also reviews the technique used in each issue. Consequently, the advantage and disadvantages are also described in this review.

## 2.1 Overview

The question answering system (QAS) can be divided into two approaches: data-intensive approach and knowledge-intensive approach (Jitkrittum et al., 2009). Data-intensive question answering system is not required to perform a complex NLP technique. The system uses a mathematical model especially probabilistic to analyze the amount of data and identify the answer from a piece of data. Knowledge-intensive question answering system highly uses NLP techniques to analyze and understand the question and find the exact answer. The system could work on background knowledge such as ontology and semantic lexicon to analyze natural language to logical form to construct the knowledge-based for querying the answer (Bordes et al., 2015; Usbeck et al., 2015; Yao, 2015).

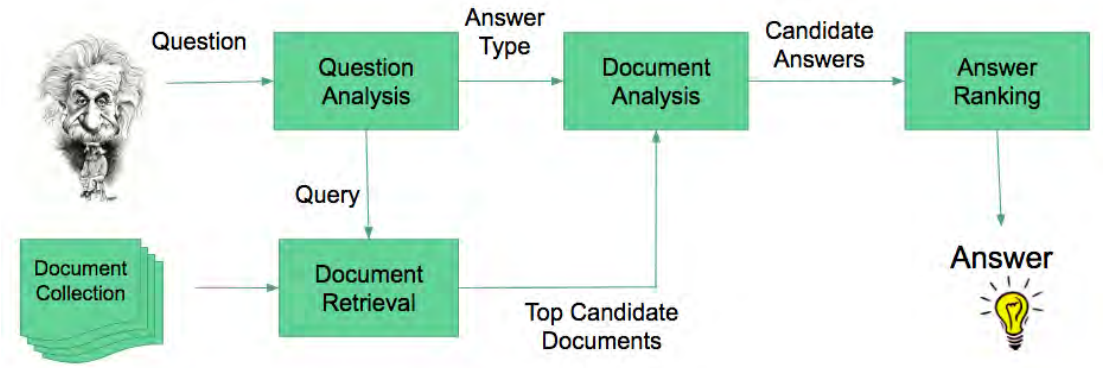


*Figure 2.1.* Overview of QAS in data-intensive approach

Figure 2.1 shows the overview of the question answering system in a data-intensive approach. The question is analyzed to generate the query statement to retrieve the documents that are relevant to the question from document collection or internet sources. After that, the candidate documents are processed by using a heuristic process to identify the candidate answers. Then, the candidate answers are ranked to find the best answer and deliver it to the users. This system is not complex and is easy to scale up. However, the performance depends on the document retrieval process and the precision of answers is a big issue for this system compared to knowledge-intensive question answering systems.

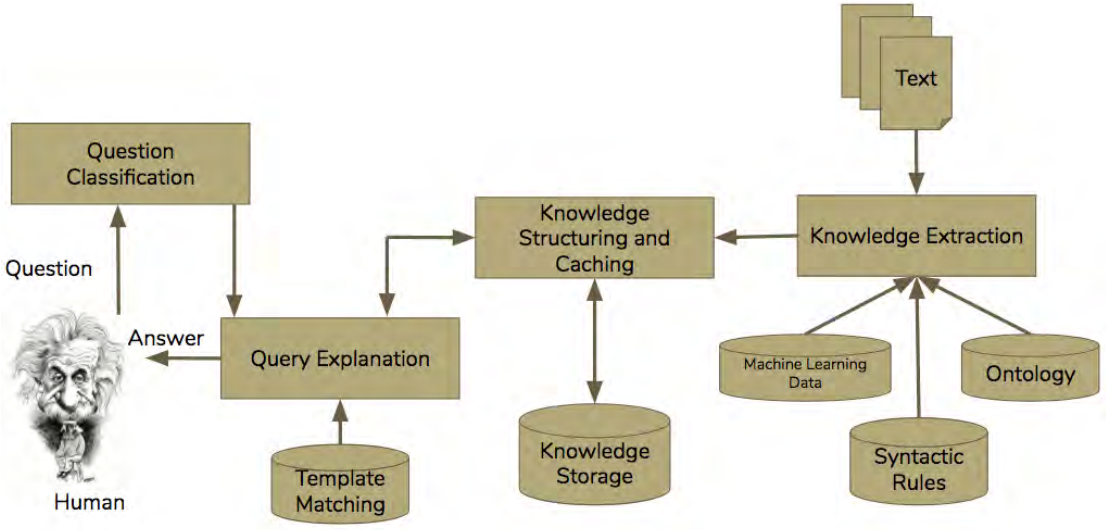


*Figure 2.2.* Overview of QAS in knowledge-intensive approach



Figure 2.2 shows the overview of the question answering system in the knowledge-intensive approach. Knowledge-intensive question answering systems translate textual data from textual sources into knowledge representation by using a knowledge extraction module and then store knowledge in knowledge storage. The knowledge extraction module consists of many NLP processes such as morphological analysis, syntactic analysis, and semantic analysis to create the knowledge-based for question answering system. The question from users, which poses in form of natural language, is analyzed to find the type of answer and translate to logical form by the question classification module. After that, the query explanation module identifies the exact answer from knowledge-based and then deliver the answer to a user in the proper form.

Morphological analysis is the process to analyze raw text from textual sources at the word level. Words will be recognized and processed to tag the part of speech. Subsequently, the named entities and anaphora resolution are significant parts to complete the structure of syntactic in its sentence. Then, The syntactic analysis will construct the syntactic information from the sentence by using the parsing technique. After that, the semantic analysis constructs knowledge based by creating the semantic structure from syntactic information.

In the Thai language, words in sentences are not separated by space or any clues. Hence, the word segmentation process is an important process to separate the words in a sentence before sending it to the next step. The named entities in the Thai language can be formed by clue words or all words are unknown words. Then, the named entities extraction module should process and extract named entities with high accuracy that influence the whole system. Most research in Thai natural language processing is hugely working in the morphological processing (Chanlekha & Kawtrakul, 2004; Kawtrakul & Thumkanon, 1997; Kruengkrai et al., 2006; Sornlertlamvanich, 1993; Sudprasert & Kawtrakul, 2003). All of the morphological analyses will be described in detail in Section 2.2.

After the word segmentation and named entities extraction are finished, the syntactic analysis plays an important role to construct the syntactic information. The semantic information will be identified from syntactic information by using a semantic dictionary and ontology as background knowledge to construct the knowledge from the sentence into knowledge representation. Eventually, The knowledge is collected in the

knowledge storage for use in the answer extraction process by the question answering system.

The data intensive approach can be done without the semantic construction that needs a complex natural language analysis. However, the precision is not high because the answer just comes from the ranking of the relation between the word in question and documents. The knowledge intensive approach needs a more complex process to construct the knowledge base however, the higher precision is worth to be done. Moreover, the knowledge base can be used for other applications such as the expert system and knowledge discovery research.



## 2.2 Morphological Analysis in Thai Text

The morphological analysis of the Thai language is an essential component of the knowledge-intensive approach question answering system that processes Thai text for word segmentation, named entities extraction, and also ellipsis and anaphora resolution with high accuracy. The accuracy of the morphological analysis is influenced by the precision of the whole system since knowledge structuring to answer extraction. The anaphora resolution will be reviewed in Section 2.4 and this section reviews word segmentation and named entity extraction in Thai NLP research.

### 2.2.1 Corpus Resources

It is more than three decades-long history (Arreerard et al., 2022), that Thai NLP research is active in developing the tools, resources, and knowledge. The process in Thai morphological analysis can be successful with reliable and comprehensive of Thai NLP resources such as the Thai corpus and Thai lexical database. The Thai NLP resource development can be developed for different purposes and domain usage. That made the various corpus development in the different methodologies and also the detail in corpus such as the tagging set and the annotation.

The very first corpus is developed for improving the process of tokenization and word segmentation such as Orchid (Sornlertlamvanich et al., 1999). The corpus contains content from encyclopedias, news, and novels and was annotated with word boundaries and sentence boundaries. The named entity information can be additionally annotated in some corpus such as Thai-Nest (Theeramunkong et al., 2010), Nattadaporn (Lertcheva, 2010), and LST20 (Boonkwan et al., 2020). The LST20 is also annotated with more information: part of speech, word boundaries, sentence boundaries, and clause boundaries. The Thai syntactic dependency information was annotated in Blackboard Treebank which featured dependency structures, syntactic constituency structures, word boundaries, named entities, clause boundaries, and sentence boundaries. TALPCo (Nomoto, 2019) is a parallel corpus that contains the Thai language. The corpus consists of Japanese sentences together with their translations into eight languages, the Thai language is one of them. There still be many corpus that are developed for various purposes and different domains. The exhaustive list of Thai corpus resources is shown in Table 2.1.

Table 2.1

*Exhaustive list of Thai corpus resources (Arreerard et al., 2022)*

| **Corpus** | **Developer** | **Data Size** |
|---|---|---|
| BEST | NECTEC | 5 million words |
| HSE Thai | HSE School of Linguistics | 50 million tokens |
| LST20 | NECTEC | 3 million words, 288,020 Named Entities |
| Nattdaporn | Nattadaporn Lertcheva | 178,474 words, 2,463 Named Entities |
| Nutcha | Nutcha Tirasaroj | 367,673 words, 16,179 Named Entities |
| Sasiwimon | Sasiwimon Kalunsima | 80,513 words, 2,954 Named Entities |
| VISTEC-2021 | VISTEC and CMU | 3.39M words, 49,997 sentences |
| Wongnai-corpus | Thongthanomkul et al. | 500,000 unique words, 39,999 reviews |
| Thai-Nest | NECTEC | 45,000+ Named Entities |
| ClickBait | Wannaphong Phatthiyaphaibun | 350 sentences |
| Mt-opus | VISTEC | 5.4 million sentence pairs |
| Orchid | NECTEC | 30,000 sentences |
| Prachathai | Phatthiyaphaibun et al. | 67,000 sentences |
| TALPCo | Nomoto et al. | 1,372 sentences |
| Thai Universal Dependency | UD Thai PUD | 1,000 sentences |
| Thai Wiki QA | NECTEC | 17,000 sentences |
| Thai Literature Corpora (TNHC set) | Jitkapat Sawatphol | 756,478 lines, 47 documents |
| Toxic tweet | Sirihattasak et al. | 3,300 tweets |
| Wisesight | Suriyawongkul et al. | 26,737 messages |
| Thai QA | NECTEC | 4,000 questions |
| Thai-joke-corpus | iApp Technology | 449 jokes |
| Prime Minister29 | Phatthiyaphaibun et al. | 6 documents, 338KB |
| Thai Plagiarism | NECTEC | 1,050 plagiarism texts, 554MB Source docs |
| Scb-mt-en-th-2020 | VISTEC and SCB | 1 million Thai-English texts |
| Blackboard Treebank | NECTEC | 130,561 Trees |
| Wikipedia dumps | Wikipedia | 2.08GB |

### 2.2.2 Word Segmentation

Work on Thai word segmentation started in the 1980s. Thai word segmentation starts with using the matching algorithm with a dictionary-based to identify the word boundary. Poowarawan (1986) used the dictionary based on the longest matching algorithm to solve word boundaries in Thai sentences. This work used a dictionary that contains all Thai words as internal data and find word boundaries for each word in a sentence by implementing the longest matching algorithm. With dictionary-based, it is no need to use any training data, and easy to implement. However, this algorithm cannot identify and works if there is an unknown word that is not contained in the dictionary. Dictionary should be built with high comprehensive to use in the real problem. Sornlertlamvanich (1993) proposed the maximum matching algorithm for Thai word segmentation. This work splits a sequence of characters into all possibilities of segmentation based on a word set and selects the segmentation path with the lowest number of segmentation tokens. This algorithm has also used the dictionary based to match the segmentation tokens in each sentence. The matching algorithm is also implemented in some Asian languages due to the low resources and computation is a major concern. Htay and Murthy (2008) proposed the syllable level longest matching to identify the word boundary in Myanmar text. This technique works well with low resources for Myanmar corpus. Srithirath and Seresangtakul (2013) presented the longest syllable level matching to recognize the Lao-named entities from the syllable level of words. However, the matching algorithm does not have high precision to perform the highly accurate NLP.

Machine learning is used to increase the precision of word segmentation. The statistical approach is started in 1997. Kawtrakul and Thumkanon (1997) proposed the statistical approach based on the tri-gram Markov model. This work used a dictionary-based in the first step to generate all possibilities of the sequence of words in the sentence. Then, the next step used the tri-gram Markov model to make a decision that which sequence of the word is the right segmentation with part of speech tagging. However, by using of tri-gram Markov model, the size of the training corpus is crucial to training the model. The training corpus must be comprehensive and big enough for training to use in that domain. Aroonmanakun (2002) proposed the maximum collocation approach to determine the word combination from the sequence of syllables. This work started with generating the sequence of syllables from sentences with a syllable pattern matching technique. After that, the word will be constructed from syllables by finding the maximum collocation of each syllable. However, the training corpus is still needed to find the maximum collocation. Sudprasert and Kawtrakul (2003) proposed the unsupervised learning model with global and local data in the corpus. This work also used dynamic programming techniques to improve the training speed of the model. Kruengkrai et al. (2006) used the Conditional Random Field (CRF) algorithm for training a word segmentation model. The CRF machine learning has shown a result that is better than other machine learning models for the task of labeling and segmenting sequence data (Lafferty et al., 2001; Peng et al., 2004). In the first step, the possibilities of word segmentation are produced by using the combination of the longest matching algorithm and backtracking technique. In the last step, this work used the CRF framework to select the optimal path from word and part of speech tagging. Kongyoung et al. (2015) also used CRF integrated with three dictionaries in pre-processing and post-processing phrases to improve the accuracy of the word segmentation with an

impressive result. The CRF also works well in some Asian language. Pa et al. (2016) experimented the word boundary identification for Myanmar text that compared between CRF technique with the method based on the maximum matching method. The results revealed that the CRF technique performed a higher F-score than the maximum matching method. Vanthanavong and Haruechaiyasak (2011) experimented with Lao word segmentation using the CRF technique. The corpus size is 100,000 words for training. The CRF technique compares the performance with the dictionary-based technique and the results revealed that the CRF has higher precision and recall score. Boonkwan and Supnithi (2017) used a recurrent neural network model of deep learning technique by using the integration of two-level backoff models and character-level contexts to process the word segmentation and part of speech tagging. They have experimented with Thai and Chinese languages. The results reveal that the model significantly improves the overall accuracy of word segmentation.



The overview of Thai word segmentation work is done by using longest matching and maximum matching with a dictionary in the first era and then the machine learning technique is coming to gain more accuracy with various kinds of machine learning. Table 2.2 shows the summary of word segmentation research.

Table 2.2

*Summary of word segmentation research*

| Technique | Description | Advantage | Disadvantage | Reference |
|---|---|---|---|---|
| Dictionary based | make predefined data into a words dictionary and used the matching algorithm to identify the sequence of the word in the sentence. | No training requires. Easy to implement. | Cannot identify the unknown word. Dictionary must be comprehensive. | Poowarawan (1986) Sornlertlamvanich (1993) |
| Tri-gram Markov Model | Used probabilistic model to make a decision that which sequence of the word is the right segmentation. | Fast segmentation process. | Needed big data training set. Training data may be not comprehensive. | Kawtrakul and Thumkanon (1997) |
| Maximum Collocation | Used maximum collocation technique to choose the right segmentation from all possibilities of the sequence of syllables. | Fast segmentation process. | Needed training data and syllable rules set that should be comprehensive and accurate. | Aroonmanakun (2002) |
| Conditional Random Field | Used Condition Random Field learning model to identify the best segmentation from the sequence of words. | High accuracy. | Needed training data that should be comprehensive. | Kruengkrai et al. (2006) Kongyoung et al. (2015) |
| Deep Learning | Used a recurrent neural network model of deep learning technique to predict the word boundary. | High accuracy. | Needed training data that should be comprehensive and high computation hardware. | Boonkwan and Supnithi (2017) |

The conditional random fields (CRF) perform very well with less effort of intense computation compare with the deep learning technique. However, the additional process to correct some wrong segmentation by merging words and approved by the dictionary can be a good idea to increase the precision of the word segmentation. After finishing the word segmentation process, named entity extraction is the next process to identify the entity that appears in the sentence by using the information from word segmentation and part of speech tagging.

### 2.2.3 Named Entity Extraction

Named entity extraction is an important part of the question answering system. The named entity is the object such as person, organization, location, etc., that can be a factoid answer of the question answering system. A good accurate named entity extraction can improve the accuracy of the question answering system.



Charoenpornsawat et al. (1998) used Winnow algorithm (Blum, 1997) for Thai Named Entity identification with the feature-based approach. The features such as context words, collocation, part of speech tagging, and heuristic information from the dictionary are used to generate named entity candidates and also solve named entity boundaries. However, this experiment used the corpus that segments words and part of speech tagging manually by the linguist.

Chanlekha and Kawtrakul (2004) used Maximum Entropy Model and heuristic information together with the dictionary to extract the Thai-named entity. The

combination of rules-based, dictionary-based, and statistical-based are used to predict the boundary of a named entity and also to identify the named entity type such as a person, organization, and location. However, this work needed a big corpus for training and the corpus must be comprehensive. The Maximum Entropy-based named entity recognition also works well in some languages. Riaz et al. (2020) experimented with the Maximum Entropy-based named entity recognition in the Urdu language. The challenges for the Urdu language are that there is no word capitalization, borrowed words, and ambiguous acronyms. The results show the good performance of precision and recall measures. Benajiba et al. (2007) presented ANERsys which is an Arabic language named entity recognition based on the Maximum Entropy model. The important characteristics of the Arabic language are that a character may have up to three different forms, Arabic does not have a capital letter and very complex morphology. The results shown that performance is good for some kind of named entity.



Sutheebanjard and Premchaiswadi (2009) used rule-based and front-rear context to identify Thai personal names without using word segmentation and also part of speech tagging. This experiment is in the corpus in the domain of political, financial, and sports articles. The result is good in the political and financial domain because the pattern of text that appears in the domain is not in a different form. However, in the sports domain, the model is not performing well.

The research in named entity extraction is ongoing research and needs to improve its accuracy. The named entity in the Thai language can be formed by a common word, an

unknown word, or both. Accuracy in Thai named entity extraction is very challenging. Table 2.3 shows the summary of Thai name entity extraction research.

Table 2.3

*Summary of Thai named entity extraction research*

| Technique | Description | Advantage | Disadvantage | Reference |
|---|---|---|---|---|
| Winnow Algorithm | Used Winnow algorithm to generate named entity candidate and solve the named entity boundary. | Can identify the proper name that contains various forms of known words and unknown strings, High accuracy | Training corpus must be segmented words and tagged POS by an expert linguist manually. | Charoenpornsawat et al. (1998) |
| Maximum Entropy Model | Used Maximum Entropy Model to predict the boundary of named entity and type of named entity. | Can detect the type of proper name without using POS tagging, Good Accuracy. | Training corpus must be big enough and comprehensive. | Chanlekha and Kawtrakul (2004) |
| Rules Based | Used Rule-based to identify Thai personal names without word segmentation. | Can extract Thai personal name without using word segmentation, Easy to implement. | Perform not well in some domains. | Sutheebanjard and Premchaiswadi (2009) |

Word boundary and named entity are identified in the overall morphological analysis. The next step is that all information is sent to the syntactic analysis to create the syntactic structure.

## 2.3 Syntactic Analysis

The syntactic parser is a process to analyze the syntactic structure in a sentence. After the morphological analysis process is done, The syntactic structure will be analyzed by using information from morphological analysis such as word boundary and name entities. Grammar rules are the essential component of the syntactic parser that is used to identify the syntactic structure in each sentence. Thai syntactic parser is very rare research and mostly used in a small work (Aroonmanakun, 1989; Tongchim et al., 2008). The parser can be categorized into 2 types: dependency parser and context-free grammar parser.



### 2.3.1 Dependency Parser

A dependency parser is a syntactic analyzer that establishes the relationship between headwords and words that modify the headword. Dependency structure can be defined as a directed graph in that words are nodes and relations are edges. Figure 2.3 shows the example of dependency structure.

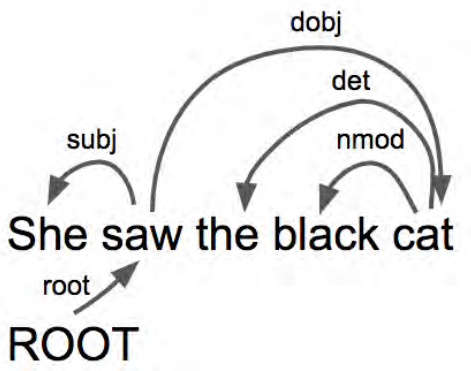


*Figure 2.3.* Example of dependency structure

LR parser (Knuth, 1965) is proposed for an efficient left-to-right parsing algorithm. LR algorithm process in bottom-up strategies by applying shift-reduce techniques.

McDonald et al. (2005) proposed non-projective dependency parsing by using spanning tree algorithm (Edmonds, 1967). This parser can reduce time complexity from $O(n^3)$ to $O(n^2)$ from Eisner (1996) probabilistic dependency parsing by utilizing the spanning tree algorithm. Figure 2.4 shows the maximum spanning tree finding algorithm.

**Chu-Liu-Edmonds**$(G, s)$

Graph $G = (V, E)$

Edge weight function $s : E \rightarrow \mathbb{R}$

1. Let $M = \{(x^*, x) : x \in V, x^* = \arg\max_{x'} s(x', x)\}$
2. Let $G_M = (V, M)$
3. If $G_M$ has no cycles, then it is an MST: return $G_M$
4. Otherwise, find a cycle $C$ in $G_M$
5. Let $G_C = \text{contract}(G, C, s)$
6. Let $\boldsymbol{y} = \text{Chu-Liu-Edmonds}(G_C, s)$
7. Find a vertex $x \in C$ s. t. $(x', x) \in \boldsymbol{y}, (x'', x) \in C$
8. return $\boldsymbol{y} \cup C - \{(x'', x)\}$

**contract**$(G = (V, E), C, s)$

1. Let $G_C$ be the subgraph of $G$ excluding nodes in $C$
2. Add a node $c$ to $G_C$ representing cycle $C$
3. For $x \in V - C : \exists_{x' \in C}(x', x) \in E$
   Add edge $(c, x)$ to $G_C$ with
   $s(c, x) = \max_{x' \in C} s(x', x)$
4. For $x \in V - C : \exists_{x' \in C}(x, x') \in E$
   Add edge $(x, c)$ to $G_C$ with
   $s(x, c) = \max_{x' \in C} [s(x, x') - s(a(x'), x') + s(C)]$
   where $a(v)$ is the predecessor of $v$ in $C$
   and $s(C) = \sum_{v \in C} s(a(v), v)$
5. return $G_C$

*Figure 2.4.* Algorithm for finding maximum spanning trees in directed graphs (McDonald et al., 2005)

Recurrent neural networks with long short-term memory units (LSTMs), which are called stack LSTMs (Ballesteros et al., 2016, 2017; Dyer et al., 2015), are introduced in dependency parsing. This parser computes character-based continuous-space vector embedding of words using bidirectional LSTMs. The results are stronger for agglutinative languages, such as Basque, Hungarian, Korean, and Turkish.

### 2.3.2 Context-free Grammar Parser

Context-Free Grammar (CFG) is a set of production rules that describe all possible strings in a given formal language. The CFG is in form of a phrase-structure grammar

that is a rule to describe how a phrase structure is constructed from words. Figure 2.5 shows the example of CFG.

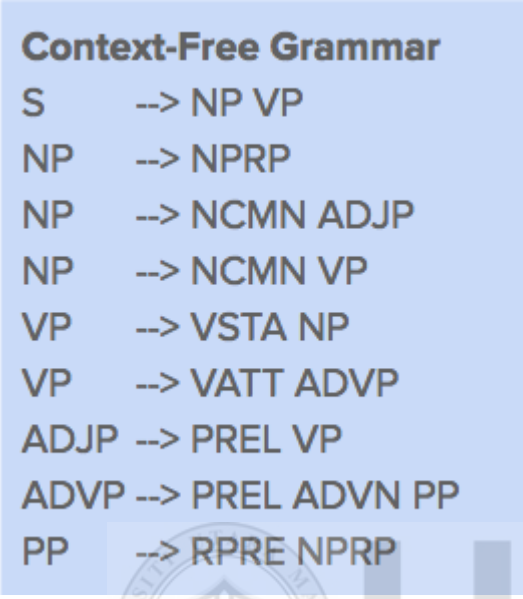


*Figure 2.5.* Example of Context-Free Grammar



Chomsky (1956) proposed the finite state machine to analyze the syntactic structure of English by applying finite state grammars in form of a phrase-structure grammar. This algorithm is a deterministic process that can be run in time complexity because the rules can be recursive.

CYK parsing algorithm (Cocke, 1970; Kasami, 1965; Lange & Leiß, 2009; Younger, 1967) is developed by applying dynamic programming techniques to reduce the time complexity of the parsing process. CYK algorithm solved the syntactic structure from CFG with bottom-up strategies. Grammar rules in the CYK algorithm must be

Chomsky Normal Form (CNF). Figure 2.6 shows the pseudocode of the CYK algorithm.

```
let the input be a string I consisting of n characters: a_1 ... a_n.
let the grammar contain r nonterminal symbols R_1 ... R_r, with start symbol R_1.
let P[n,n,r] be an array of booleans. Initialize all elements of P to false.
for each s = 1 to n
  for each unit production R_v -> a_s
    set P[1,s,v] = true
for each l = 2 to n -- Length of span
  for each s = 1 to n-l+1 -- Start of span
    for each p = 1 to l-1 -- Partition of span
      for each production R_a -> R_b R_c
        if P[p,s,b] and P[l-p,s+p,c] then set P[l,s,a] = true
if P[n,l,1] is true then
  I is member of language
else
  I is not member of language
```

*Figure 2.6.* Pseudocode of CYK algorithm (Lange & Leiß, 2009)

Earley (1970) proposed the parsing algorithm that applies dynamic programming to reduce time complexity with top-down strategies. Grammar rules in the Earley algorithm are not necessary for CNF. Figure 2.7 shows the pseudocode of Earley algorithm.

```
DECLARE ARRAY S;

function INIT(words)
    S ← CREATE-ARRAY(LENGTH(words))
    for k ← from 0 to LENGTH(words) do
        S[k] ← EMPTY-ORDERED-SET

function EARLEY-PARSE(words, grammar)
    INIT(words)
    ADD-TO-SET((γ → •S, 0), S[0])
    for k ← from 0 to LENGTH(words) do
        for each state in S[k] do  // S[k] can expand during this loop
            if not FINISHED(state) then
                if NEXT-ELEMENT-OF(state) is a nonterminal then
                    PREDICTOR(state, k, grammar)         // non-terminal
                else do
                    SCANNER(state, k, words)             // terminal
            else do
                COMPLETER(state, k)
        end
    end
    return chart

procedure PREDICTOR((A → α•Bβ, j), k, grammar)
    for each (B → γ) in GRAMMAR-RULES-FOR(B, grammar) do
        ADD-TO-SET((B → •γ, k), S[k])
    end

procedure SCANNER((A → α•aβ, j), k, words)
    if a ⊂ PARTS-OF-SPEECH(words[k]) then
        ADD-TO-SET((A → αa•β, j), S[k+1])
    end

procedure COMPLETER((B → γ•, x), k)
    for each (A → α•Bβ, j) in S[x] do
        ADD-TO-SET((A → αB•β, j), S[k])
    end
```

*Figure 2.7.* Pseudocode of Earley algorithm (Jurafsky & Martin, 2009)

Pingali and Bilardi (2015) proposed a graphical model for CFG parsing. They presented a graphical representation of CFG called the Grammar Flow Graph (GFG) that permits parsing problems to be phrased as path problems in graphs. Figure 2.8 shows the Grammar Flow Graph.

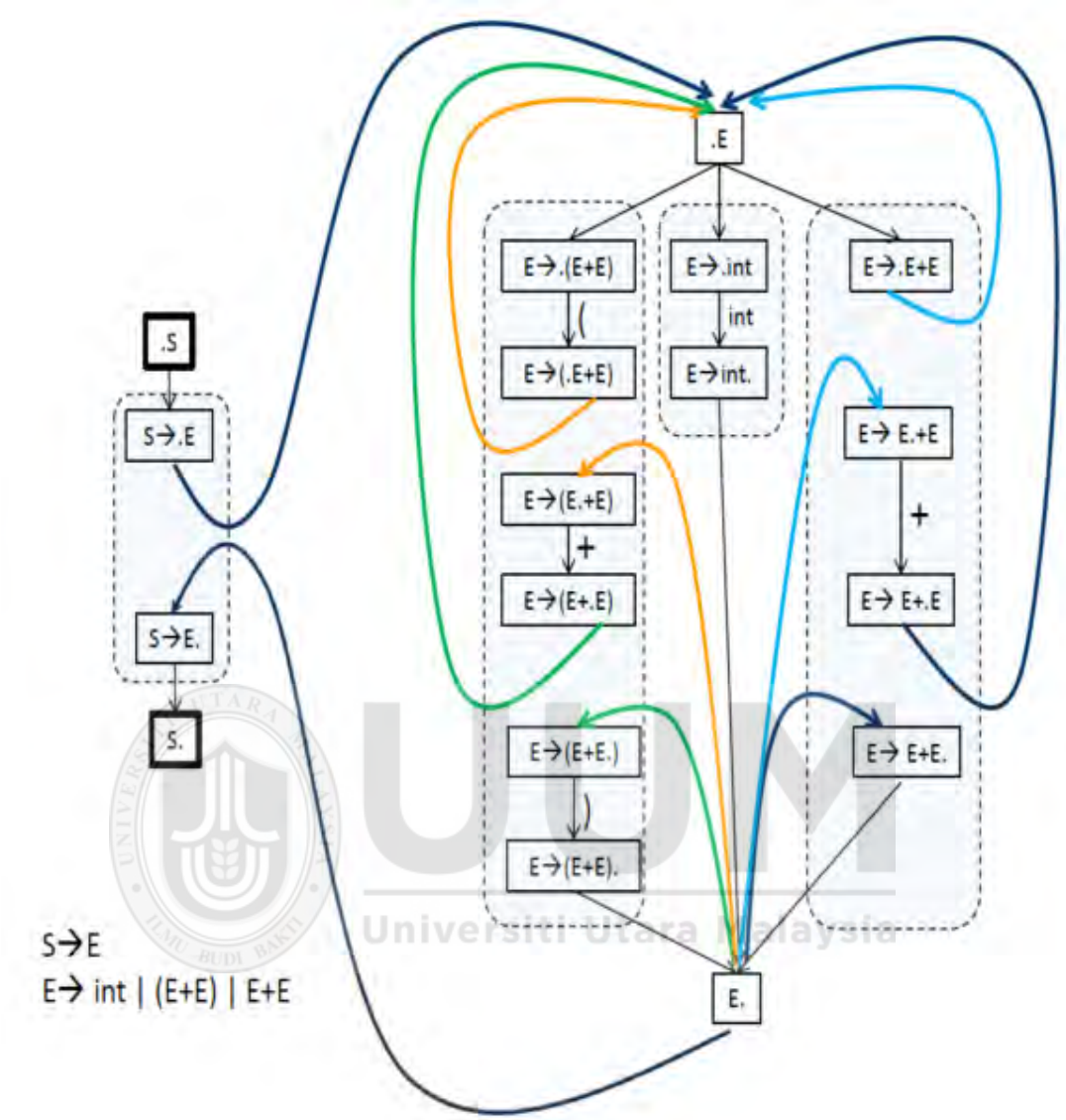


*Figure 2.8.* Grammar flow graph (Pingali & Bilardi, 2015)

### 2.3.3 Syntactic Analysis in Thai text

The research in natural language processing in Thai is mostly focused on morphological analysis until nowadays. Research in Thai syntactic analysis is very limited. The very first work in Thai syntactic analysis is in the master thesis by Aroonmanakun (1989). However, this work is done with a very small corpus (50 sentences).

Satayamas et al. (2005) have developed the processes for Thai tree bank construction. A tree bank of 400 sentences is used in this work which is quite small.

Tongchim et al. (2008) has developed a dependency parser for Thai that is a part of an ongoing project in developing a syntactically annotated Thai corpus. This work used some machine learning such as support vector machines (Chang & Lin, 2001; Yamada & Matsumoto, 2003) and maximum entropy model (Ratnaparkhi, 1999) to determine the sequences of actions in the parser. To find the best dependency structure, the search algorithm is proposed that is shown in Figure 2.9



$\mathcal{T}_s \leftarrow \{w_1, w_2, ..., w_n\}$
$\mathcal{T}_u \leftarrow \{\}$
$l = \text{argmax}_{i \in [1,n]} Prob_{root}(i)$
Remove $w_l$ from $\mathcal{T}_s$
Add $w_l$ to $\mathcal{T}_u$
**while** $\mathcal{T}_s$ *is not empty* **do**
    Find $w_o$, $w_o \in \mathcal{T}_s$ and its head $w_p, w_p \in \mathcal{T}_u$ which maximize the conditional probability
    Remove $w_o$ from $\mathcal{T}_s$
    Add $w_o$ to $\mathcal{T}_u$
**end**

*Figure 2.9.* Algorithm to find the best dependency structure (Tongchim et al., 2008)

Research in this area is still rare and needs more work to succeed. However, the development of syntactic annotates corpus is ongoing in several organizations that will be a very good source for making the better parser and syntactic analysis research area.

### 2.4 Ellipsis and Anaphora Resolution

Anaphora is a language instrument that is used for referring to the object that used to be introduced before in a previous sentence. The use of anaphora is a normal phenomenon in natural language. To construct the knowledge from the sentence, the anaphora must be resolved that which object it is referred to.

There are many kinds of anaphora in natural language. The very interesting kind of anaphora in the Thai language is zero anaphora. Zero anaphora is the disappearing of a noun phrase, that is used to introduce before, in a sentence for referring to the object. Figure 2.10 shows the example of zero anaphora in Thai language.



1) เพลี้ยไฟเป็นแมลงขนาดเล็ก
The aphid is a small insect.

2) Φ มีสีเหลืองหรือสีน้ำตาลอ่อน
Φ has a yellow or light brown color.

Figure 2.10. Example of zero anaphora in the Thai language

There is a research of Thai zero anaphora resolution (Aroonmanakun, 1997, 2000) by utilized the centering approach (Di Eugenio, 1990, 1996; Grosz et al., 1983; Walker et al., 1990, 1996). The centering theory is a computation model to consider the local coherence in a discourse segment. However, this work is tested on a very small corpus.

Textual ellipsis is a phenomenon in the Thai language in that some prepositions in a noun phrase disappear from the sentence but are still able to understand what it refers to. Textual ellipsis occurs many times in the Thai language as usually. Figure 2.11 shows the example of textual ellipsis in Thai language.

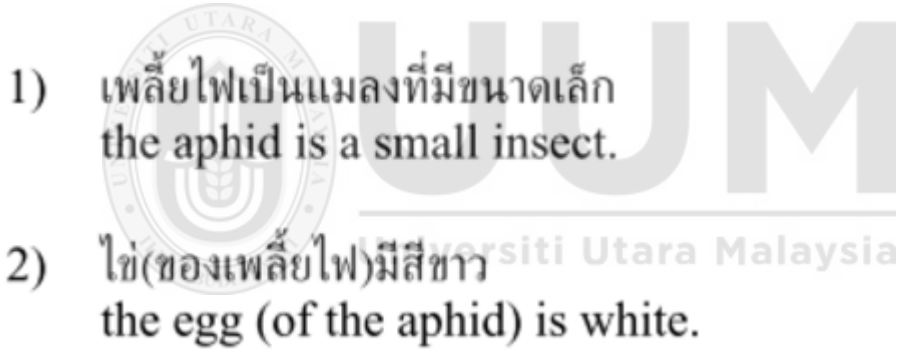


Figure 2.11. Example of textual ellipsis in the Thai language

To construct the complete knowledge from the sentence, zero anaphora and textual ellipsis have to resolve to find the complete semantic. Kongwan and Kawtrakul (2005) developed the system that extracts object properties from Thai text. This project proposed the ellipsis resolution algorithm in Thai text that was tested in agriculture domain text. Figure 2.12 shows the ellipsis resolution algorithm.

```
Input : S is the current sentence
Output : S is the solved sentence

EllipsisRes(S):
1 N = getSubject(S)
2 if N is no possession preposition:
3     Sp = getPreviousSubject(Sp)
4     if N is “part, piece” or “plant part” sense:
5         while Sp != null:
6             I = getPreviousSubject(Sp)
7             if N is same I:
8                 if I is no prep(I):
9                     Sp = getPreviousSentence(Sp)
10                    continue
11                addPrep(N,prep(I))
12                return S
13            else if N is meronym of I:
14                addPrep(N,I)
15                return S
16            else if N is meronym of prep(I):
17                addPrep(n,prep(I))
18                return S
19            else:
20                Sp = getPreviousSentence(Sp)
21    else: return S
```

*Figure 2.12.* Ellipsis resolution algorithm (Kongwan & Kawtrakul, 2005)

However, there is very little research in anaphora and ellipsis resolution that have done in this area. An improved algorithm is needed to be done and is still in ongoing research in this field. Machine learning can be utilized with the relevance feature from the context of the text and could be an interesting study to improve the anaphora and ellipsis resolution in Thai text.

## 2.5 Semantic Analysis

Semantic analysis is a process to determine the correct word sense from surface words and construct the semantic structure from the textual source. The semantic analysis can be classified into 2 approaches: The syntactic-driven approach and the semantic-driven approach.

### 2.5.1 Syntactic-Driven Approach

Syntactic-Driven Approach is a semantic analysis process that starts with putting the sentence into the syntactic analysis process. Then, the syntactic structure from the syntactic analysis process is submitted to the semantic analyzer to produce the semantic structure by the appropriate semantic representation.

Miller et al. (1993, 1996) proposed a hidden understanding model for parsing that extend the statistic syntactic parsing model to incorporate a semantic label on each node. And Miller et al. (2000) have adapted a probabilistic context-free parser by utilizing a head-driven model for information extraction on MUC-7.

Zettlemoyer and Collins (2005) proposed the algorithm that automatically induces that map sentence to the semantic structure along with a probabilistic parsing model. The grammar formalism is utilized by the Combinatory Categorial Grammar (CCG) (Steedman, 1996, 2000) in their work.

Bos (2015) proposed the Boxer which is a semantic parser for English texts based on Discourse Representation Theory. This work is utilized the CCG in syntactic driven analysis.

### 2.5.2 Semantic-Driven Approach

Semantic-Driven Approach is a semantic analysis process that constructs the semantic structure directly from the surface sentence. Semantic grammar (Jurafsky, Daniel and Martin, 2000) approach work on translating sentence directly to semantic representation. A syntactic parsing algorithm can be directly used to parse semantics in the semantic grammar approach. (Kate et al., 2005) developed the system called SILT by utilizing a semantic grammar approach. The semantic grammar is induced by the learning method.



The semantic analysis can be done by both approaches that depend on the information and the needs of the application. In a less complex system, the semantic-driven approach could be more suitable. For the deep semantic analysis, the syntactic-driven could be a significant part of the knowledge construction from the whole text.

## 2.6 Knowledge Representation

Knowledge representation is an essential part of the question answering system. Knowledge is extracted from knowledge sources such as text, and web pages and formulated into knowledge representation that is able to retrieve or query the object or

answer in the future. This section reviews the knowledge representation that is frequently used to build the knowledge-based application.

### 2.6.1 Resource Description Framework (RDF)

Resource Description Framework (RDF) (Gibbins, 2016; Lassila & Swick, 1999) is a standard model for data interchange on the web. This standard model is recommended by W3C to use in semantic web technology. Nevertheless, RDF is able to use to be a knowledge representation for some working on question answering systems. There are 3 syntax types that are used for representing the knowledge in RDF: N-Triples, Turtle, and XML. Figure 2.13 shows example of RDF graph.

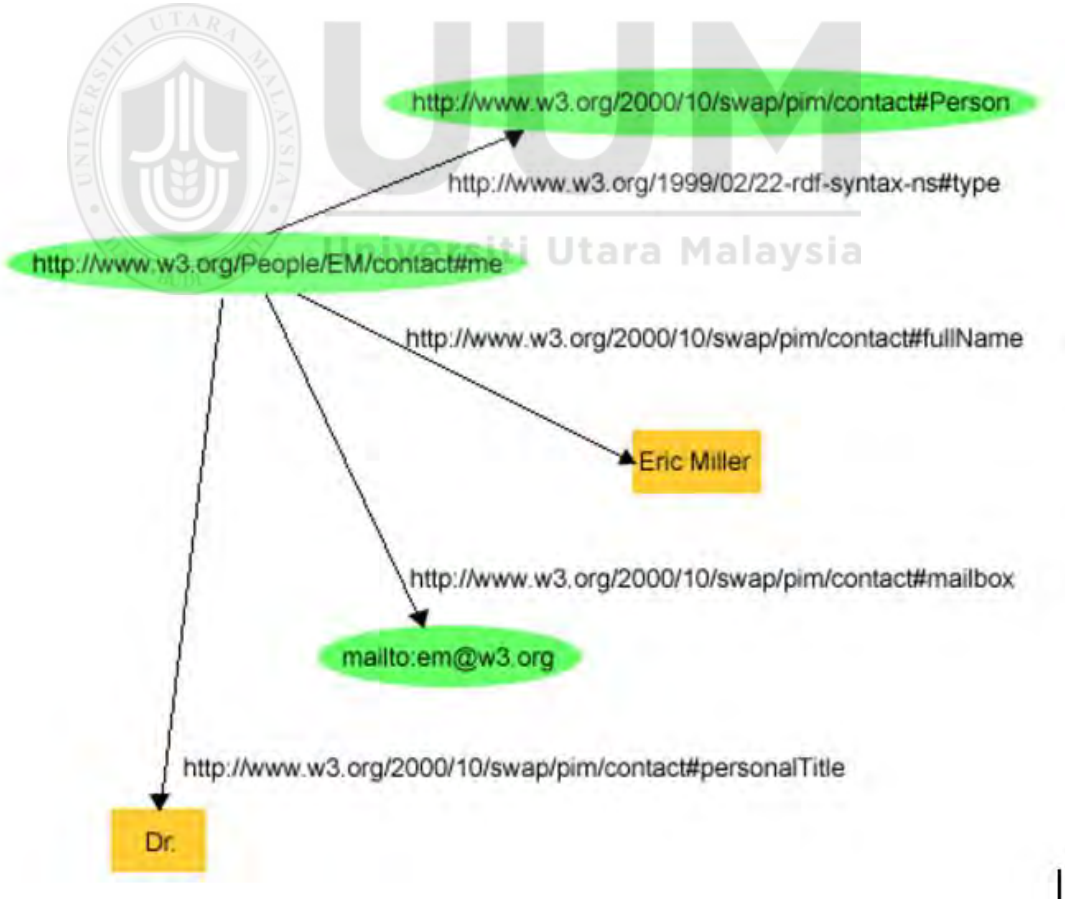


*Figure 2.13.* Example of RDF graph (Wikipedia, 2007)

From Figure 2.13, the RDF graph shows that “Eric Miller” is the contact name and “em@w3.org” is his e-mail and his title is “Dr.”. Figure 2.14 shows RDF in N-Triples syntax. Figure 2.15 shows RDF in Turtle syntax. Figure 2.16 shows RDF in XML syntax.

```
<http://www.w3.org/People/EM/contact#me>
<http://www.w3.org/2000/10/swap/pim/contact#fullName> "Eric Miller" .

<http://www.w3.org/People/EM/contact#me>
<http://www.w3.org/2000/10/swap/pim/contact#mailbox>
<mailto:e.miller123(at)example> .

<http://www.w3.org/People/EM/contact#me>
<http://www.w3.org/2000/10/swap/pim/contact#personalTitle> "Dr." .

<http://www.w3.org/People/EM/contact#me>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://www.w3.org/2000/10/swap/pim/contact#Person> .
```

*Figure 2.14.* Example of N-Triples syntax (Wikipedia, 2007)

```
@prefix eric:    <http://www.w3.org/People/EM/contact#> .
@prefix contact: <http://www.w3.org/2000/10/swap/pim/contact#> .
@prefix rdf:     <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .

eric:me contact:fullName "Eric Miller" .
eric:me contact:mailbox <mailto:e.miller123(at)example> .
eric:me contact:personalTitle "Dr." .
eric:me rdf:type contact:Person .
```

*Figure 2.15.* Example of Turtle syntax (Wikipedia, 2007)

```
<?xml version="1.0" encoding="utf-8"?>
<rdf:RDF xmlns:contact="http://www.w3.org/2000/10/swap/pim/contact#"
xmlns:eric="http://www.w3.org/People/EM/contact#"
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
  <rdf:Description rdf:about="http://www.w3.org/People/EM/contact#me">
    <contact:fullName>Eric Miller</contact:fullName>
  </rdf:Description>
  <rdf:Description rdf:about="http://www.w3.org/People/EM/contact#me">
    <contact:mailbox rdf:resource="mailto:e.miller123(at)example"/>
  </rdf:Description>
  <rdf:Description rdf:about="http://www.w3.org/People/EM/contact#me">
    <contact:personalTitle>Dr.</contact:personalTitle>
  </rdf:Description>
  <rdf:Description rdf:about="http://www.w3.org/People/EM/contact#me">
    <rdf:type rdf:resource="http://www.w3.org/2000/10/swap/pim/contact#Person"/>
  </rdf:Description>
</rdf:RDF>
```

*Figure 2.16.* Example of XML syntax (Wikipedia, 2007)

Jitkrittum et al. (2009) proposed the QAST system. QAST is the question answering system for Thai Wikipedia. The system constructs the knowledge from structured information from Wikipedia into RDF format and unstructured test store as a search index. RDF triples are generated from Wikipedia's infoboxes (Auer & Lehmann, 2007; Isbell & Butler, 2007) that are used as the first knowledge source for the user's question. Unstructured text is used as the secondary source of knowledge when the first knowledge source is not able to identify the answer. SPARQL query (Prud'Hommeaux & Seaborne, 2008) is generated from the user's question-by-question pattern to query the answer from RDF knowledge-based. Figure 2.17 shows the system architecture of QAST system.

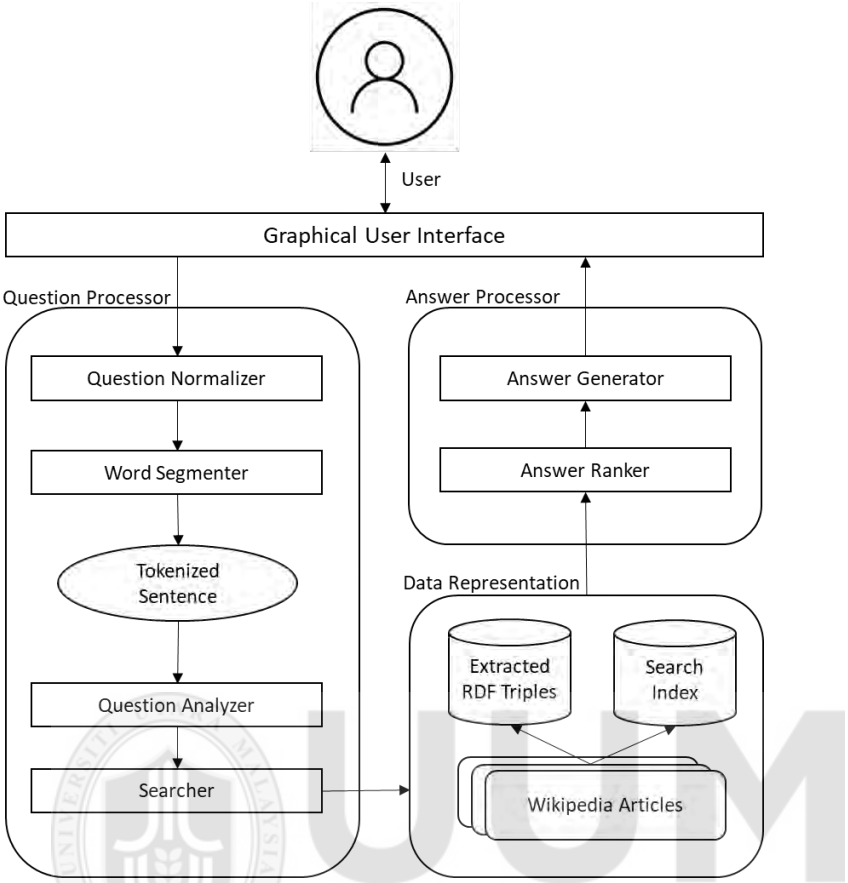


*Figure 2.17.* System architecture of QAST system (Jitkrittum et al., 2009)

Kongthon et al. (2011) proposed the semantic-based question answering system for Thailand tourism information. This work constructed the tour ontology (Fodor & Werthner, 2005; Pranter et al., 2007) that is formulated in RDF form. The question is analyzed by pattern matching technique and then generate the SPARQL query to extract the answer from the tour ontology. Figure 2.18 shows the proposed system.

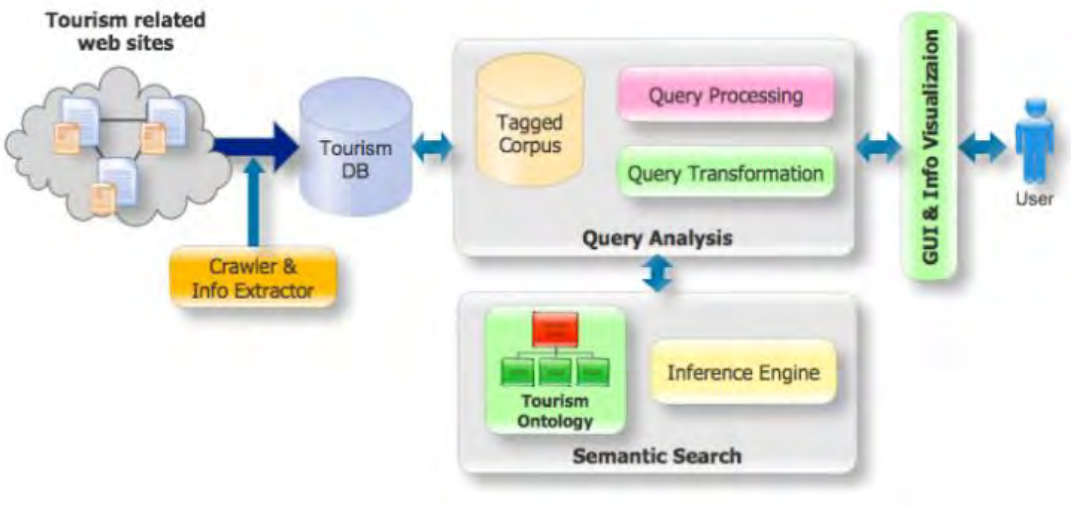


*Figure 2.18.* The proposed system of Thailand tourism information system (Kongthon et al., 2011)

The RDF is a good knowledge representation for the semantic web. However, the RDF is hard to fill the whole semantic of the sentence and also is not support handling the fuzzy factoid. The RDF is more suitable to create the specific domain rather than the open domain question answering system.



### 2.6.2 Universal Networking Language (UNL)

Universal networking language (UNL) (Dhindsa & Sharma, 2016; Kumar & Goel, 2016; Uchida et al., 2005) is a declarative formal language that is used to represent the semantic data from natural language and language independence. There are some works in natural language processing that is related to UNL representation. Figure 2.19 shows the example of UNL.

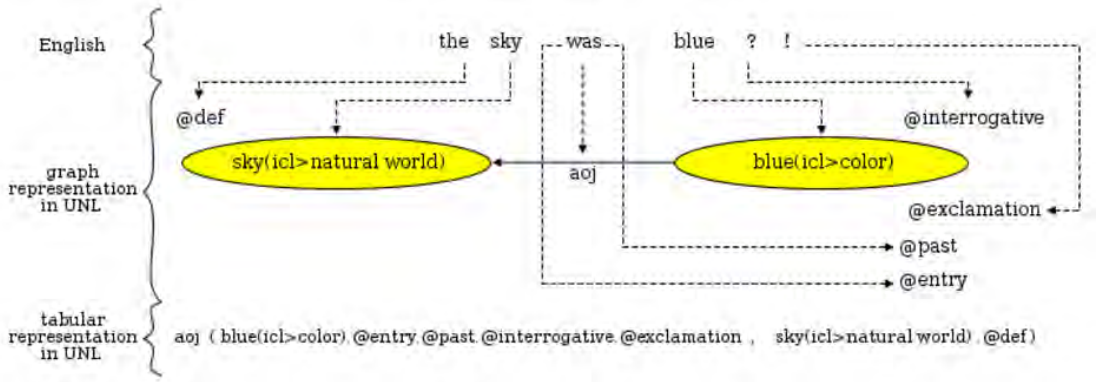


*Figure 2.19.* Example of UNL (Wikipedia, 2005)

Ali et al. (2008) worked on morphological analysis in Bangla. This work presented some rules to develop the morphological analysis of simple and compound words of Bangla that can be used to construct the dictionary of Bangla-UNL. This dictionary is able to be used to translate Bangla sentences to UNL documents.



Boguslavsky et al. (2000) presented the NLP modules that work on a multifunction NLP environment and interface with UNL. This module is involved with machine translation, natural language interface to SQL type database, the system of synonymous paraphrasing of the sentence, syntactic error correction tool, computer-aided language learning tool, tree bank workbench, and a new UNL converter. Figure 2.20 shows the layout of the machine translation (MT) module.

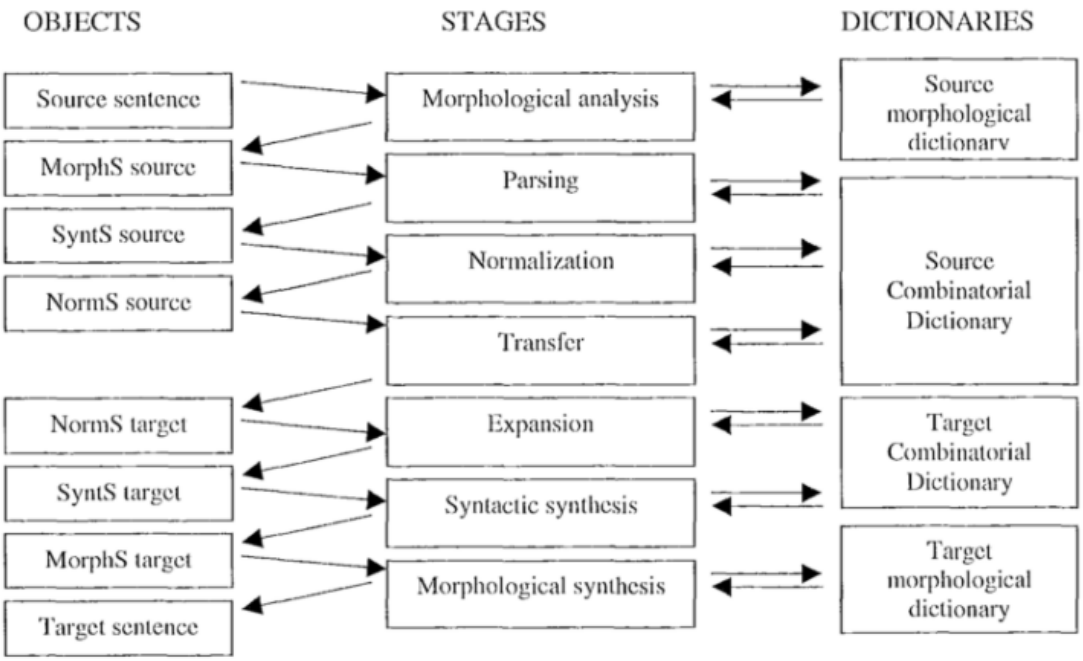


*Figure 2.20.* Layout of MT module (Boguslavsky et al., 2000)

Choudhary and Bhattacharyya (2002) presented the text clustering by using the feature that is generated from UNL. This work used the Self Organizing Maps (SOM) (Kohonen, 2001) to learn the feature of the document that extract from UNL. The semantic is used to cluster rather than a bag of words and performs better than other algorithms.

Ripon et al. (2014) developed the analysis tool for UNL-based knowledge representation. This tool provided a graphical tool that represented a visual of semantic data represented in the native text. Figure 2.21 shows the step of tools to convert from native language to UNL graphical representation.

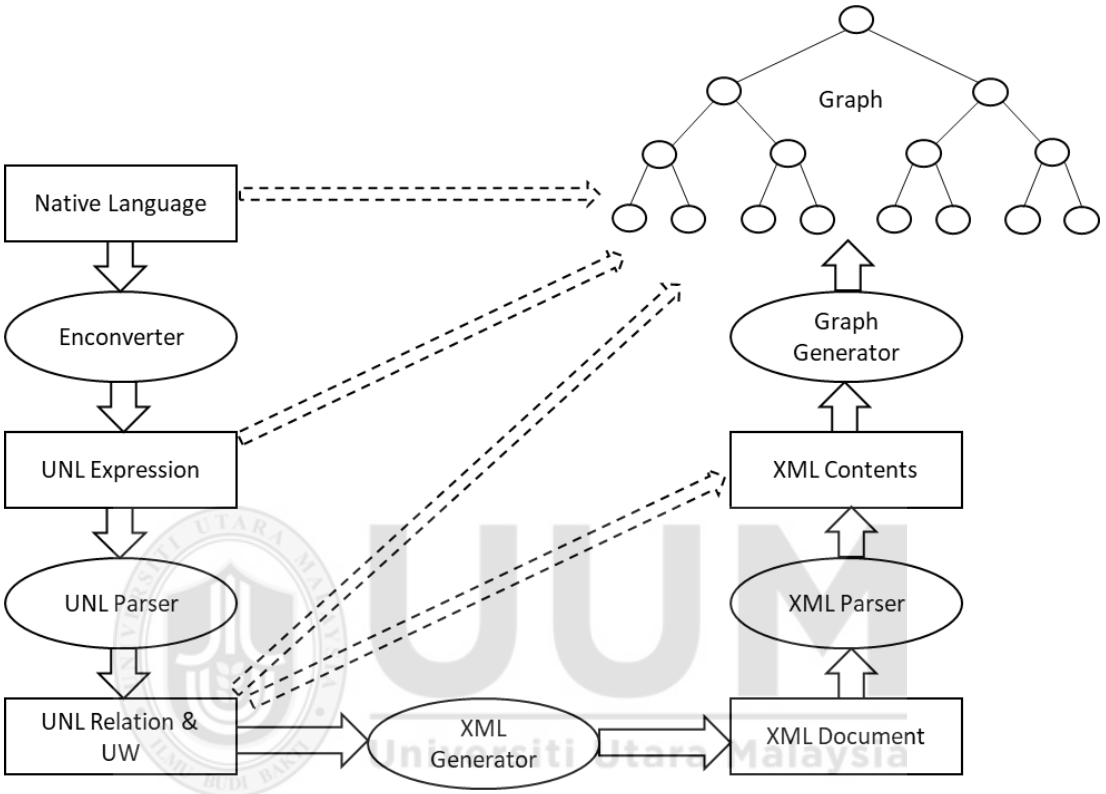


*Figure 2.21.* Step of tools to convert from native language to UNL graphical representation (Ripon et al., 2014)

Mridha et al. (2014) worked on resolving the semantic problem of phrases in the Bangla language. The ambiguity of root words in the Bangla language is resolved by semantic analysis and constructing a suitable universal word for the Bangla language.

The UNL is widely used in machine translation. The building of UNL must follow the specification of the UNL syntax. The rule to construct the UNL from a sentence needs

the mapping of universal words with Thai surface words. There is still lacking in the mapping of universal words to Thai surface words and it is a huge work to be done.

### 2.6.3 Natural Language Annotation

Natural language annotation (Katz, 1997; Katz et al., 2002; Lin et al., 2002) is used in the START question answering system at MIT artificial intelligence laboratory. The English sentence is formulated with embedded ternary expressions (T-expressions). This representation has 3 parts: subject, relation, and object. The answer can be extracted by reasoning from T-expressions and then constructed into natural language for the user. Figure 2.22 shows the example of T-expressions.

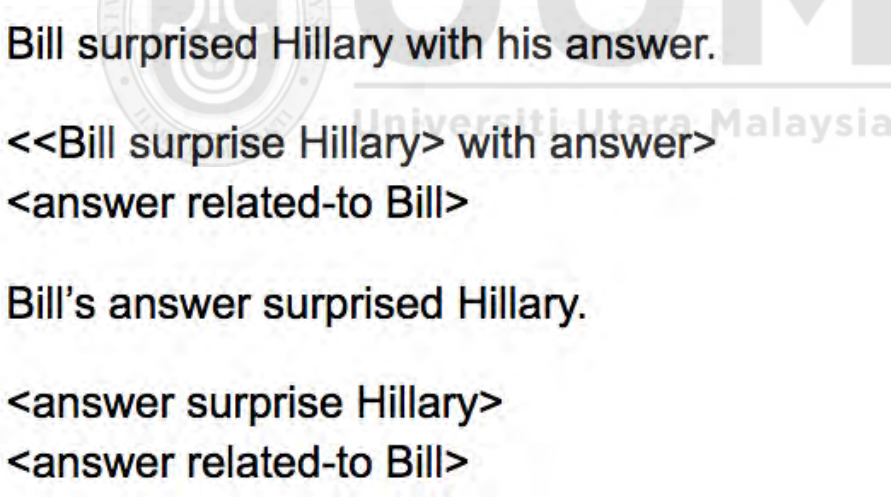


Figure 2.22. The example of T-expression (Katz, 1997)

Natural language annotation cannot create a complete semantic. The question answering system that needs to process a complex natural language process cannot be

done well with this representation. This representation is more suitable for the specific domain that does not need complex semantic analysis.

### 2.6.4 Semantic Frame Representation

The Berkeley FrameNet project (Baker et al., 2015; Fillmore & Baker, 2001) is research that produced rich linguistic resources that contain a frame-semantic description of English lexical items. The conceptual structure is constructed to take the word meaning and the participant in frame semantics. Frame semantics have a lot of properties of stereotyped scenarios that are expected to contain the events to occur and states to obtain from the English sentences and stored as a knowledge-based. Frames also can be inherited elements from their parent. Figure 2.23 shows the example of frame semantics and its heritance.



```
frame (CommercialTransaction)
frame-elements{BUYER, SELLER, PAYMENT, GOODS}
scenes (BUYER gets GOODS,
SELLER gets PAYMENT)

frame (RealEstateTransaction)
inherits (CommercialTransaction)
link(BORROWER = BUYER, LOAN = PAYMENT)
frame-elements{BORROWER, LOAN, LENDER}
scenes (LOAN (from LENDER) creates PAYMENT,
BUYER gets LOAN)
```

Figure 2.23. The example of frame semantics and its heritance (Lowe et al., 1997)

Das et al. (2010) proposed a computational and statistical model for frame-semantic parsing. This work aims to predict a frame-semantic representation by using a probabilistic framework that cleanly integrates the FrameNet lexicon and training data. Figure 2.24 shows the related to the CAUSE_TO_MAKE_NOISE frame and MAKE_NOISE frame.

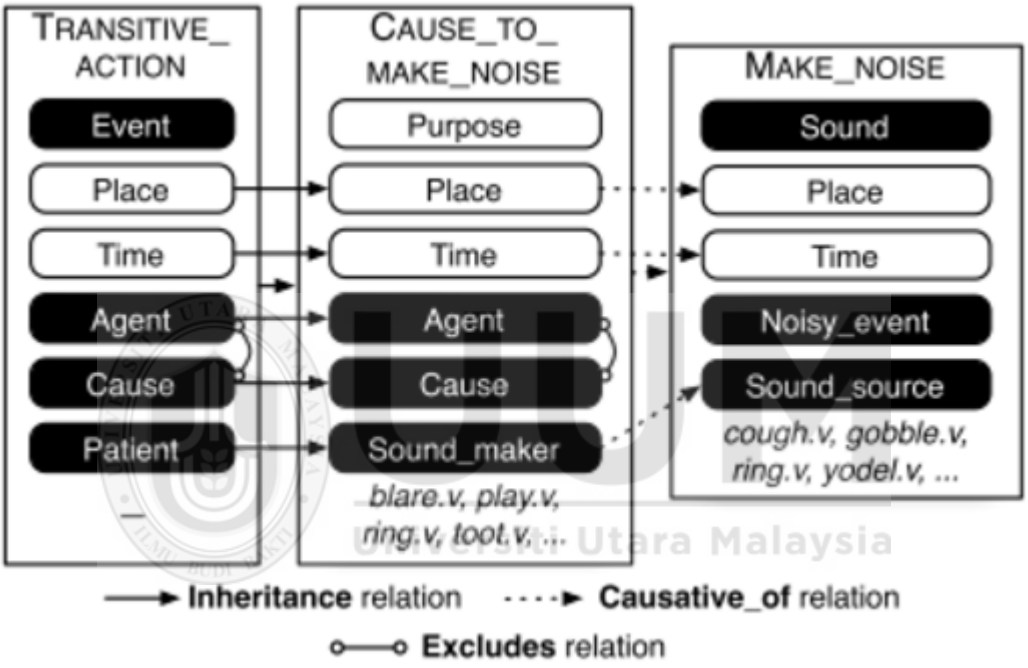


*Figure 2.24.* The related to the CAUSE_TO_MAKE_NOISE frame and MAKE_NOISE frame (Das et al., 2010)

Lestari and Nugraha (2017) used the frame-based approach on a speech-based QA system. This QAS is intended to provide information about scheduled routes and schedules for the departure of the Jakarta Metropolitan Are Commuter Line. In the frame-based approach, the dialogue that occurs between the system and the user is analogous to the process of filling in information on a form that is not visible to the user. The form is represented as a frame.

In summary, all knowledge representation is mostly good for explicit factoid answers. For the knowledge that is imprecise and needs specific computation, knowledge representation like UNL, RDF, and text annotation cannot handle it. Frame representation can contain any kind of information such as symbolic, numerical, function, and even frame. knowledge piece, that is imprecise, needs specific computation, and is able to form of function. Frame representation should be a good option to hold the imprecise knowledge.

## 2.7 Question Answering System

A question answering system (QAS) is the system that finds or generates the exact answer to the users by giving the question in natural language with accuracy, usefulness, and precision. The question answering system is able to solve the problem for the users by giving the exact answers that users need. The answer from the question answering system can be fact, list, definition, reason, procedure, etc. The question answering system consists of many components such as morphological analysis, syntactic analysis, semantic construction, question classification, and then answer extraction.

Question answering system research, in early time, started in the closed domain and closed corpus. BASEBALL (Green et al., 1961) is a database-oriented question answering system that answers the questions associated with results, locations, and dates of a baseball game. The question analysis is scoped in the closed and known domain of a baseball game and then the system identifies the answer from the database.

LUNAR (Woods, 1973) is a question answering system that provides information about lunar geology that gathers from Apollo lunar exploration. This system used English as a query language by transforming the question into a formal representation of the meaning of the query to the system. The answer is generated for the user after the query process.

BASEBALL and LUNAR both are good systems in their closed domain. However, there is a very limit to the question type that can be used in the system. Data in the database is acquired by humans and manually maintained.

The open domain question answering system that used unstructured data sources is started by TREC in 1999 (Voorhees, 2001, 2003). The first TREC provides a list of 200 questions and a document collection. The document collections, questions, and answer evaluations are increasing in size and complexity for more challenges every year.

Burger et al. (2001) presented a road map paper to guide research in question answering system that enables meaningful and useful capabilities if they follow the following standards:

> **Timeliness:** When the question is posted to the system, the answer should be provided in real-time even if the system is accessed by a thousand users.

**Accuracy:** The precision of the question answering system is extremely important. The incorrect answers are worse than no answers.

**Usability:** Knowledge in question answering systems should be tailored to the specific needs of a user. The question answering system must be able to mine answers regardless of the data source format and must deliver the answer in any format desired by the user.

**Completeness:** Complete answers to a user's question are desirable. Answer fusion in coherent information is required. The generation of the complete answer must rely on implicatures, due to the economic way in which people express themselves and due to the data sparseness.

**Relevance:** The answer to a user's question must be relevant within a specific context.



The evaluation criteria for question answering system are proposed to evaluate the answer by Breck et al. (2000) that there are six following criteria:

**Relevance:** The answer must be relevant and solve the given question.

**Correctness:** The answer must be correct and precise for the given question.

**Conciseness:** The answer should not contain irrelevant information to the given question.

**Completeness:** The answer should be complete and comprehensive to the given question.

**Coherence:** The answer should be coherent and be simply a phrase or sentence that makes it easier to understand for the users.

**Justification:** The answer should provide sufficient context to give more trust and confidence to the users.

The criteria, that is objective and can be numerical, could be easier to define the evaluation method. Correctness and completeness can be determined by the precision and recall evaluation method. The criteria that are more subjective can be evaluated by a human expert.



Semantic-based and knowledge-based question answering system (Höffner et al., 2017; Wang et al., 2015; Zhou et al., 2016) is more interested in the question answering system research field to improve the precision of the answer to the users.

CubeQA (Höffner & Lehmann, 2014; Konrad et al., 2016) is a semantic question answering system using multi-dimensional statistical linked data with RDF datacube vocabulary. The CuteQA algorithm converts a natural language question to a SPARQL query using a linear pipeline to generate the result set containing the answer. Figure 2.25 shows the pipeline of the CubeQA system.

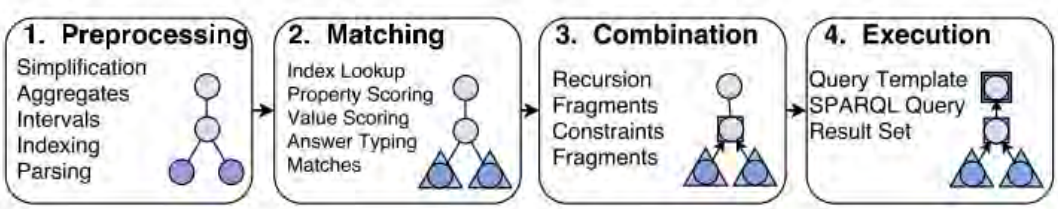


*Figure 2.25.* CubeQA pipeline (Konrad et al., 2016)

DEV-NLQ (Frost et al., 2014) presented the expression of the lambda calculus that worked on an event-based triple store using only triple-based retrieval operations. DEV-NLQ claims this question answering system is able to solve chained, arbitrarily-nested, complex, preposition phrases.

Freitas and Curry (2014) introduced a distributional compositional semantic model which is used as the central element for the construction of a vocabulary-independent Natural Language Interface (NLI) for Linked Data. The construction of a distributional semantic model is based on the extraction of co-occurrence patterns from large corpora, which defines a distributional semantic vector space. The distributional semantic vector space uses concept vectors to semantically represent data and queries, by mapping datasets elements and query terms to concepts in the distributional space. Figure 2.26 shows high level components diagram of the vocabulary independent query approach and distributional inverted index structure.

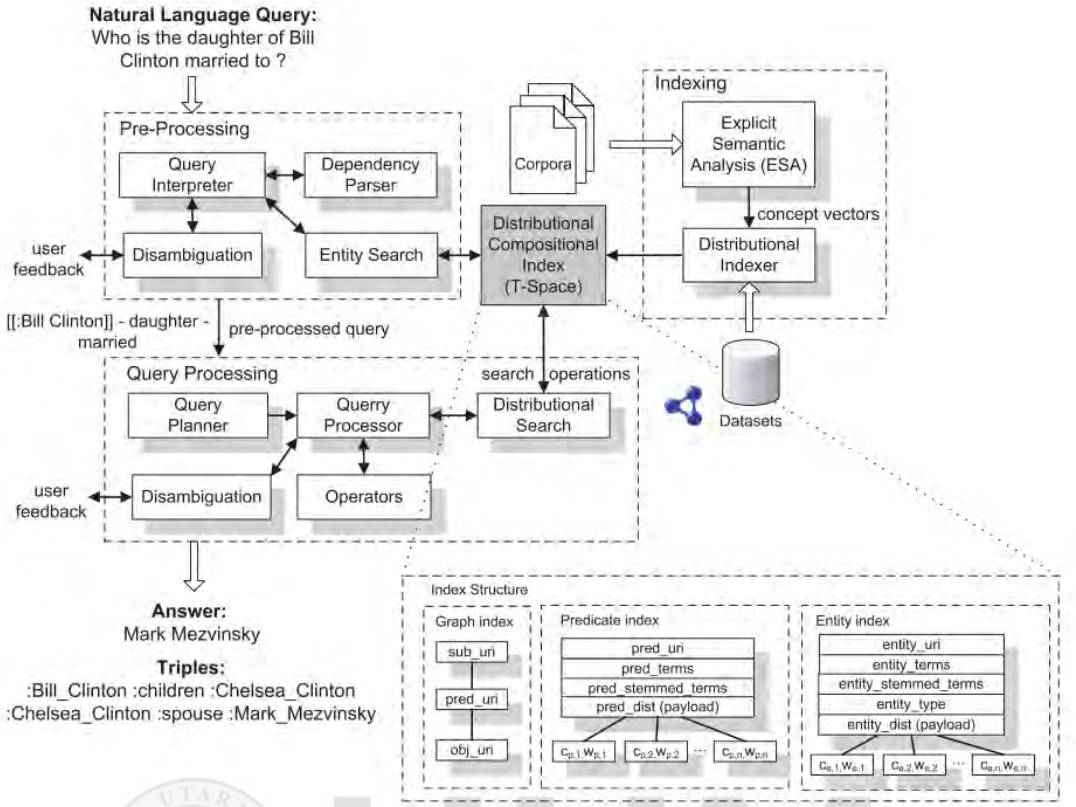


*Figure 2.26.* High-level components diagram of the vocabulary independent query approach and distributional inverted index structure (Freitas & Curry, 2014)



QuASE (Sun et al., 2015) is an open domain question answering system based on web search and the Freebase knowledge base with three stages of work. First, QuASE uses entity linking, semantic feature construction, and candidate ranking on the input question. Then, it selects the documents and according sentences from a web search with a high probability to match the question and presents them as answers to the user. Figure 2.27 shows QuASE framework.

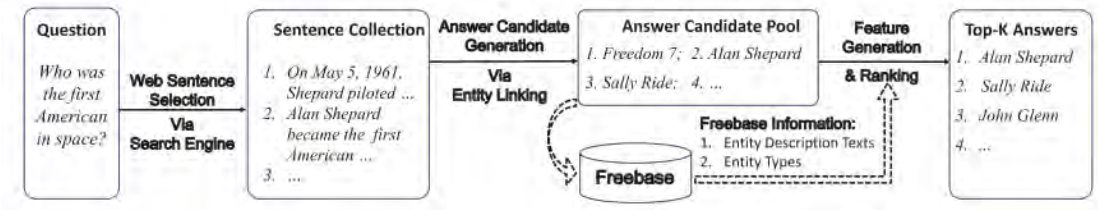


*Figure 2.27.* QuASE framework (Sun et al., 2015)

WabiQA (Noraset et al., 2021) is the open domain question answering system using Thai Wikipedia as the knowledge source. The WabiQA system consists of 2 parts: the document retriever and document reader. After the user gives the question in natural language form, the document retriever finds the candidate document that possible contains the answer by ranking the document based on the relatedness with BM25F scoring function (Robertson, Zaragoza, et al., 2009) between the question and the document. Then, the document reader locates the possible answer from the top-ranked document and ranks the answer by a confidence score for the user. Figure 2.28 shows the WabiQA system architecture.



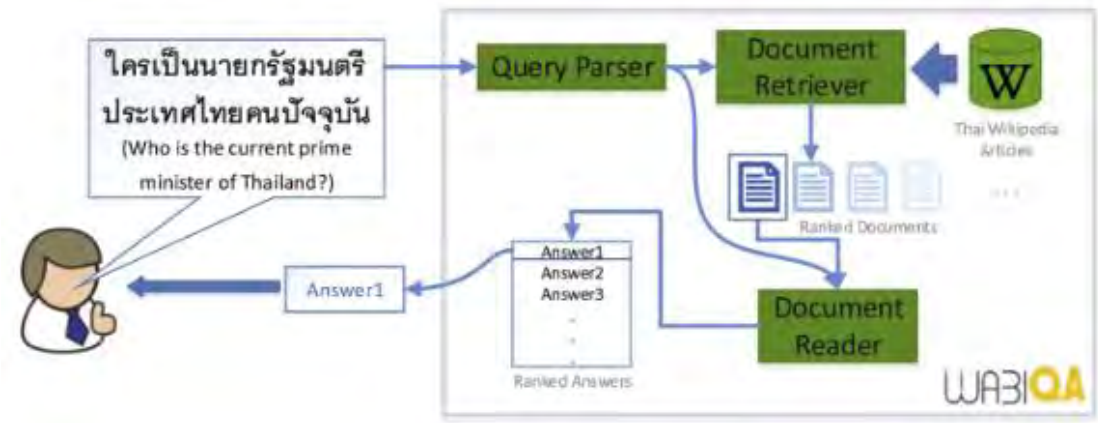


*Figure 2.28.* WabiQA system architecture (Noraset et al., 2021)

Classification of the question answering system can be classified in many aspects: content-based, data source based, language paradigm based, question type based, the approach based on question analysis and source document, and technique based (Lebedeva & Zaitseva, 2014; Mishra & Jain, 2016).

### 2.7.1 Content Based

A question answering system can be developed by the requirement of the application that is related to the scope of the content. Some applications handle the information on the general topic and some applications handle specific information or restricted topic. By content-based, question answering systems can be classified into the general or open domain, and restricted or closed domain.

**General or Open Domain:** In the open domain, the question answering system is dealing with a large set of document collections and provides the answer to any questions with domain independence. In most cases, the open domain question answering system needs general ontology and world knowledge in its methodology. Questions are mostly asked by casual users.

**Restricted or Closed Domain:** A restricted domain question answering system provides the answer to fixed topics and deals with some specific questions. Domain-specific ontology and terminology are used in question answering systems. Question patterns are very limited used, however, the quality of the answer can be more accurate and good elaborate. Questions are mostly asked by experts and specific users.

### 2.7.2 Data Source Based

Data sources are an essential part of the knowledge acquisition process for making the knowledge-based of question answering system. The source of data can be a factor to make the complexity of the methodology in question answering system. The data source of the question answering system can be classified as the structured data source, semi-structured data source, and unstructured data source.

**Structured Data Source:** The structured data can be from a database or structured document. Similar data entities are collected in the same structures and store the same attributes. The description of the structure of the entity that is collected in a unit is called schema. The matching of query in the structure data source is exact and accurate. Mostly, a structured data source question answering system is used in a restricted domain.

**Semi-structured Data Source:** The semi-structured data is data that there is no such partition between stored data and the schema. It provides a flexible format for making data exchange between different types of databases. XML is an example of semi-structured data that can be used in question answering systems. However, it is intensive labor to build the data source.

**Unstructured Data Source:** Unstructured data is data that can be of any type. Free text is an example of an unstructured data source. Dealing with unstructured data, question answering system requires the use of natural language processing and information retrieval techniques intensively to analyze and find the precise answer. Normally, unstructured data is the data source of open domain question answering systems and is used by casual users.

### 2.7.3 Language Paradigm Based

Classification based on language paradigm can be divided into 3 paradigms: Monolingual, Cross-lingual, and Multilingual question answering systems.

**Monolingual:** A monolingual question answering system is a question answering system that processes the data sources, queries the data, and finds the answer expressed in one language. Most question answering systems are working on a monolingual paradigm.

**Cross-lingual:** Cross-lingual question answering system is a question answering system that processes the data source in the first language and is able to query the data and find the answer from the second language by translating it to the first language before. This paradigm can start from monolingual and then uses the language-translated module to convert the query and the answer between two languages.

**Multilingual:** A multilingual question answering system is a question answering system that processes the data sources in the first language and is able to query the data and find the answer from other languages by translating it to the first language before. The knowledge can be represented in language-independent representation and query the data by translating the other language to a language-independent query before. The number of languages that the system supported depends on the language analyzed module to analyze the question and generate the answer.

### 2.7.4 Question Type Based

The question answering system is able to work on certain question types. The answer from different question types uses different processes and methodologies to query and extract. The question types can be listed as factoid, list, hypothetical, causal, and confirmation.

**Factoid:** The factoid question type is the question that is a simple question based on fact and requires an answer in a word, a short phrase, or a short sentence. These questions are involved with the "what, when, where, who" words. The answer in factoid is about the definition, the object, and the property of the object. The most answer of factoid is a general named entity, therefore, the named entity extraction is an essential part to identify the answer from data sources.

**List:** The list question type is the question that requires a list of named entities or facts in the answer. The threshold value for the quantity of the entity has to fix to pleasure the users. The technique that applies successfully in factoid questions is able to work with list questions with good accuracy.

**Hypothetical:** A hypothetical question is a question that asks for information related to any hypothetical event. The answers are subjective and no specific correct answers to these questions. Question answering systems need knowledge retrieval techniques to generate the answer by using intensive natural language processing techniques.

**Causal:** A causal question is a question that requires the explanation of the object or entity. These questions are involved the “how, why” word. The advanced natural language processing technique is required to analyze the text in the data source at the pragmatic and discourse level to extract and generate the answer. The output of the causal question is the explanation of the reason or procedure for the event or manufacture.

**Confirmation:** A confirmation question is a question that requires an answer in form of Yes or No. The inference system is an essential part of world knowledge and commonsense reasoning to generate the answer.

### 2.7.5 Approach Based on Question Analysis

Question analysis is an important part to derive the correct answer. Different question types require different methodologies to obtain the focus of the question and then the exact answer. There are three approaches for making the question analysis: statistical-based approach, pattern-based approach, and hybrid approach.

**Statistical Based Approach:** A statistical-based approach is a data-driven approach that uses quantitative relations to discover statistical relations existing in questions and documents. Utilization of probabilistic modeling, linear algebra, and information theory is used to analyze the relationship. A large data collection is needed to correct statistical learning that could produce a promising result.

**Pattern Based Approach:** A pattern-based approach is an approach that combines the use of linguistic rules and human knowledge in the information retrieval process. Predefined patterns are built for the question to extract the answer by performing on pattern matching technique. The pattern could be a lexico-syntactic or a lexico-semantic pattern. The training data set is less required to build the pattern. However, expert and domain knowledge is much required.

**Hybrid Approach:** A hybrid approach is a combination of the usage of a statistical-based approach together with a pattern-based approach. The requirement for a large data source could be minimized by a hybrid approach. However, patterns are still needed an expert to build to extract the answer.



### 2.7.6 Technique Based

Question answering system process the data source, query the data and generate the answer that could be done by various techniques. There are different levels of natural language processing needed in different question types and different kinds of answers. Question answering system that is classified by the technique: data mining, information retrieval, natural language understanding, and knowledge retrieval and discovery.

**Data Mining:** A question answering system based on data mining works by using a bag of words for searching the answer and replying to the short answer to the users. The data source mostly is the structure data or database.

**Information Retrieval:** A question answering system based on information retrieval works by using a bag of words for searching the answer and reply the relevant data as the answer to the users. The answer was mostly retrieved from the text document collection.

**Natural Language Understanding:** A question answering system based on natural language understanding works with a bag of concepts for searching for the answer and reply the well-described answer to the users. The knowledge-based of question answering system is created from text document collection and represented into knowledge representation by using the analysis of syntactic and semantic information.

**Knowledge Retrieval and Discovery:** A question answering system based on knowledge retrieval and discovery works by using a bag of knowledge for searching the answer and replying a new knowledge and reasonable answer to make the user understand. The pragmatic level of the natural language process is needed in the question answering system process to analyze the question and construct the knowledge for the users.

Mishra and Jain (2016) described the question answering system based on techniques that are shown in table 2.4.

Table 2.4

*Classification of QAS based on techniques (Mishra & Jain, 2016)*

| Aspects | QAS based on data mining | QAS based on information retrieval | QAS based on natural language understanding | QAS based on knowledge retrieval and discovery |
|---|---|---|---|---|
| Searching | Searching for factual data | Querying for factual information | Querying for information could be subjective opinionated or fact based | Understanding knowledge, creating knowledge, and searching for useful correct answers |
| Matching | Exact | Best match | Best match | Best correct match |
| Technology | Artificial intelligence and database | Information retrieval and natural language processing | Natural language processing and understanding | Natural language understanding, knowledge acquisition, mining |
| Knowledge source | Data base | Syntactic web | Syntactic and pragmatic web | Semantic and pragmatic web |
| Models used in retrieval process | Bag of word | Bag of word | Bag of concepts | Bag of knowledge |
| Reliability | Very good as the schema is designed by domain expert | Fewer lots of fake information is seen on the web | Less | Good |

A question answering system based on information retrieval gives an answer in good condition. However, a better answer could be by merging and validating answers from various sources of knowledge. Data from various textual sources will be processed and transformed into knowledge-based. Question answering systems based on natural language understating should be the next step.

According to some comprehensive surveys (Abbasiantaeb & Momtazi, 2021; Antoniou & Bassiliades, 2022; Pereira et al., 2022; Zhu et al., 2021), the question answering system also can classify by technique in the knowledge intensive and data intensive view. Figure 2.29 shows the classification of the question answering system based on knowledge and data intensive view.

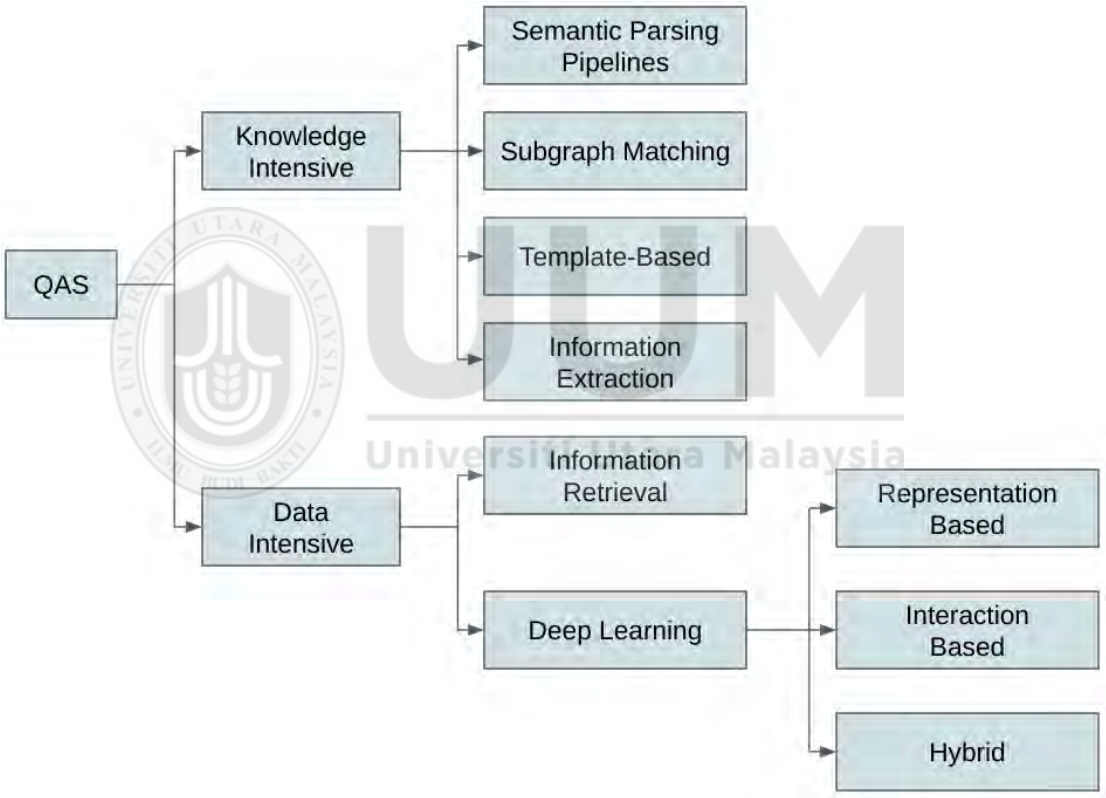


*Figure 2.29.* Classification of QAS based on knowledge and data intensive view

From Figure 2.29, the knowledge intensive technique can be semantic parsing pipelines, subgraph matching, template-based, and information extraction. The semantic parsing pipeline is the use of a semantic parser to convert the question into

the target query system and find the answer from a database or any knowledge base. The subgraph matching is the building of the query subgraph using a semantic tree. The template-based is the use of a template that matches the knowledge representation in the knowledge base and then extracts the exact answer. Information extraction is the use of some form of a deep neural network to identify the answer from the knowledge base. The data intensive technique can be information retrieval and deep learning. Information retrieval is the technique that depends on the specific domain trying to identify the piece of answer from text data. Deep learning is a new technique that uses of the deep neuron network to identify the piece of the answer from the text data.

## 2.8 Question Classification

Question Classification is a component of a question answering system that classifies and identifies the question type. Question Classification is useful to prune out the irrelevant information in the answer extraction process. The accuracy of question classification can affect the performance of the question answering system.

Li and Roth (2002, 2006) presented the hierarchical classifier for question classification. This work predefined the hierarchical classifier and used the Winnow algorithm within the SNoW (Carlson et al., 1999) to learn the coarse and fine classifier by using the feature from syntactic and semantic information. Figure 2.30 shows the hierarchical classifier used in this work.

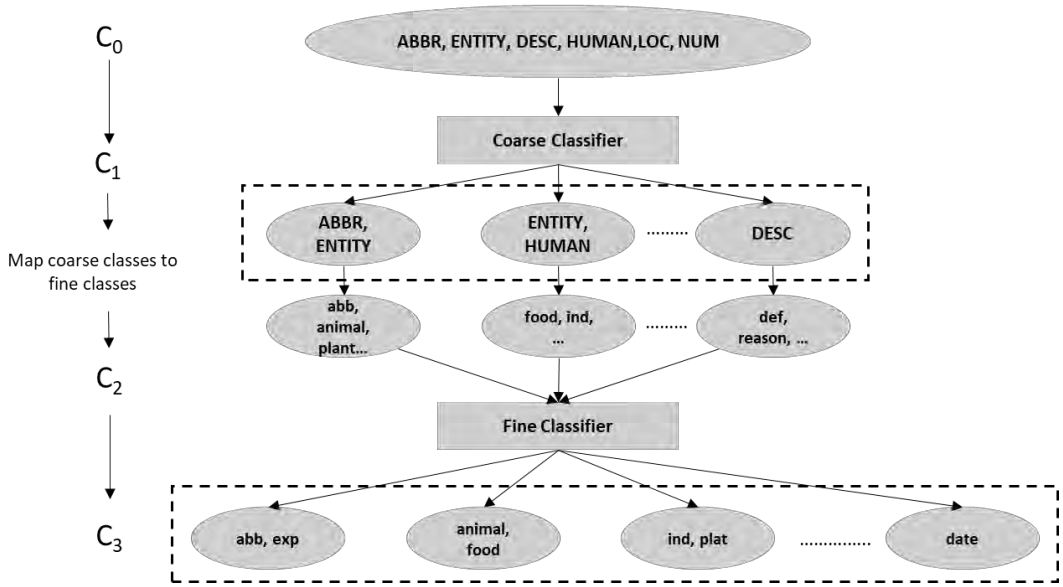


*Figure 2.30.* The hierarchical classifier (Li & Roth, 2006)

Zhang and Lee (2003) proposed the Support Vector Machine (SVM) for the question classifier. This work compared the various machine learning technique with SVM (Cristianini & Shawe-Taylor, 2000) to a learning classifier and used syntactic information as a feature. This work indicated that SVM is learning with higher accuracy than the other machine learning. Table 2.5 shows the coarse and fine-grained question categories that used in this work.

Table 2.5

*The coarse and fine-grained question categories*

| Coarse | Fine |
|---|---|
| ABBR | abbreviation, expansion |
| DESC | definition, description, manner, reason |
| ENTY | animal, body, color, creation, currency, disease/medical, event, food, instrument, language, letter, other, plant, product, religion, sport, substance, symbol, technique, term, vehicle, word |
| HUM | description, group, individual, title |
| LOC | city, country, mountain, other, state |
| NUM | code, count, date, distance, money, order, other, percent, period, speed, temperature, size, weight |



Day et al. (2007) presented an integrated genetic algorithm and machine learning for question classification in a cross-language question answering system in English-Chinese. The genetic Algorithm (GA) is used to select the near-optimal feature subset for the Conditional Random Fields (CRF) algorithm. After the feature is selected, the CRF question informer prediction model predicts the question informers. And then SVM question classification used the question informers as a key feature to learning the question classification. Figure 2.31 shows the proposed architecture that integrated genetic algorithm and machine learning approach for question classification in English-Chinese cross-language question answering.

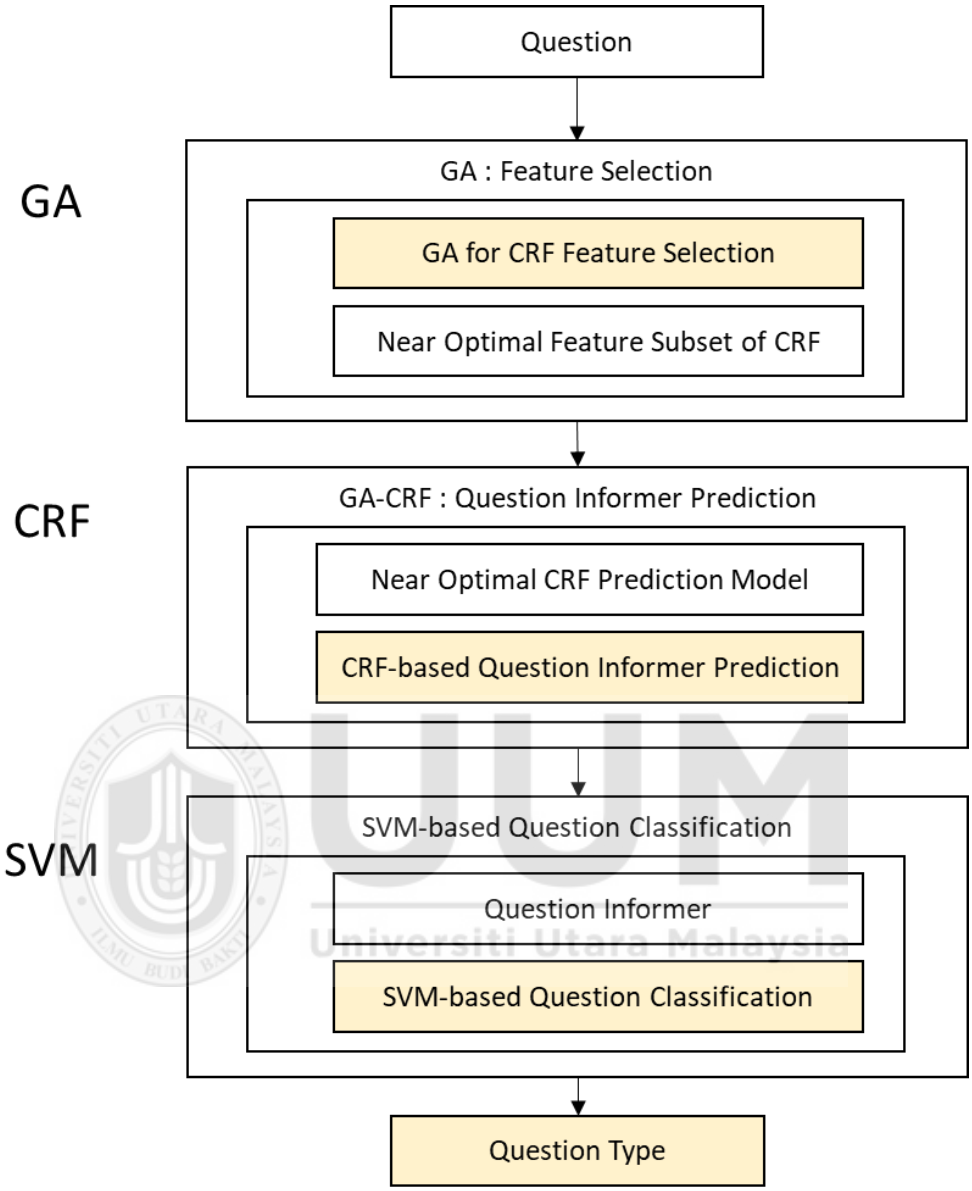


*Figure 2.31.* The proposed architecture that integrated genetic algorithm and machine learning approach for question classification in English-Chinese cross-language question answering (Day et al., 2007)

The question classification is needed in some methodologies that used the relation of the word in question and the document source to identify the piece of the answer. The

irrelevance paragraph can be pruned out by the question classification. However, for the deep semantic application, the question classification could be not important.

## 2.9 Fuzzy Factoid

A fuzzy factoid is the answer type of question answering system that return data involved with an object such as a person, organization, location, etc that is dealing with fuzzy linguistic property. The property of the object that is stated in natural language normally be fuzzy such as young, dark, warm, and property with range numerical. Knowledge representation that deals with fuzzy factoid have to be computable in fuzzy formalism. There are several topics in fuzzy that have concern in fuzzy factoid: Fuzzy sets, Fuzzy logic, and Fuzzy frame-based knowledge representation formalism.

**Fuzzy Sets:** Fuzzy sets are introduced by Zadeh (1965). A fuzzy set is a class with a continuum of grades of membership that range between zero and one. The formalism of a fuzzy set has been defined as a complement, union, intersection, etc.

**Fuzzy Logic:** Fuzzy logic (Zadeh, 1988) is the formal principle of approximate reasoning with precise reasoning viewed as a limit case. Fuzzy logic is able to express the imprecise meaning in natural language as elastic constraints on a variable and reasoning through the propagation of elastic constraints. Fuzzy logic may be viewed as an extension of multivalued logic.

**Fuzzy Frame-Based Knowledge Representation Formalism:** Tettamanzi (2003) presented the formalism of fuzzy frame-based knowledge representation

for representing imprecise knowledge. Knowledge in this formalism consists of three basic types of objects: knowledge element, linguistic values, and relation. Formalism has been described which is based on the combination of three concepts: frame-based knowledge representation formalism, unification, and fuzzy set theory.

The question answering system that can handle the fuzzy factoid can produce more precision and recall of the answer. The fuzzy value is the important attribute that frequently is the answer. The fuzzy factoid answer could be come up with the confidence value of the result to make the positiveness to the user.



## 2.10 Summary

This chapter reviews the component of the question answering system and technique. Thai morphological analysis is still active in the research area to improve the accuracy of word segmentation and named entity extraction. The corpus that exists is developed for a specific purpose with the specific annotate. That corpus may be not suitable for our works due to the different annotations and incomprehensive annotations. The new corpus development is needed to serve the study of our work. The Anaphora and ellipsis resolution is needed for making the precise QAS. The anaphora resolution for nominal anaphora is still not done in Thai morphological analysis. Then, the complete ellipsis and anaphora resolution for all types of anaphora is still needed.

Knowledge representation tends to use RDF to question answering knowledge base whereas the natural language annotation is used in practical question answering systems such as the START question answering system. However, RDF and natural language annotation cannot handle fuzzy factoid answers. Frame-based knowledge representation potentially deals with a fuzzy factoid.

Some question answering systems for the Thai language use data-intensive and pre-defined knowledge from structured sources that the system can work on a restricted area of the domain and not perform an automatic knowledge extraction in open domain or factoid question. Semantic-based question answering system with knowledge extraction from textual sources could be a good option to develop a precise question answering system.

# CHAPTER THREE

# RESEARCH METHODOLOGY

This chapter describes the methodology of the research. The corpus preparation is the first step and the next are morphological analysis, ellipsis and anaphora resolution, semantic parser, knowledge representation, and answer extraction.

## 3.1 Stage of Development

This research consists of six stages of work. The first is corpus preparation. The second is morphological analysis. The third is ellipsis and anaphora resolution. Fourth is the semantic parser. The fifth is Answer Extraction and the last is evaluation. Figure 3.1 is shown the overview of the research design stages.

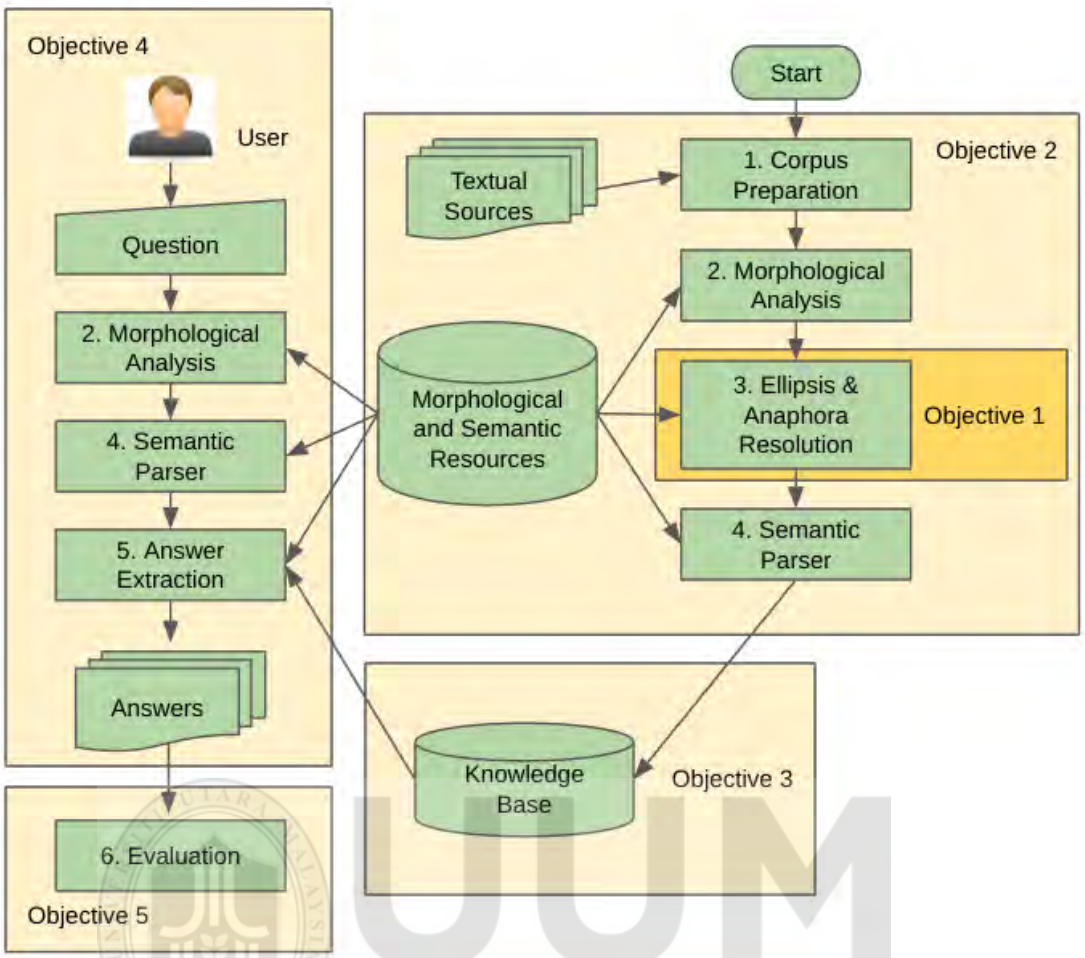


*Figure 3.1.* Research design stages

Corpus preparation is the stage of collecting and cleaning the documents that used to be the source of knowledge and testing in the morphological analysis process. The stage of morphological analysis is to implement the morphological process for the Thai language such as Thai word segmentation, Thai named entity identification, and Thai EDU segmentation. The stage of ellipsis and anaphora resolution is the process that resolves the anaphoric reference and textual ellipsis to complete the sentence for semantic construction. The stage of the semantic parser is to implement the semantic parser that analyzes the complete syntactic information and then formulates it into the knowledge representation for the question answering system. The stage of answer

extraction is to implement the question analysis that identifies the question type and then extract the precise answer that matches the constraint in the question type from the knowledge base. The last step is to evaluate the question-answering system.

### 3.2 Corpus Preparation for Morphological Process

This stage is to prepare the corpus that is used in the morphological process and also to be a source of knowledge for the question-answering framework. Thai Wikipedia is considered to be one of the sources of corpus because it consists of a vast knowledge that is in form of Thai text (Mesgari et al., 2015). Thai Wikipedia consists of a lot of articles in the open domain unstructured text. This can be a very good source of corpus development in the open domain in this research. Figure 3.2 shows the example of the Thai Wikipedia page.

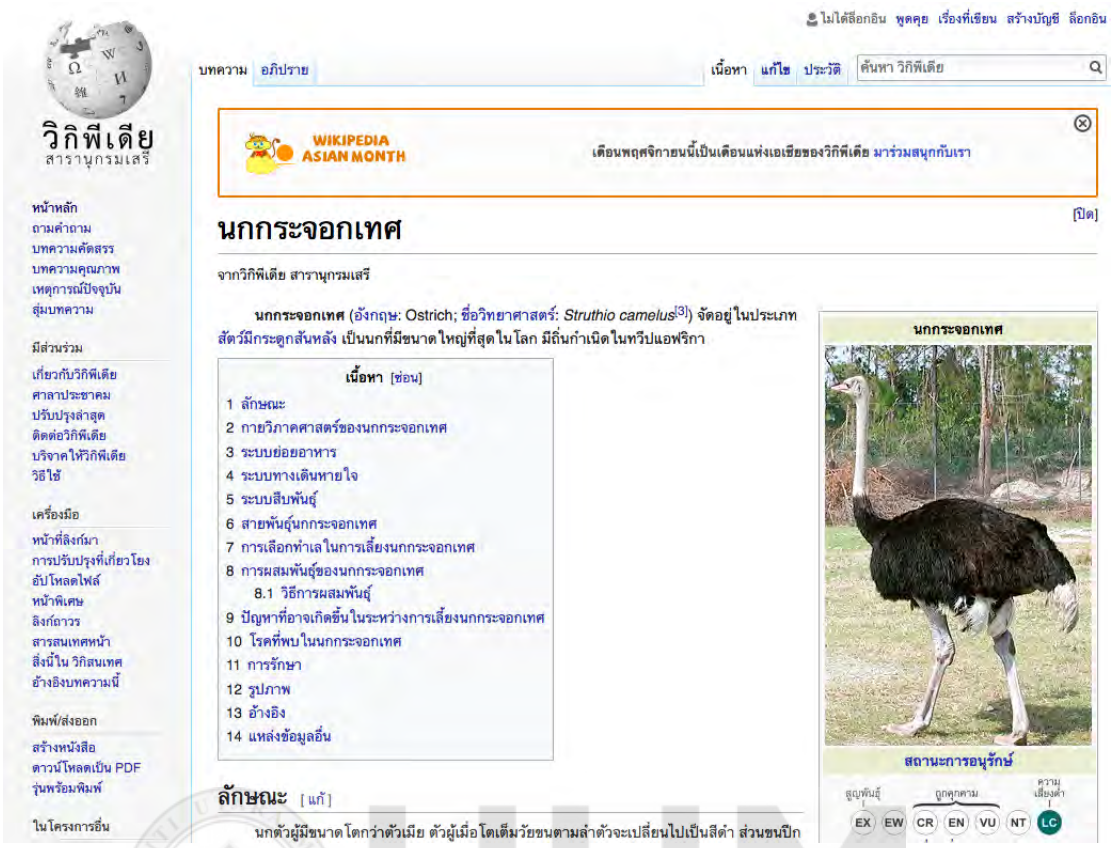


*Figure 3.2.* Example of the Thai Wikipedia page



Web Crawler is developed to gather the web pages from Thai Wikipedia into the database. After that, HTML tags will be removed in the background process. Figure 3.3 shows the example of HTML source.

```
<link rel="EditURI" type="application/rsd+xml" href="//th.wikipedia.org/w/api.php?action=rsd"/>
<link rel="license" href="//creativecommons.org/licenses/by-sa/3.0/"/>
<link rel="canonical"
href="https://th.wikipedia.org/wiki/%E0%B8%99%E0%B8%81%E0%B8%81%E0%B8%A3%E0%B8%B0%E0%B8%88%E0%B8%
AD%E0%B8%81%E0%B9%80%E0%B8%97%E0%B8%A8"/>
<link rel="dns-prefetch" href="//login.wikimedia.org"/>
<link rel="dns-prefetch" href="//meta.wikimedia.org" />
<!--[if lt IE 9]><script src="/w/load.php?
debug=false&amp;lang=th&amp;modules=html5shiv&amp;only=scripts&amp;skin=vector&amp;sync=1">
</script><![endif]-->
</head>
<body class="mediawiki ltr sitedir-ltr mw-hide-empty-elt ns-0 ns-subject page-นกกระจอกเทศ
rootpage-นกกระจอกเทศ vector-nav-directionality skin-vector action-view">        <div id="mw-page-
base" class="noprint"></div>
        <div id="mw-head-base" class="noprint"></div>
        <div id="content" class="mw-body" role="main">
            <a id="top"></a>

                            <div id="siteNotice" class="mw-body-content"><!-- CentralNotice -->
<div id="mw-dismissablenotice-anonplace"></div><script>(function(){var
node=document.getElementById("mw-dismissablenotice-anonplace");if(node){node.outerHTML="\u003Cdiv
class=\"mw-dismissable-notice\"\u003E\u003Cdiv class=\"mw-dismissable-notice-
close\"\u003E[\u003Ca href=\"#\"\u003Eปิด\u003C/a\u003E]\u003C/div\u003E\u003Cdiv class=\"mw-
dismissable-notice-body\"\u003E\u003Cdiv id=\"localNotice\" lang=\"th\"
dir=\"ltr\"\u003E\u003Cdiv class=\"mw-parser-
output\"\u003E\u003Cp\u003E\u003C/p\u003E\n\u003C/div\u003E\u003C/div\u003E\u003C/div\u003E\u003C
/div\u003E";}}());</script></div>
                        <div class="mw-indicators mw-body-content">
</div>
            <h1 id="firstHeading" class="firstHeading" lang="th">นกกระจอกเทศ</h1>
                                    <div id="bodyContent" class="mw-body-content">
                                    <div id="siteSub" class="noprint">จากวิกิพีเดีย สารานุกรมเสรี</div>
                                <div id="contentSub"></div>
                                                <div id="jump-to-nav" class="mw-jump">
                                            <a href="#mw-head">ป้ายบอกทาง</a>,
                    ไปที่:
<a href="#p-search">ค้นหา</a>
                </div>
                <div id="mw-content-text" lang="th" dir="ltr" class="mw-content-ltr"><div
class="mw-parser-output"><div>
<table class="infobox biota" style="text-align: left; width: 200px; font-size: 100%">
```

*Figure 3.3.* Example of HTML source

The Thai text will be processed and converted all symbols and space into a symbol tag. For example, a space character is converted to ”<space>”, a colon symbol is converted to ”<colon>”, a left parenthesis is converted to ”<left_parenthesis>”, and a comma symbol is converted to ”<comma>”. The overview of corpus preparation process is shown in Figure 3.4 and Table 3.1 shows the example of symbol tag.

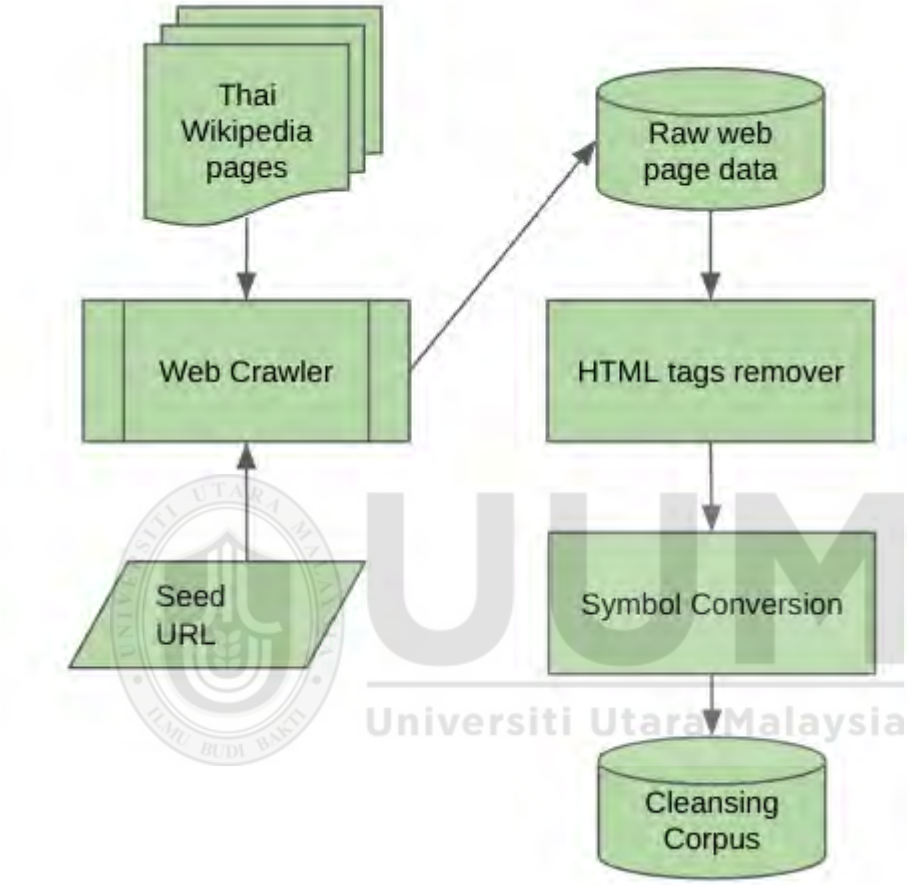


*Figure 3.4.* Overview of corpus preparation process

Table 3.1

*Examples of symbol tags*

| Symbol Name | Symbol Character | Symbol Tag |
|---|---|---|
| Colon | : | <colon> |
| Semi Colon | ; | <semi_colon> |
| Comma | , | <comma> |
| Left Square Bracket | [ | <left_square_bracket> |
| Right Square Bracket | ] | <right_square_bracket> |
| Left Parenthesis | ( | <left_parenthesis> |
| Right Parenthesis | ) | <right_parenthesis> |
| Space | | <space> |

In this process, some space characters, that are insignificant and ineffective, will be removed such as space before and after a number, and a space before the repeater sign. This cleaned corpus is the data source for training and experiment in the next process. The example of a clean corpus is shown in Figure 3.5.

นกกระจอกเทศเป็นสัตว์กินพืช<left_parenthesis>Herbivorous
<right_parenthesis>กระเพาะของนกจะแบ่งเป็น2ส่วน<space>คือ
<space>ส่วนที่เป็นกระเพาะบด<left_parenthesis>Gizzard
<right_parenthesis>เหมือนไก่<space>แต่ไม่มีกระเพาะพัก
<left_parenthesis>Crop<right_parenthesis>และส่วนที่สองเป็นกระ
เพาะแท้<left_parenthesis>Proventriculus<right_parenthesis>เหมือน
สัตว์เคี้ยวเอื้อง<left_parenthesis>Ruminant<right_parenthesis>บาง
ชนิดเช่น<space>โคและกระบือ<space>เป็นต้น

*Figure 3.5.* Example of a clean corpus

The other domain Thai corpus is also considered to be the source of knowledge. The documents will have a preprocessing to clean some unused information and reorganize its content before being sent to the morphological processing. The corpus in this process contains 18,248 words after finished.

### 3.3 Morphological Analysis

Morphological analysis is the process to analyze words and resolve some phenomena in the language. 3 processes in the morphological analysis consist of Thai word segmentation with Part of Speech (POS) tagging, Thai name entities identification, and Thai EDU segmentation. The cleansing corpus is processed by Thai word segmentation with POS tagging in the first step. Thai named entity identification is the second step to identifying the object in sentences. And Then Thai EDU segmentation is the last step to separate the paragraph in the source text into the Thai EDU segment. The overview of morphological analysis is shown in Figure 3.6.

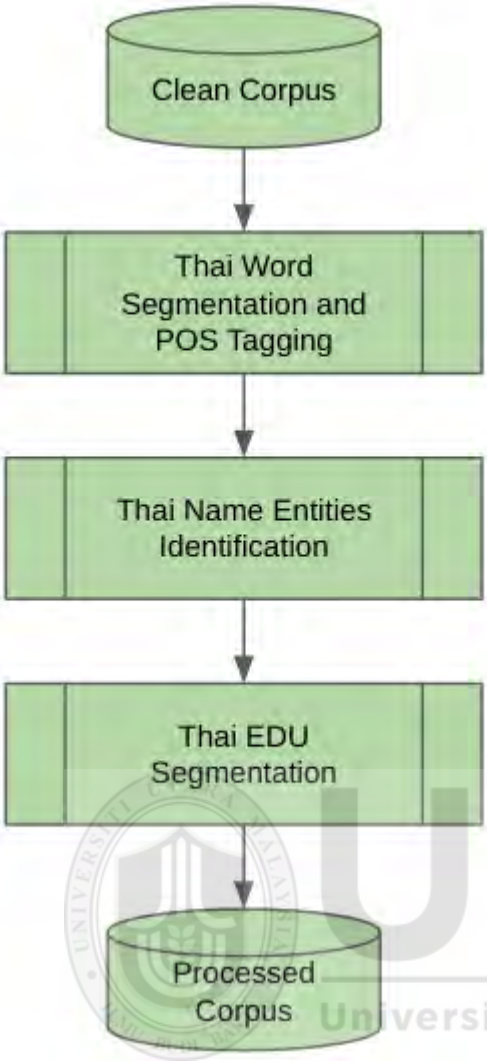


*Figure 3.6.* Overview of morphological analysis

### 3.3.1 Thai Word Segmentation and POS Tagging

Thai word segmentation is the first step in the morphological analysis for finding word boundaries in Thai text and also tagging POS in each word. The dictionary-based approach is used to find all possible combinations of sequences of words in each EDU. Conditional Random Field (CRF) algorithm (Lafferty et al., 2001; Peng et al., 2004) is considered to be used to make a decision that which sequences of words will be a better

sequence of words segmentation with POS tagging information. The examples of the POS tagging label are shown in Table 3.2.

Table 3.2

*Examples of the POS tagging label*

| POS Label | Description | Example Words |
|---|---|---|
| NCM | Common Noun | นก ลิง เศรษฐกิจ |
| NCA | Noun and Classifier for Attribute, Kind, and Group | สี อายุ ประเภท ชนิด |
| NNA | Noun for Amount | สิบ ร้อย พัน หมื่น |
| VRB | Transitive Verb | กิน มอง รัก |
| VRI | Intransitive Verb | ยิ้ม วิ่ง เดิน |
| VAT | Attribute Verb | ใหญ่ หนัก สวย |
| FVN | Prefix to transform Noun and Verb to be Noun | การ |
| NBO | Ordinal Number Word | แรก สุดท้าย ต่อไป |



In the training process for the CRF algorithm, the manually tagged corpus is needed. The selected corpus is tagged POS manually and then used for the training of the CRF algorithm. POS tag-set is designed for this work total of 45 tags within 10 groups of tag-set. The example of a tagged corpus is shown in Figure 3.7.

[นก]<NCM>[ตัวผู้]<NCA>[มี]<VRB>[ขนาด]<NCA>[โต]<VAT>
[กว่า]<PRP>[ตัวเมีย]<NCA><space>[ตัวผู้]<NCA>[เมื่อ]<SUB>
[โต]<VAT>[เต็มวัย]<ADV>[ขน]<NCM>[ตาม]<PRP>[ลำตัว]<NCM>
[จะ]<VAX>[เปลี่ยน]<VRB>[ไป]<VPT>[เป็น]<VPO>[สี]<NCA>
[ดำ]<VAT><space>[ส่วน]<SUB>[ขน]<NCM>[ปีก]<NCM>
[และ]<CON>[ขน]<NCM>[หาง]<NCM>[จะ]<VAX>[เป็น]<VRB>
[สี]<NCA>[ขาว]<VAT>[สวยงาม]<VAT>[มาก]<ADV><space>
[สำหรับ]<SUB>[ตัวเมีย]<NCA>[จะ]<VAX>[มี]<VRB>[ขน]<NCM>
[ตาม]<PRP>[ตัว]<NCM>[สี]<NCA>[น้ำตาล]<VAT>[เทา]<VAT>
[อ่อน]<ADV>

*Figure 3.7.* Example of a tagged corpus

Dictionary is developed for use in the word segmentation process. There are surface words with a possible POS type in the dictionary. A dictionary is developed by gathering words from the tagged corpus and also its POS. Dictionary is an essential resource to generate all possible segment sequences of the word segmentation for the word segmentation process. The example of dictionary is shown in Table 3.3.

Table 3.3

*The example of dictionary*

| **Word** | **POS** |
|---|---|
| นกพิราบ (pigeon) | NCM |
| อยู่ (live, stay, -ing) | VRB VPO VPT VPA VPR |
| อย่าง (thing, -ly) | FAV CLS |
| บาง (some, thin) | DQE VAT |
| สวยงาม (beautiful) | VAT ADV |
| กำลัง (power, -ing) | VAX NCM |
| ให้ (give, to) | VPO VRB VPR VPT VPA |



POS feature is needed for the CRF model to train and also to determine the best POS tag in word segment sequence. POS feature is the sequence pattern of POS of words that appear in corpus data. POS features are extracted from the tagged corpus and then collected in the database. After that, the CRF model is trained by using data from the tagged corpus and POS feature database to adjust the weight of the POS feature.

CRF is the probability of the label sequence that globally normalizes to avoid the label-bias problem and also provides the flexibility to use non-independent features (Lafferty et al., 2001). Let $\boldsymbol{l}$ be a label of POS sequence, $\boldsymbol{s}$ be a word segmentation, $\boldsymbol{f}$ be a feature function, and $\boldsymbol{\lambda}$ be the weight of the feature function. Then, we can define the score

function *score(l|s)* to determine the score of the POS sequence to the sentence as shown in Equation 3.1.

$$score(l|s) = \sum_{j=1}^{m} \sum_{i=1}^{n} \lambda_j f_j(s, i, l_i) \tag{3.1}$$

The conditional probability distribution *p(l|s)* for a linear-chain CRF can be defined as shown in Equation 3.2.

$$p(l|s) = \frac{\exp[score(l|s)]}{\sum_{l'} \exp[score(l'|s)]} = \frac{\exp\left[\sum_{j=1}^{m} \sum_{i=1}^{n} \lambda_j f_j(s, i, l_i)\right]}{\sum_{l'} \exp\left[\sum_{j=1}^{m} \sum_{i=1}^{n} \lambda_j f_j(s, i, l_i')\right]} \tag{3.2}$$



To estimate the weight of feature function $\boldsymbol{\lambda}$ for training data, Equation 3.3 can be defined where $\boldsymbol{\alpha}$ is a learning rate.

$$\lambda_i = \lambda_i + \alpha \left[ \sum_{j=1}^{m} f_i(s, j, l_j) - \sum_{l'} p(l'|s) \sum_{j=1}^{m} f_i(s, j, l_j') \right] \tag{3.3}$$

The features of learning for POS tagging are the POS sequence pattern that is extracted from the tagged corpus. There are 11 types of POS sequence pattern combination in this work that is shown in Table 3.4. The notation ***V*** is the POS in the position that is determined, ***X*** is the POS in the previous position that is determined, ***Y*** is the POS in the next position that is determined, ***W*** is a surface word, and then, none is the position that is not determined.

Table 3.4

*POS pattern rules*

| Feature Type | POS Pattern |
|---|---|
| 1 | none : X1 : V : none : none |
| 2 | X2 : X1 : V : none : none |
| 3 | none : none : V : Y1 : none |
| 4 | none : none : V : Y1 : Y2 |
| 5 | none : X1 : V : Y1 : none |
| 6 | X2 : X1 : V : Y1 : none |
| 7 | none : X1 : V : Y1 : Y2 |
| 8 | X2 : X1 : V : Y1 : Y2 |
| 9 | none : X1 : V : Y1 : none#W |
| 10 | none : X1 : V : none : none#W |
| 11 | none : none : V : Y1 : none#W |

After finishing training, the CRF model is ready to determine word segmentation and POS tagging. Raw text after the cleaning process is the input for the word segmentation process. By using a dictionary, all possibilities of word segmentation sequence are generated. Due to the time complexity, the word segmentation sequence will be generated limited to only 4 words each time. CRF is used to determine the best POS tag for each sequence by selecting the best probability. The best word segmentation sequence is selected by the best score. The first word in the best word segmentation sequence will be selected for the answer sequence. After that, the next word segmentation sequence will be generated and re-operated again until the last word is determined.

Words in the Thai language can be composed of more than one other word and become a new meaning. For example ”พื้นที่(area)” is composed of 2 words are ”พื้น(floor)” and ” ที่(at)”. Word segmentation by the CRF model tends to segment words into more chunks. Dictionary can be a useful resource to correct word segmentation (Nararatwong et al., 2018). To correct word segmentation, we apply POS pattern matching together with a dictionary. POS patterns are used to match and find the word in Dictionary with the POS target. After that, words with POS pattern is replaced by word with POS target from a dictionary. Table 3.5 shows the examples of the POS pattern for word correction.

Table 3.5

*Examples of the POS pattern for word correction*

| POS Pattern | POS Target Word |
|---|---|
| CON NCM | VAC |
| VAC NCM | VAC |
| NCM PRP | NCM |
| NCM PRL | NCM |

After the word correction process, there are some mistaken POS tagging that could be fixed. All POS tagging will be cleared and reconsidered by the POS re-tagging process. This process can improve the precision of POS tagging in the word sequence. All processes of Thai word segmentation and POS tagging are shown in Figure 3.8.

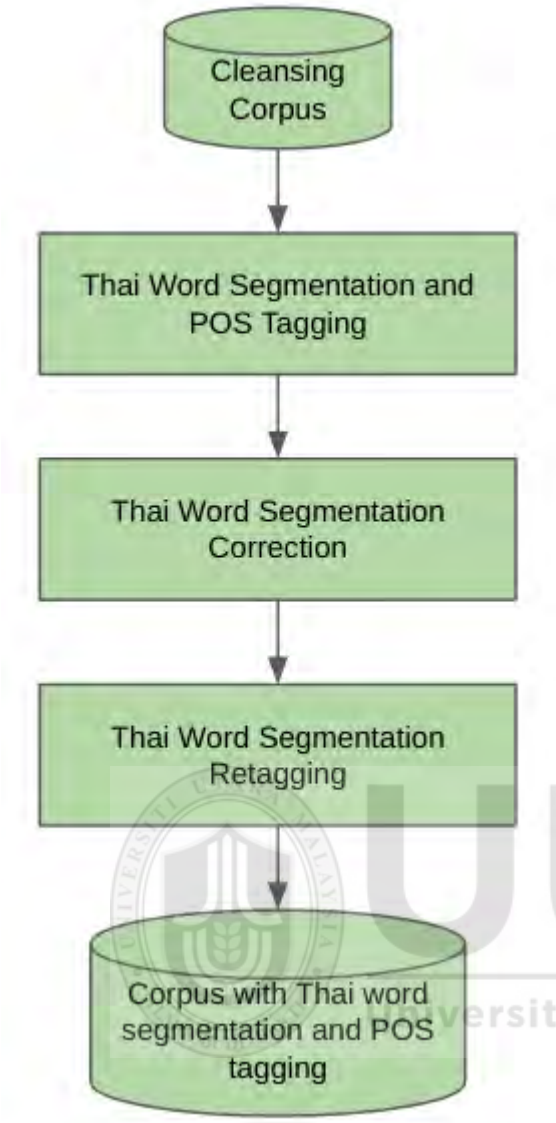


*Figure 3.8.* Processes of Thai word segmentation and POS tagging

### 3.3.2 Named Entities Identification

Thai named entity identification is the process to identify the named entity boundary and also identify the type of named entity. The template matching technique is used to identify the named entities in the text. POS tag and surface word are the component of a matching template for extracting named entities. Corpus is tagged manually to create the matching template. POS tag is the main component of the template and surface word

is an optional component. Figure 3.9 shows an example of the corpus with a manual named entity tagging.

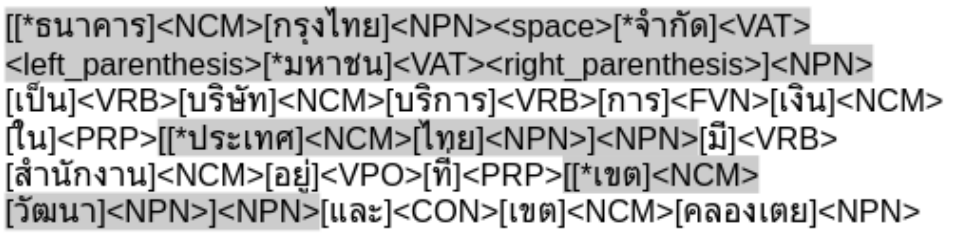


*Figure 3.9.* Example of the corpus with a manual named entity tagging

In figure 3.9, the surface word with the ”*” symbol means to indicate that the surface word is also used in matching. After that, the matching template is created by extracting it from the tagged corpus. Table 3.6 shows the examples of the named entity matching template.



Table 3.6

*Examples of the named entity matching template*

| **Pattern** | **Tag** |
|---|---|
| [เขต]<NCM>[*]<NPN><br>district | NPN |
| [ธนาคาร]<NCM>[แห่ง]<PRP>[ประเทศ]<NCM>[*]<NPN><br>bank of nation | NPN |
| [ธนาคาร]<NCM>[*]<NCM><space>[จำกัด]<VAT><br>bank limited | NPN |
| [บริษัท]<NCM><space>[*]<NPN>[*]<NCM><space>[จำกัด]<VAT><br>company limited | NPN |

In Table 3.6, the ”*” symbol in the matching template means to indicate that it is matching with any surface word. The matching algorithm works with one-to-one matching on words in the whole line of text. The size of the template is an important factor in matching. The longest template will be activated first and then the next longest.

### 3.3.3 Thai EDU Segmentation

Thai EDU segmentation is the process to analyze and segment Thai text in a paragraph into the EDU segment. In this work, we divided the Thai EDU segmentation into 4 steps: the Elementary Discourse Unit (EDU) segmentation by clue markers, the shallow parser, the EDU segmentation Segmentation by Syntactic Pattern, and EDU Reconstruction by rule-base. There are some issues in EDU segmentation that have to be a concern and the definition of Thai EDU could be defined in this work.



#### 3.3.3.1 Issues in Thai EDU Segmentation

From previous studies and our observation, there are some interesting issues in Thai EDU segmentation that we have to be concerned about. In general, the Thai language is a language structure of Subject-Verb-Object (S-V-O) like English and many other languages. But some features in the Thai language can make a sentence more complicated and ambiguous. Those features are discussed in this section as follows:

**Lack of Explicit EDU Boundary Marker:** In Thai text, there is no explicit marker or punctuation to indicate the ending of a sentence or EDU. Thai text

can be seen as a stream of continuous characters in a paragraph without any space character or punctuation. In some writing, a space character is used to indicate the ending of a sentence. However, a space character is just an option and does not appear in all sentence endings. Moreover, a space character can appear in many parts of sentence for example before and after a number, before conjunctions ”และ(and)” ”หรือ(or)”, and before the repeater sign ”ๆ”.

**Ambiguity of Word Marker:** Some words which are subordinate conjunction words can be considered to be a word marker to segment EDU for example ”ซึ่ง(that)”, ”โดย(by)”, ”ดังนั้น(therefore)”, ”เพราะ(because)”, ”เพื่อ(for)”. However, some words can be more than one POS depending on their function and their context. In this issue, the accurate word segmentation and POS tagging process are essential components to reduce ambiguity.



**Zero Anaphora:** Sentence in the Thai language can use a gap or omits the subject to refer back to the previous object. Then, the structure of the Thai sentence can be a Verb-Object (V-O) structure.

**Sentential Noun Phrase:** Some noun phrases in the Thai language can be structured similarly to a sentence. A verb can be a part of a noun phrase that can cause ambiguity in a sentence and noun phrases. For example ”ห้อง(room)นอน(sleep)” is meant to a bedroom. To overcome this issue, the accurate noun phrase

chunker or shallow parser can be a key role to disambiguate sentential noun phrases and sentences.

**Relative Clause in Noun Phrase:** Noun phrase in the Thai language can be embedded with a relative clause. A sentence with a relative clause will complicate the task to identify the boundary of EDU. Syntactic information can be useful to indicate the boundary of a relative clause.

#### 3.3.3.2 Definition of Thai EDU

EDU is the minimal discourse unit from discourse structure. A sentence can consist of several EDUs. In our work, we define our EDU as follows.

**Simple EDU:** EDU with a simple S-V-O structure. This EDU can consist of a subject, a verb, and an optional object. A preposition is possible to embed in this EDU structure also.



**Zero Anaphoric EDU:** EDU with the omission of the subject can be a V-O structure. Same as a simple EDU, a preposition can also embed in this EDU structure.

**Relative Clause EDU:** A relative clause, that is embedded in a noun phrase, is considered to be a separated EDU from its main structure. The function of a relative clause is a noun modification and its structure is similar to zero anaphoric EDU.

**Noun List EDU:** Noun list EDU is part of a sentence that describes a list or an example of a noun group. This EDU mostly starts with word ”ได้แก่(for instance), เช่น (such as)”

### 3.3.3.3 EDU Segmentation by Clue Markers

EDU segmentation in Thai text can be partially done by using a clue marker word to break the whole line of text into a piece of EDU. Some clue words with their POS tag are used to identify the origin of EDU and then segment it into a smaller discourse unit. Words and POS tag patterns such as space, subordinate, and conjunction with verbs are examples of clue marker patterns.

### 3.3.3.4 Shallow Parser

A shallow parser is a process to identify the non-recursive of various phrase types (Sha & Pereira, 2003). The various phrase in a sentence can be chunked and is useful to be a precursor to a full parser or information extraction. In this work, CRF is applied to identify a non-recursive phrase. Corpus with manual chunked tagging is used for training the CRF model. There are 11 types of phrases, that are chunked, consisting of head noun, verbal noun, adjective, transitive verb, intransitive verb, adverb, preposition, amount, time, determiner, and classifier phrase. Some symbols are used to tag into a corpus to indicate a type of phrase in the text. Figure 3.10 shows an example of a corpus with a phrase tagged.

[ธรรมชาติ]<*NCM>[ก็จะ]<#VAX>[วิวัฒนาการ]<##VRB>[ให้]<!VPO>
[นิ้ว]<*NCM>[หาย]<#VRB>[ไป]<!VPO>[ที]<-NCT>[ละ]<--COC>
[นิ้ว]<@CLS>[สอง]<@@NUM>[นิ้ว]<@@CLS>
[จน]<SUB>[เหลือแต่]<#VRB>[เพียง]<!PRP>[นิ้ว]<@CLS>
[เดียว]<@@NBM>
[เช่น]<DRF>[เท้า]<*NCM>[ของ]<!PRP>[ม้า]<*NCM>
[มี]<#VRB>[เพียง]<!PRP>[นิ้ว]<@CLS>[เดียว]<@@NBM>[ที่]<PRL>
[เรียก]<#VRB>[ว่า]<!VPO>[กีบ]<*NCM>[เท้า]<**NCM>[ม้า]<**NCM>

*Figure 3.10.* Example of a corpus with a phrase tagged

The symbols, that is used to indicate phrase type, consist of ”*”, ”&”, ”+”, ”#”, ”%”, ”$”, ”!”, ”@”, ”=”, ”?” and ”-” for head noun pattern, verbal noun pattern, adjective pattern, transitive verb pattern, intransitive verb pattern, adverb pattern, preposition pattern, amount pattern, time pattern, determiner pattern and classifier pattern respectively. Single symbol tagged for the starting element of phrase type and double symbol for the following element of phrase type. Some POSs maybe not be tagged if it is not a member of a non-recursive phrase for example ”SUB” subordinated, ”PRL” relative pronoun.



Dictionary for the shallow parser is used to indicate all possible phrase types to POS. All data in a dictionary are gathered from the tagged corpus. Table 3.7 shows examples of a dictionary for the shallow parser.

Table 3.7

*Examples of a dictionary for the shallow parser*

| POS | Phrase Types |
|---|---|
| VAT | + ++ && $ $$ ?? == |
| FAV | $ $$ && |
| NPN | * ** = == ?? ++ |
| SUB | SUB |
| VRB | # ## && == ** $$ |
| VPA | $ $$ |

3 types of features are used for training. A combination of phrase type(PT), POS, and surface word is constructed to make a shallow parser feature. Table 3.8 shows the shallow parser feature pattern.



Table 3.8

*Shallow parser feature pattern*

| Feature Type | Combination of Pattern |
|---|---|
| 1 | PT1 : PT0 : POS1 : POS0 : none : none |
| 2 | PT1 : PT0 : POS1 : POS0 : none : WORD0 |
| 3 | PT1 : PT0 : POS1 : POS0 : WORD1 : none |

The notation ”PT” is the phrase type and the subscription is the position that is determined. The notation ”POS” is the POS and the notation ”WORD” is the surface word.

### 3.3.3.5 EDU Segmentation by Syntactic Pattern

The result of the shallow parser is a phrase chunked in a given text. The syntactic information from phrases chunked is source data for EDU segmentation. From observation, some points of syntactic structure can indicate the point of EDU segmentation. Table 3.9 shows examples of the syntactic pattern.

Table 3.9

*Examples of the syntactic pattern*

| **Syntactic Pattern** |
|---|
| VRBpat:PRPpat:NP::VRBpat |
| VRBpat:NP::VRIpat |
| VRBpat:NP:ADVpat:PRPpat:NP::VRIpat |
| VRIpat::PRPpat:VRBpat |
| VRIpat:PRPpat:NP::PRL:VRBpat |
| NP2::NP:VRBpat |
| NP2::CON:VRBpat |



The pattern consists of a sequence of phrase types that connect with the colon symbol. The EDU segment point is indicated by the double colon symbol. NP and NP2 in the pattern are the encapsulated phrase type that consists of the head noun, preposition, verbal noun, amount, time, adjective, and determiner pattern. In some contexts, the EDU can be constructed by only head nouns and adjectives. That means some EDUs are looked like noun phrases. In this work, we define NP as a noun phrase on the EDU and NP2 is a noun phrase that is an EDU. We encapsulate NP and NP2 by using rules that NP2 can be composed of the head noun, adjective, and adverb that no preposition

is before the head noun and it is not followed by a verb. NP can be composed of the head noun, preposition, verbal noun, amount, time, adjective, and determiner.

#### 3.3.3.6 EDU Reconstruction by Rule-Based

The noun list can be separated by space that needs to be reconstructed. Moreover, space is used in some writing style to separate some word such as subordinate word, noun phrase, or some parts of EDU that breaks the EDU structure by EDU segmentation by clue marker process. Some EDUs need to be reconstructed to increase the precision of the EDU structure. Rule-based is applied to analyze the partial EDU and then construct the new EDU structure. The rule consists of 3 parts: starting condition, combined condition, and commit condition. Table 3.10 shows examples of the EDU reconstruction rule.



Table 3.10

*Examples of the EDU reconstruction rule*

| No. | Starting | Combined | Commit |
|---|---|---|---|
| 1 | Line is end with ”PRPpat” | Line is NP | Line is not NP and not bound |
| 2 | Line is end with ”VRBpat” | Line is NP | Line is not NP and not bound |
| 3 | Line is NP | Line is NP | Line start with ”VRBpat” and bound |

The EDU reconstruction process starts by finding the line that matches the starting condition. After that, if the next line matches the combined condition then merge this line with the starting line. Until the combined condition is not matched then check with the commit condition. If the commit condition is matched then check that the commit

line could be bound with the new EDU structure. If it is bounded then merge the last line with the new EDU structure. Lastly, if the commit condition is not matched then all the processes are canceled, and then try to the next rules.

### 3.4 Ellipsis and Anaphora Resolution

Anaphora resolution is an NLP task that solves the referent objects in text. The anaphora resolution research in Thai text is still rare (Aroonmanakun, 2000; Pathanasin, 2018). To find the complete semantics in Thai text, the anaphora resolution with acceptable precision is an essential key to success. Some phenomena are interesting problems that appear in Thai text on anaphora resolution. In the text, some anaphoras do not refer to any object but refer to the reader or the generalized object. The anaphora, that do not refer to any object in the text, is called non-referential anaphora. To resolve the reference of the anaphora, the anaphora have to resolve whether it is non-referential anaphora or not before. Moreover, some parts of the object can be omitted in Thai text such as the preposition of the owner. The omission of the preposition of the owner is called the ellipsis of the owner. The ellipsis of the owner is the language phenomenon that needs to be resolved to get complete information from the text. Due to the complicated of sentence breaking, The anaphora resolution in the EDU segmentation can be more useful. Discourse relation is the relationship between the discourse segment. Discourse relation is needed in the NLP application such as text summarization. However, it is possible to resolve the reference of the anaphora by do not use the discourse relation.

### 3.4.1 Anaphora in Thai Texts

The anaphora is a linguistic tool for referencing a thing mentioned earlier in a discourse. A phenomenon like non-referential anaphora is an interesting item that affects the anaphora resolution in this work. The interesting information on the use of anaphora in Thai text is described in this section.

#### 3.4.1.1 Anaphora Types

In this work, we define anaphora in 4 types which are zero anaphora, pronominal anaphora, nominal anaphora, and ellipsis of the owner. All types of anaphora are described as follows.

> **Zero Anaphora:** Zero anaphora is the use of a gap in the subject of a sentence that references the object in the prior sentence. There is normally a lot of use of zero anaphora in Thai text. The example of the use of zero anaphora is shown in Figure 3.11.



| นกกระจอกเทศ | เป็น | นก | ขนาด | ใหญ่ |
|---|---|---|---|---|
| Ostrich | is | bird | size | big |
| $\phi$ | มี | ถิ่นกำเนิด | ใน | ทวีปแอฟริกา |
| | has | origin | in | Africa |

*Figure 3.11.* Zero anaphora

In the first sentence, the word ”นกกระจอกเทศ(Ostrich)” is introduced in the subject of the sentence. In the next sentence, the subject is omitted by using the gap (φ) and there is only a verb phrase appears. Due to the use of zero anaphora, a Thai sentence can be formed by only a verb phrase. In the process of EDU segmentation, the embedded relative clause EDU can form a zero anaphora after EDU segmentation. The example of the embedded relative clause is shown in Figure 3.12.

| แมว | ที่ | เลี้ยง | ตาม | บ้าน | จะ | มี | รูปร่าง | ขนาด | เล็ก |
|---|---|---|---|---|---|---|---|---|---|
| cat | that | pet | at | house | will | has | shape | size | small |

*Figure 3.12.* Embedded relative clause

The clause contains 2 EDUs with the relative pronoun ”ที่(that)” on the first EDU and the second EDU can segment at the verb ”จะ(will) มี(has)”. The given clause can be segmented into 2 EDUs with the zero anaphora attached to the second EDU as Figure 3.13 below.

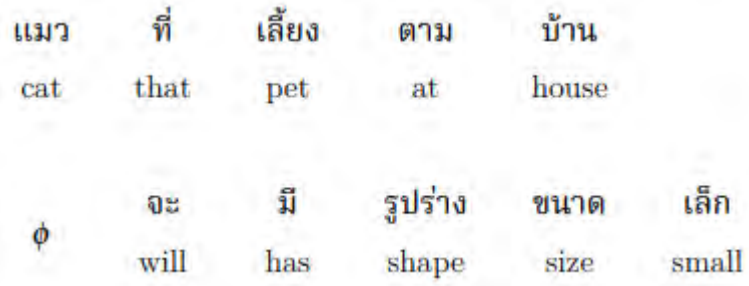


*Figure 3.13.* Zero Anaphora on the embedded relative clause after EDU segmentation

After the EDU segmentation, the new zero anaphora will be exposed to the new EDU. This phenomenon happens many times in the corpus. The anaphora resolution could capture this phenomenon to make an accurate resolution.

**Pronominal Anaphora:** Pronominal anaphora is the use of pronouns to refer to the object in the prior sentence. A pronoun is a fundamental linguistic tool to refer to the thing that has been introduced in the antecedent. The example of the use of the pronoun is shown in Figure 3.14.

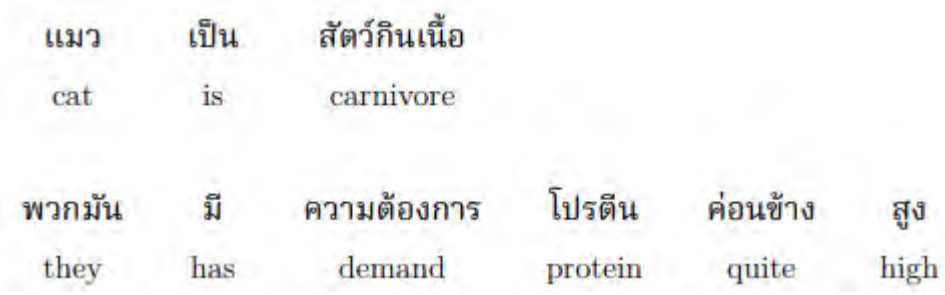


*Figure 3.14.* Pronominal anaphora

The pronoun ”พวกมัน(they)” in the second sentence refers to the word ”แมว(cat)” that has been introduced in the first sentence. The use of the pronoun is widely used in the corpus. The resolution of the pronoun may need additional information such as gender, and number to resolve the reference.

**Nominal Anaphora:** Nominal anaphora is the use of nouns with a determiner to refer to the object in the prior sentence. A Noun that is nominal anaphora can be a supertype (hyponymy) of the reference. A determiner can be used as an indication to identify the nominal anaphora. The example of the nominal anaphora is shown in Figure 3.15.


*Figure 3.15.* Nominal anaphora

The word ”นก(bird)” with the determiner ”ชนิด(kind)นี้(this)” in the second sentence is a nominal anaphora that refers to the word ”นกกระจอกเทศ(ostrich)” in the first sentence. Moreover, the same headword can be used as a nominal anaphora to refer to the same word with a related adjective as Figure 3.16.

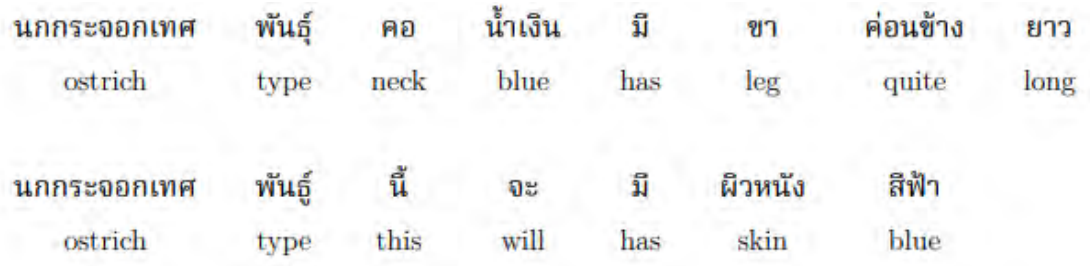


*Figure 3.16.* Nominal anaphora on the same head word

The word ”นกกระจอกเทศ(ostrich)” with the determiner ”พันธุ์(type)นี้(this)” in the second sentence refers to the same word ”นกกระจอกเทศ(ostrich)” with the related adjective ”พันธุ์(type)คอ(neck)น้ำเงิน(blue)” in the first sentence. These phenomena occur many times in the corpus. This anaphora can be resolved with the utilization of the semantic ontology to resolve the hyponymy.



**Ellipsis of the owner:** Nouns in Thai text can omit the preposition of the owner that was introduced in the antecedent. Mostly, a part-of or meronymy is a semantic relation that attaches between a noun and the ellipsis. The example of the ellipsis of the owner is shown in Figure 3.17.


*Figure 3.17.* Ellipsis of the owner

The word ”นกกระจอกเทศ(ostrich)” is introduced in the first sentence. In the second sentence, the word ”ผิวหนัง(skin)” omits the preposition of the owner ”ostrich”. However, the reader still knows that the skin is the skin of the ostrich. Additional information like the ontology of meronymy is needed for resolving the ellipsis of the owner.

#### 3.4.1.2 Referential and Non-Referential Anaphora

Anaphora generally refer to the reference object in the antecedent. There is an interesting phenomenon that anaphora may not refer to any object in text. Therefore, the anaphora can be tagged into 2 kinds that are referential and non-referential anaphora.

**Referential Anaphora:** Referential anaphora means any type of anaphora that refers to the object in the text. Mostly, the anaphora that appear in the text is the referential anaphora. From the observation in the corpus, the pronoun, zero anaphora, and ellipsis of the owner mostly refer to the existing entities in the text. However, there is a lot of nominal anaphora that does not refer to any object in the text. Before resolving the referential anaphora, the anaphora could be identified that is referential or non-referential anaphora.

**Non-Referential Anaphora:** Non-referential anaphora means any type of anaphora that does not refer to any explicit entity in text. Any type of anaphora can be a non-referential anaphora. In zero anaphora, non-referential anaphora

occurs mostly from the use of the verb of occurrence. The example of non-referential anaphora in zero anaphora is shown in Figure 3.18.

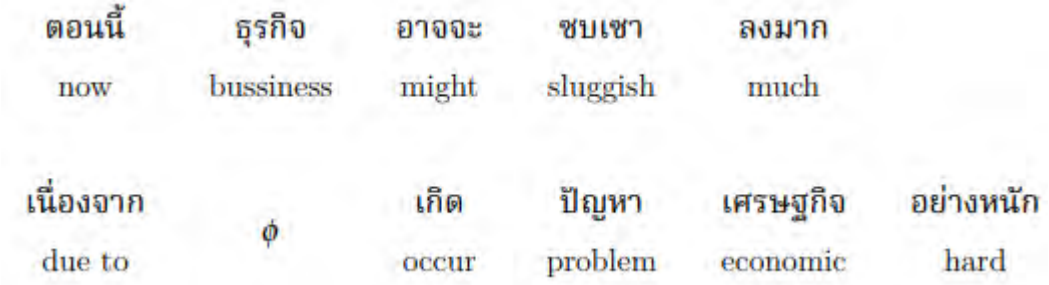


*Figure 3.18.* Non-referential in zero anaphora

The word ”เกิด(occur)” in the second sentence is a verb of occurrence that makes this sentence forms only a verb phrase. The zero anaphora in this sentence does not refer to any object in the antecedent. There are some verbs that can generate the non-referential anaphora in zero anaphora such as ”เกิด(occur, birth)”, ”มี(happen, has)”, and ”เป็น(be)”. There is the pronoun ”เรา(we)” that can refer to the reader or general people that do not refer to any object in text. The example of non-referential anaphora in pronominal anaphora is shown in Figure 3.19.

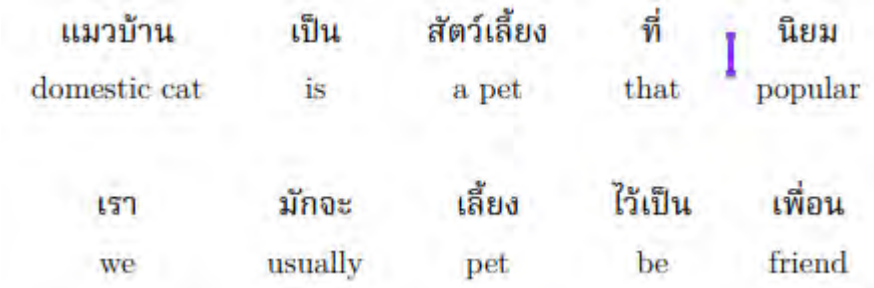


*Figure 3.19.* Non-referential in pronominal anaphora

The word ”(we)” in the second sentence refers to the general people that do not exist in the text. In nominal anaphora, there is the word with some determiner that refers to the general object that is not specified to any object in the text. Figure 3.20 shows the example of the non-referential in nominal anaphora.

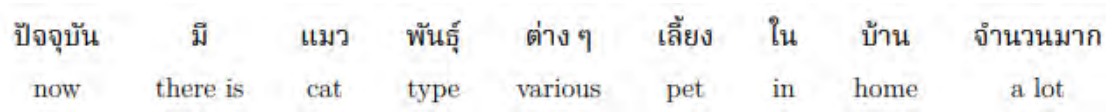


*Figure 3.20.* Non-referential in nominal anaphora

The word ”ต่าง ๆ(various)” makes the noun ”แมว(cat)” be general and not specify to any object in the text. The surface word could be used for learning to identify which nominal anaphora could be non-referential anaphora.



### 3.4.2 The Resolution

There are 3 steps of processes in the anaphora resolution: anaphora determiner, resolution for non-referential anaphora, and resolution for referential anaphora. Anaphora determiner is the algorithm for determining the anaphora type in EDU. After that, the resolution for non-referential anaphora is applied to distinguish the anaphora which is the non-referential or referential anaphora. Finally, the resolution for referential anaphora is applied to find the reference of the referential anaphora from the antecedent EDU. The ontology is a background knowledge that contains semantic

concepts and semantic relations such as meronymy and hyponymy. The ontology is significant in the anaphora determiner and is a component of the feature set for the anaphora resolution process. Figure 3.21 shows the overview of the anaphora resolution processes.

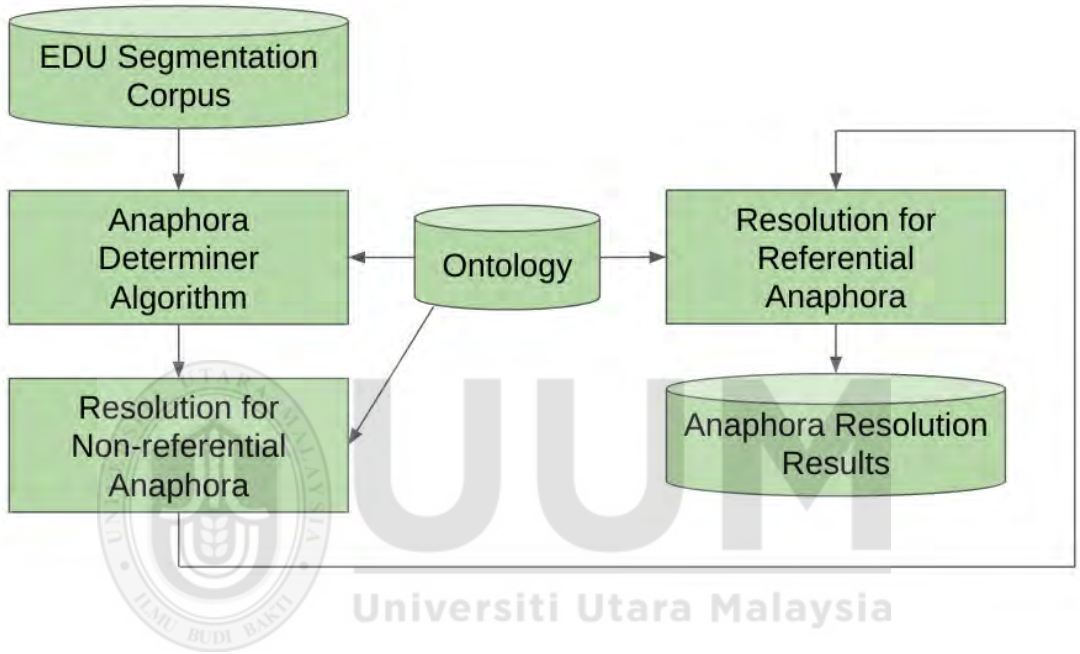




*Figure 3.21.* Overview of the anaphora resolution processes

In Figure 3.21, the anaphora determiner algorithm is the process of determining the anaphora type for instance zero, nominal, pronominal, and also the ellipsis. The ellipsis is determined by using the semantic relation in the ontology to identify which noun phrase is part of something. After all anaphora is identified, the resolution for non-referential anaphora is worked to determine that which anaphora is the non-referential anaphora. Consequently, the resolution for referential anaphora is worked to determine the referent of all anaphora and also the ellipsis.

### 3.4.2.1 Corpus Preparation for Anaphora Resolution

The corpus for anaphora resolution has come from the result of The EDU segmentation process. The corpus will be tagged with the additional information for training in the anaphora resolution training model. The entities in the corpus will be tagged the number for reference. Each anaphora will be tagged with the number and the reference number. The zero number will tag in the reference number in the case of the non-referential anaphora. Figure 3.22 shows the example of the anaphora tagging in the corpus.


*Figure 3.22.* Anaphora tagging in the corpus

### 3.4.2.2 Anaphora Determiner Algorithm

The anaphora determiner algorithm is the algorithm to indicate that each phrase in EDU is the entity or the anaphora and also identify the anaphora type to the anaphora. The rule-based is applied to make a decision to indicate the entity and identify the anaphora type. The anaphora determiner algorithm is shown in Figure 3.23.

```
input: Q us an array of EDU
begin
   foreach E in Q do
      if E has no subject with (VRBpat,VRIpat,ADJpat) then
         Mark Zero at the subject
      end
      foreach H is (HNpat,VNNpat,AMTpat,DETpat) in E do
         if There is pronoun in H then
            Mark Pronominal
         else if H is (HNpat,VNNpat) and connect withnDETpat then
            Mark Nominal
         else if H is HNpat and has part-of relation and is a subject then
            if H follows by the preposition of the owner then
               Mark Entity
            else
               Mark Ellipsis
            end
         else if H is (DETpat,AMTpat) with no
         (HNpat,TIMEpat,CLSpat,DETpat,ADJpat) before then
            Mark Entity
         else if H is (HNpat,VNNpat) then
            Mark Entity
         else
            continue
         end
      end
   end
end
```

*Figure 3.23.* Anaphora determiner algorithm

The non-recursive phrases that appear in the algorithm are head noun (HNpat), verbal noun (VNNpat), time (TIMEpat), classifier (CLSpat), determiner (DETpat), adjective (ADJpat), amount (AMTpat), transitive verb (VRBpat), and intransitive verb (VRIpat). After finished the anaphora determiner process, the entities and all anaphora will be tagged with the identification number for reference.

#### 3.4.2.3 Resolution for Non-Referential Anaphora

In this work, the first step to resolving the anaphora is to identify whether the anaphora is a non-referential or referential anaphora. The ranking model (Denis & Baldridge, 2008) is selected to resolve the non-referential and also the referential anaphora. The ranking model is shown in the equation 3.4.

$$P(\phi_i|\pi) = \frac{\exp(\sum_j w_j f_j(\pi, \phi_i))}{\sum_k \exp(\sum_j w_j f_j(\pi, \phi_k))} \tag{3.4}$$

In equation 3.4, $\boldsymbol{\pi}$ stands for the anaphora type, $\boldsymbol{\varphi_i}$ for the antecedent candidate, $\boldsymbol{f_j}$ for the feature function, $\boldsymbol{w_j}$ for the weight of the feature function, and $\boldsymbol{k}$ for the iterator of all candidates. In the training process, the weight adjustment is defined in the equation 3.5.

$$w_j = w_j + \alpha \left[ f_j(\pi, \phi_i) - \sum_k P(\phi_k|\pi) f_j(\pi, \phi_k) \right] \tag{3.5}$$

The features are extracted from the tagged corpus and then store in the database for training purposes. The structure of the feature consists of 3 parts that are feature type, feature value, and weight. Table 3.11 shows the example of the features of non-referential anaphora in the database.

Table 3.11

*Example of the features of non-referential anaphora in the database*

| **Features** | **Weight** |
|---|---|
| elip0N6:CON | 1.25329154714611 |
| zero0N4:บาง_ที | 2.65211400253568 |
| pro0N1:ไม่_ได้_ต่อต้าน | 1.05137012602686 |
| zero0N7:เรียนรู้ | 1.02143300771681 |
| zero0N7:ยัง_ชอบ_กิน | 2.61727791578152 |
| pro0Y5:CON | 1.8745485653644 |

The feature type and the feature value are encapsulated to the string with the colon connector. The first part of the string is the feature type and the second part is the feature value. The feature type ”zero0N4” encapsulated 3 meanings. ”zero” means zero anaphora. ”0N” means is not non-referential anaphora. And ”4” means the fourth kind of feature value. 16 kinds of feature values are used to indicate the non-referential anaphora. Table 3.12 shows the kinds of feature values for non-referential anaphora.

Table 3.12

*Kinds of feature values for non-referential anaphora*

| | | |
|---|---|---|
| 1.verb | 2.verb pos | 3.verb phrase type |
| 4.word in front | 5.word pos in front | 6.word phrase type in front |
| 7.word behind | 8.word pos behind | 9.word phrase type behind |
| 10.syntactic position | 11.head or part of noun | 12.word |
| 13.pos | 14.phrase type | 15. start paragraph |
| 16. end paragraph | | |

Verb, syntactic information, and word that surround the anaphora are used as the features for the training model. Due to the non-referential anaphora having no reference, then only surface words and some syntactic information are considered to be used to indicate the non-referential anaphora.



### 3.4.2.4 Resolution for Referential Anaphora

The features for referential anaphora are also extracted from the tagged corpus and then store in the database. The feature structure consists of 4 parts that are feature type, feature value, distance, and weight. Table 3.13 shows the example of the features of referential anaphora in the database.

Table 3.13

*Example of the features of referential anaphora in the database*

| **Features** | **Weight** |
|---|---|
| zeroXA7:มักจะ_ทำ:1 | 1.70678539113928 |
| elipXB15:ขน:นกกระจอกเทศ | 1.09692873658293 |
| zeroXC4:ย่อมที่จะ:สำหรับ:1 | 1.0 |
| proXB10:Dobject:2 | 1.09317637004032 |
| nomXB1:ซ่อน:1 | 1.00003546264457 |
| zeroXB12:นก:8 | 4.11060974191702 |

The feature value can be one value or pair value of anaphora and reference. Then, the feature value for referential anaphora can be divided into 3 groups: anaphora value, reference value, and pair of anaphora and reference value. The first group is the value of the anaphora and the surrounding information. 16 kinds of the first feature values on the anaphora side are used to indicate the referential anaphora. Table 3.14 shows the first group of feature values on the anaphora side.

Table 3.14

*First group of feature values on the anaphora side*

| |
|---|
| 1.verb (anaphora) : distance |
| 2.verb pos (anaphora) : distance |
| 3.verb phrase type (anaphora) : distance |
| 4.word in front (anaphora) : distance |
| 5.word pos in front (anaphora) : distance |
| 6.word phrase type in front (anaphora) : distance |
| 7.word behind (anaphora) : distance |
| 8.word pos behind (anaphora) : distance |
| 9.word phrase type behind (anaphora) : distance |
| 10.syntactic position (anaphora) : distance |
| 11.head or part of noun (anaphora) : distance |
| 12.word (anaphora) : distance |
| 13.pos (anaphora) : distance |
| 14.phrase type (anaphora) : distance |
| 15. start paragraph (anaphora) : distance |
| 16. end paragraph (anaphora) : distance |



The second group is the value of the reference and the surrounding information. 17 kinds of the second feature values on the reference side are shown in Table 3.15.

Table 3.15

*Second group of feature values on the reference side*

| |
|---|
| 1.verb (reference) : distance |
| 2.verb pos (reference) : distance |
| 3.verb phrase type (reference) : distance |
| 4.word in front (reference) : distance |
| 5.word pos in front (reference) : distance |
| 6.word phrase type in front (reference) : distance |
| 7.word behind (reference) : distance |
| 8.word pos behind (reference) : distance |
| 9.word phrase type behind (reference) : distance |
| 10.syntactic position (reference) : distance |
| 11.head or part of noun (reference) : distance |
| 12.word (reference) : distance |
| 13.pos (reference) : distance |
| 14.phrase type (reference) : distance |
| 15.word (anaphora) : word (reference) |
| 16.is-head-word-match : distance |
| 17. is-hyponymy : distance |



The third group is the pair value of the anaphora and reference and the surrounding information. 14 kinds of the third feature values on both sides of anaphora and reference are shown in Table 3.16.

Table 3.16

*Third group of feature values on both sides of anaphora and reference*

| |
|---|
| 1.verb (anaphora) : verb (reference) : distance |
| 2.verb pos (anaphora) : verb pos (reference) : distance |
| 3.verb phrase type (anaphora) : verb phrase type (reference) : distance |
| 4.word in front (anaphora) : word in front (reference) : distance |
| 5.word pos in front (anaphora) : word pos in front (reference) : distance |
| 6.word phrase type in front (anaphora) : word phrase type in front (reference) : distance |
| 7.word behind (anaphora) : word behind (reference) : distance |
| 8.word pos behind (anaphora) : word pos behind (reference) : distance |
| 9.word phrase type behind (anaphora) : word phrase type behind (reference) : distance |
| 10.syntactic position (anaphora) : syntactic position (reference) : distance |
| 11.head or part of noun (anaphora) : head or part of noun (reference) : distance |
| 12.word (anaphora) : word (reference) : distance |
| 13.pos (anaphora) : pos (reference) : distance |
| 14.phrase type (anaphora) : phrase type (reference) : distance |

A total of 47 kinds of feature values are used in the resolution for referential anaphora. The distance is set to the maximum of 10 EDUs between the anaphora and the reference. The ranking model is used to find the best probabilistic on the antecedent candidates that are up to 10 EDUs.

### 3.5 Semantic Parser

The semantic parser is the process to construct the semantic frame after all of the text is finished the processes of the Thai morphological analysis and the anaphora resolution. Syntactic phrase structure and surface word are the crucial data that are used to construct the semantic frame structure: semantic label, semantic ID, and semantic frame type. The semantic parser consists of 2 parts: word sense disambiguation, and semantic frame construction. The process of word sense disambiguation is the process to identify the correct semantic ID for the surface word in syntactic phrase structure. After the correct semantic ID is identified, the semantic frame is constructed by the semantic frame construction process. The example of transforming from syntactic phrase structure with semantic ID to semantic frame structure is shown in Figure 3.24.

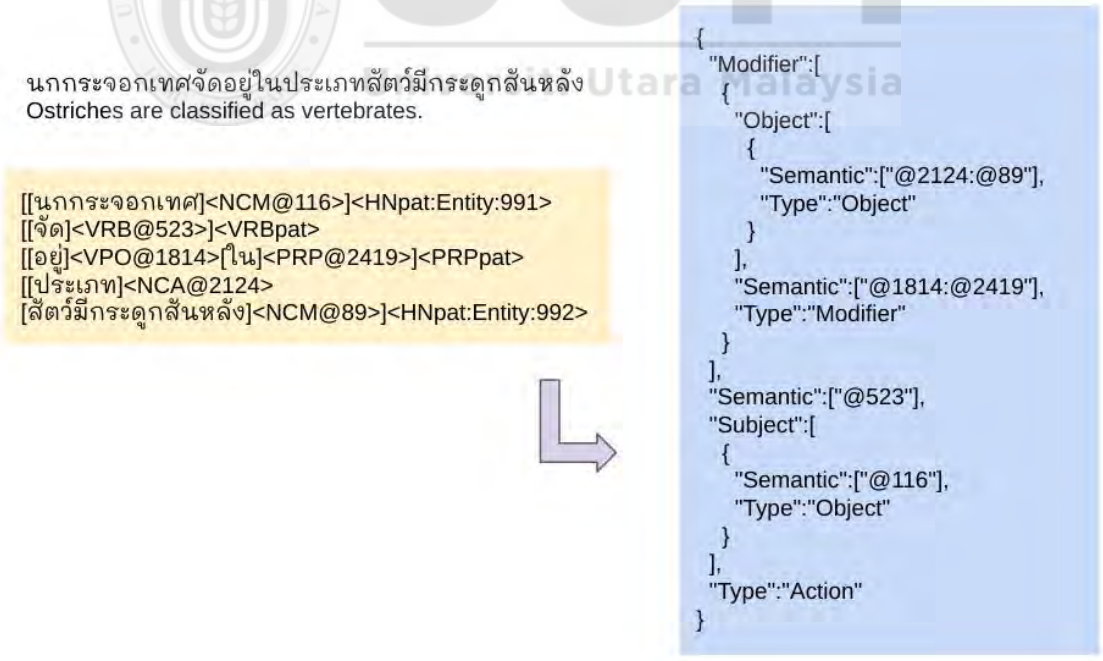


*Figure 3.24.* The example of transforming from syntactic phrase structure with semantic ID to semantic frame structure

### 3.5.1 Word Sense Disambiguation

Word sense disambiguation is the process to recognize the actual semantics of the target surface word. Some surface words can vary in meaning depending on their context. In this work, the word sense disambiguation process is divided into 2 parts: corpus tagging, and semantic disambiguation.

#### 3.5.1.1 Corpus Tagging

The tagged corpus by the semantic ID is the important data source for the training process. Every surface word in the corpus will be tagged the semantic ID. The semantic ID is the running number for representing the semantic concept that we have recognized in the corpus. The running number of the semantic ID is predefined manually during the corpus tagging. When the word with a new semantic is recognized, the new running number is defined in the corpus by the human. The semantic ID will be tagged with the ”@” symbol together with the running number. The semantic ID is tagged along with the POS tagging in the corpus. Figure 3.25 shows the example of a tagged corpus by the semantic ID.

[[นกกระจอกเทศ]<NCM@116>]<HNpat:Entity:991>[[จัด]<VRB@523>]<VRBpat>
[[อยู่]<VPO@1814>[ใน]<PRP@2419>]<PRPpat>[[ประเภท]<NCA@2124>
[สัตว์มีกระดูกสันหลัง]<NCM@89>]<HNpat:Entity:992>[@]<Zero:993:991>
[[เป็น]<VRB@936>]<VRBpat>[[นก]<NCM@114>]<HNpat:Entity:994>
[ที่]<PRL@858>[@]<Zero:995:994>[[มี]<VRB@483>]<VRBpat>
[[ขนาด]<NCA@1349>]<HNpat:Entity:996>[[ใหญ่]<VAT@2217>]<ADJpat>
[[ที่สุด]<ADV@1833>]<ADVpat>[[ใน]<PRP@2419>]<PRPpat>
[[โลก]<NPN@1913>]<HNpat:Entity:997>
[@]<Zero:998:993>[[มี]<VRB@483>]<VRBpat>
[[ถิ่นกำเนิด]<NCM@159>]<HNpat:Entity:999>[[ใน]<PRP@2419>]<PRPpat>
[[[ทวีป]<NCM@184>[แอฟริกา]<NPN@2369>]<NPN>]<HNpat:Entity:1000>

*Figure 3.25.* The example of tagged corpus by the semantic ID

The semantic ID will be tagged after the POS of the surface word with the ”@” symbol to represent the semantic concept. The semantic dictionary is also an important source of the semantic disambiguation process that contains the semantic ID, surface word, and word POS. The semantic dictionary can be generated from the tagged corpus. Table 3.17 shows the examples of semantic dictionary.

Table 3.17

*The examples of semantic dictionary*

| Semantic ID | Word | POS |
|---|---|---|
| 232 | เสือชีตาห์ (cheetah) | NCM |
| 232 | เสือชีต้า (cheetah) | NCM |
| 236 | เสือปลา (fishing cat) | NCM |
| 266 | โค (cow) | NCM |
| 266 | วัว (cow) | NCM |
| 271 | ใต้ (under) | NPP |
| 272 | ใต้ (under) | PRP |



### 3.5.1.2 Semantic Disambiguation

Semantic disambiguation is the process that recognizes the most matched semantic ID to the surface word. The learning model utilized the semantic dictionary and tagged corpus to find the probability of each semantic ID for the surface word. The context surrounding the surface word is captured to define the features of the training model. The ranking model is applied as the learning model that is shown in the equation 3.6.

$$P(\eta_i|\Theta) = \frac{\exp(\sum_j w_j f_j(\Theta, \eta_i))}{\sum_k \exp(\sum_j w_j f_j(\Theta, \eta_k))} \quad (3.6)$$

In equation 3.6, $\boldsymbol{\Theta}$ stands for the semantic ID, $\boldsymbol{\eta_i}$ for the surface word, $\boldsymbol{f_j}$ for the feature function, $\boldsymbol{w_j}$ for the weight of the feature function, and $\boldsymbol{k}$ for the iterator of all candidates. In the training process, the weight adjustment is defined in the equation 3.7.

$$w_j = w_j + \alpha \left[ f_j(\Theta, \eta_i) - \sum_k P(\eta_k|\Theta) f_j(\Theta, \eta_k) \right] \quad (3.7)$$

After finishing the training, the weight will be adjusted for each feature and stored in the database. There are 7 feature type is defined in the semantic concept training. Table 3.18 shows the feature types for the semantic concept training.

Table 3.18

*The feature types for the semantic concept training*

| |
|---|
| 1 : Word : POS : Semantic ID : Verb Word : POS |
| 2 : Word : POS : Semantic ID : Phrase Types |
| 3 : Word : POS : Semantic ID : Other Word : POS |
| 4 : Word : POS : Semantic ID : Next Phrase Type |
| 5 : Word : POS : Semantic ID : Head Word Next Phrase : POS |
| 6 : Word : POS : Semantic ID : Previous Phrase Type |
| 7 : Word : POS : Semantic ID : Head Word Previous Phrase : POS |

The features are extracted from the tagged corpus for each feature type. The learning model will adjust the weight of each feature for finding the probability of each semantic ID for the surface word. Table 3.19 shows the examples of the feature that are extracted from the tagged corpus.

Table 3.19

*The examples of the feature that are extracted from the tagged corpus*

| **Features** | **Weight** |
|---|---|
| 7:ใน:PRP:1462:ไข่:NCM | 0.204908678545273 |
| 1:ใน:PRP:1462:ผสมพันธุ์:VRB | 0.21813734675515 |
| 1:เป็น:VPO:937:แบ่ง:VRB | 0.223162095723574 |
| 3:เป็น:VPO:2394:เป็น:VPO | 0.238678805127176 |
| 2:เป็น:VPO:2394:PRPpat | 0.238678805127176 |
| 5:ใน:PRP:2419:สถานที่:NCM | 0.260730558129846 |
| 5:จาก:PRP:1474:none:none | 0.271167453222577 |
| 4:จาก:PRP:1474:none | 0.271167453222577 |
| 7:จาก:PRP:1474:สาเหตุ:NCM | 0.271167453222577 |
| 5:ใน:PRP:1462:เลือกตั้ง:VRB | 0.272657467303031 |
| 7:เป็น:VPO:937:แบ่ง:VRB | 0.285647147370223 |



In Table 3.19, the weight of the feature is affected to the determined of the semantic ID process. The higher weight of the feature makes it more probability to determine the semantic ID of that surface word. Also, the lower weight of the feature can cause the low probability of that semantic ID to the surface word.

### 3.5.2 Semantic Frame Construction

Semantic frame construction is the process to transform the syntactic phrase structure into the semantic frame as the knowledge representation. The semantic frame will be generated by the activating of semantic frame construction rules that contain the semantic construction information.

#### 3.5.2.1 Frame-based Knowledge Representation

The syntactic phrase structure will be formulated into a frame-based representation for representing the semantic structure. The frame can be formulated as the below formulas.

$$Fr = (S,V) \tag{3.8}$$

$$\{Sm,Su,Ob,Md,Nd,At,Ax,T\ m,Dt,Fo,To,Nm,Nm2,Cl,Ty,Co\} \in S \tag{3.9}$$

$$\{Se,Num,Fr,St\} \in V \tag{3.10}$$

Where:

Fr is the semantic frame

S is the slot name

Sm is the semantic slot

Su is the subject slot

Ob is the object slot

Md is the modifier slot

Nd is the noun modifier slot

At is the attribute slot

Ax is the auxiliary slot

Tm is the time slot

Dt is the determiner slot

Fo is the from slot

To is the to name

Nm is the number slot

Nm2 is the number2 slot

Cl is the classifier slot

Ty is the type slot

Co is the comparative slot

V is the slot value

Se is the semantic ID

Num is the number value

St is the string value

In this work, there are 6 semantic frame types: action, object, modifier, attribute, time, and determiner frame. The syntactic phrase structure will be transformed into these 6 semantic frame types to represent the semantic structure.

The action frame consists of 7 slots: semantic, subject, object, modifier, attribute, auxiliary, and time. Figure 3.26 shows the structure of the action frame. The ”@” symbol refers to the semantic ID and the other slot may contain the other frame type. The semantic slot contains the semantic ID of the action verb in EDU. The subject slot contains the list of object frames that are the subject of the action. The object frame contains the list of the object frame. The modifier slot contains the list of the modifier that infer from the preposition in the EDU. The attribute frame contains the list of the attribute that infer from the adverb in the EDU. The auxiliary frame contains the list of the semantic ID of the auxiliary word that occurs in the action verb. And the time frame contains the list of time series that occur in the EDU.



```
Action: {
        Semantic:   @
        Subject:    [ Object ]
        Object:     [ Object ]
        Modifier:   [ Modifier (prep)]
        Attribute:  [ Attribute ]
        Auxiliary: [ @ ]
        Time : [ Time ]
}
```

*Figure 3.26.* The structure of the action frame

The object frame consists of 5 slots: semantic, nmodifier, modifiers, attribute, and determiner. Figure 3.27 shows the structure of the object frame. The semantic slot contains the semantic ID of the object that is the head noun in the head noun phrase in EDU. The nmodifier slot contains the list of the semantic ID of the noun modifier that occurs along with the head noun in the head noun phrase. The modifier slot contains the list of modifier frames that infer from the preposition phrase that attaches to the head noun phrase. The attribute slot contains the list of the attribute frame that infer from the adjective phrase that attaches to the head noun phrase. The determiner slot contains the list of the determiner frame that infer from the determiner phrase that attaches to the head noun phrase.

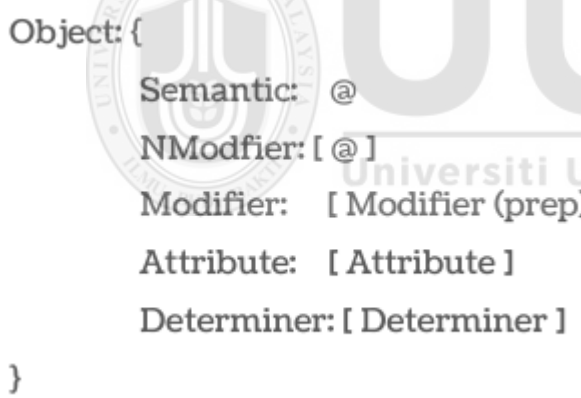


*Figure 3.27.* The structure of the object frame

The modifier frame consists of 2 slots: semantic, and object. Figure 3.28 shows the structure of the modifier frame. The semantic slot contains the semantic ID of the preposition that occurs in the preposition phrase. The object slot contains the object in the preposition phrase.

```
Modifier: {
        Semantic:   @ (prep)
        Object:   Object
}
```

*Figure 3.28.* The structure of the modifier frame

The attribute frame consists of 9 slots: fuzzy, semantic, comparative, attribute, classifier, num, num2, object, and auxiliary. Figure 3.29 shows the structure of the attribute frame. The fuzzy slot contains the semantic ID of the word that makes the value in the attribute be the approximate or obscure value such as ”ประมาณ (approximate)”, ”ราว ๆ(about). The semantic slot contains the semantic ID of the attribute word that occurs in the adjective or adverb phrase. The comparative slot contains the list of the object frame that is the comparative object in the adjective phrase. The attribute slot contains the list of the attribute frame that is the modifier of the attribute such as ”มาก(very)”. The classifier slot contains the semantic ID of the classifier which is the unit of measurement that occurs along with the numeric data in the adjective phrase. The num and num2 contain the numeric data that occur in the adjective phrase. The object slot contains the list of the object that occur along with the attribute in the adjective phrase. The auxiliary slot contains the list of the semantic ID of the auxiliary word that occurs in the adjective phrase.

```
Attribute: {
        Fuzzy: @
        Semantic: @
        Comparative: [ Object ]
        Attribute: [ Attribute ]
        Classifier: @
        Num: Number
        Num2: Number
        Object: [ Object ]
        Auxiliary: [ @ ]
}
```

*Figure 3.29.* The structure of the attribute frame



The time frame consists of 5 slots: from, to, num, classifier, and semantic. Figure 3.30 shows the structure of the time frame. The from slot contains the series of the semantic ID of time which is the starting time in the action. The to slot contains the series of the semantic ID of time which is the ending time in the action. The num slot contains the numeric date of the time value. The classifier slot contains the classifier that is the time measurement that occurs in the time phrase. The semantic slot contains the semantic ID of time that is wordy time such as ”เร็ว ๆ นี้(soon)”.

```
Time: {
        From:  [ @ ]
        To:  [ @ ]
        Num: Number
        Classifier: @
        Semantic: @
}
```

*Figure 3.30.* The structure of the time frame

The determiner frame consists of slots: num, classifier, semantic, and time. Figure 3.31 shows the structure of the determiner frame. The num slot contains the numeric data that occurs in the determiner phrase. The classifier slot contains the semantic ID of the measurement unit that occurs along with the numeric value in the determiner phrase. The semantic slot contains the semantic ID of the determiner word. The time slot contains the list of the time frame that occurs in the determiner phrase.

```
Determiner: {
        Num: Number
        Classifier:  @
        Semantic:  @ (บาง, ทุก, ที่, ลำดับที่)
        Time:        [ Time ]
}
```

*Figure 3.31.* The structure of the determiner frame

#### 3.5.2.2 Semantic Frame Construction Rules

The semantic frame construction rule is the set of rules that contain the construction information for constructing the semantic frame. There are 3 types of semantic frame construction rules: frame activation rule, constraint rule, and slot extraction rule.

The frame activation rule is the rule for activating to create the semantic frame. The syntactic phrase structure will be applied to each frame activation rule until it satisfies the pattern matching in the rule and then that rule will be activated. The frame activation rule consists of 3 parts: rule name, pattern matching, and constructed information. The rule name starts with the ”!” symbol and follows with the rule name. Pattern matching is the sequence of phrase structures in EDU that will be matched to activate the rule. The constructed information is the constraint rule or the slot extraction rule along with the pattern input as constructed information. Each part of the rule will be separated by the ”;” symbol. The frame activation rule will be active when the pattern matching is matched with the input EDU. And then the semantic frame will be created. After that, the constructed information will be active to extract information to the frame slot. The syntax of the frame activation rule is defined below.

rn[:fn];pt[;#rn;ip][;@rn:sl;ip]

The ”rn” is the rule name and the ”fn” is the option for the frame name. If the frame name is not provided, the rule name will be the frame name. The ”pt” is the pattern matching. And the last part can follow by the number of the constraint rule or the slot

extraction rule. The constraint rule is the rule with the ”#” symbol and the slot extraction rule is the rule with the ”@” symbol. The ”ip” is the input for the constraint rule and the slot extraction rule that is referring to the position of the pattern matching. An example of the frame activation rules is shown below.

!Time;PRPpat,TIMEpat,AMTpat,TIMEpat,TIMEpat;#Time;$0;

@Time_semantic:From;$0,$1,$2,$3,$4

The first part is ”!Time” which is the rule name and can activate to create the time frame. The second part is ”PRPpat,TIMEpat,AMTpat,TIMEpat,TIMEpat” which is the pattern matching in phrase type. If the input EDU is matched with this pattern matching, this rule will be activated. The next part is the constructed information. The ”#Time;$0” is the constraint rule with the input of position 0 of pattern matching. The constraint rule ”#Time” is used to check the condition in its input. If the result of the constraint rule is ”true”, the frame activation rule can be active. The ”$” symbol means to send the content in the phrase to the rule that means to the POS in the ”PRPpat” phrase. The ”@Time_semantic:From” is the slot extraction rule. The ”$0,$1,$2,$3,$4” is the input of the slot extraction rule. The result of the extraction rule will be the information for the ”From” slot in the ”Time” frame if the input pattern is matched.

The constraint rule is the rule for checking conditions and the result is the ”true” or ”false” value. The constraint rule consists of 2 parts: rule name and pattern matching. The rule name starts with the ”#” symbol and follows with the rule name. The syntax

of the constraint rule is defined below. #rn;pt The ”rn” is the rule name and the ”pt” is the pattern matching. An example of the constraint rules is shown below.

#VModifier;VPO@1087,PRP@2419

The first part is ”#VModifier” which is the rule name. The second part is ”VPO@1087,PRP@2419” which is pattern matching. The POS along with the semantic ID is used for matching the input pattern. If the input pattern is 128 matched, the result of the constraint rule is a ”true” value.

The slot extraction rule is the rule for identifying the semantic ID in the pattern as the output of the rule. The slot extraction consists of 4 parts: rule name, pattern matching, semantic identifying, and repeated extraction rule. The rule name starts with the ”@” symbol and follows with the rule name. The syntax of the slot extraction is defined below.

@rn;pt;op[;op][;@[re]]

The ”rn” is the rule name, the ”pt” is the pattern matching, and the ”op” is the position of pattern matching that will result in semantic ID as the output of the rule. The semantic ID can be constructed as the compound semantic ID if it is the compound semantic such as the verbal noun. The ”@” symbol at the last is the recursive slot extraction if that

frame slot contains the series of the semantic ID. An example of the slot extraction rule is shown below.

@Time_semantic;PRP@1594,NCT,FXO,NUM,*;1;3;@

The first part is ”@Time_semantic” which is the rule name. The second part is ”PRP@1594,NCT,FXO,NUM,*” which is the pattern matching. The ”*” symbol in the pattern matching means zero or more patterns. the third part is ”1;3” which means the output of this rule is the semantic ID at position 1 and position 3 of the pattern. The last part is the ”@” symbol means that if there are more 129 patterns, the rest pattern will be sent to another slot extract rule. In pattern matching, If there is the ”*” symbol in the front of the pattern, that means that pattern can exist or not. For example, if the pattern matching is ”HNpat, *DETpat”, the pattern ”HNpat” is still matched. The pattern matching can be recursive in the frame activation rule. For example, if the pattern matching is ”HNpat,!Attribute”, the pattern ”!Attribute” means the recursive pattern in the ”!Attribute” rule. Figure 3.32 shows the example of semantic frame construction rule.

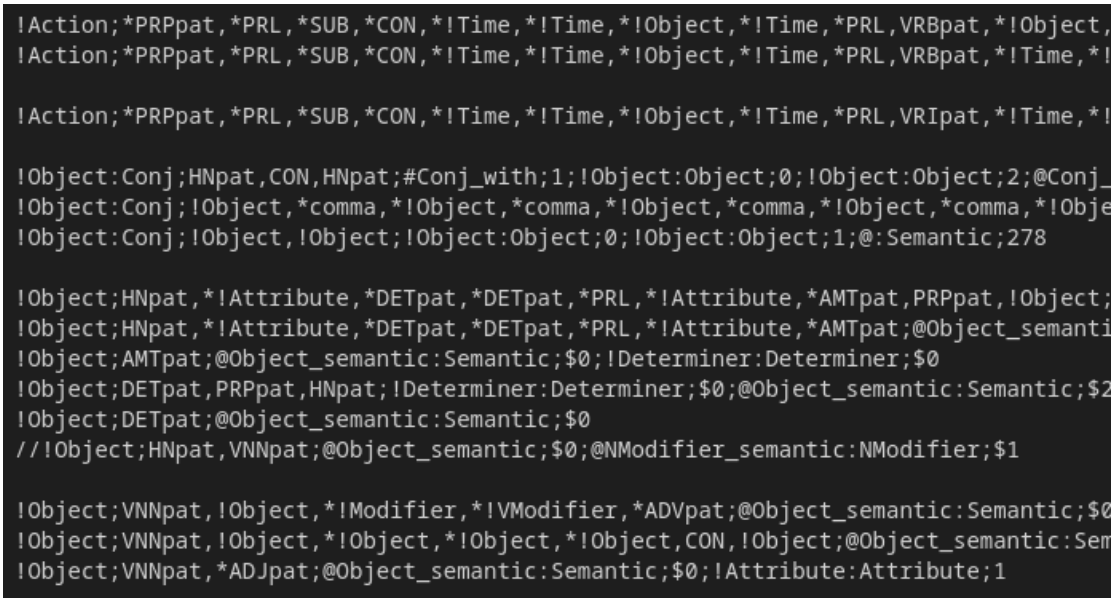


*Figure 3.32.* The example of semantic frame construction rule

### 3.5.2.3 Knowledge Base

The syntactic phrase structure in the corpus will transform into the semantic frame as the knowledge of the question-answering system. Each EDU will be matched to each activation rule to construct the semantic frame with the information in each slot. The semantic frame is converted to the JSON format and stored in the text file as the knowledge base. The syntactic phrase structure is stored along with the semantic frame as a reference. Figure 3.33 shows the example of the knowledge base.

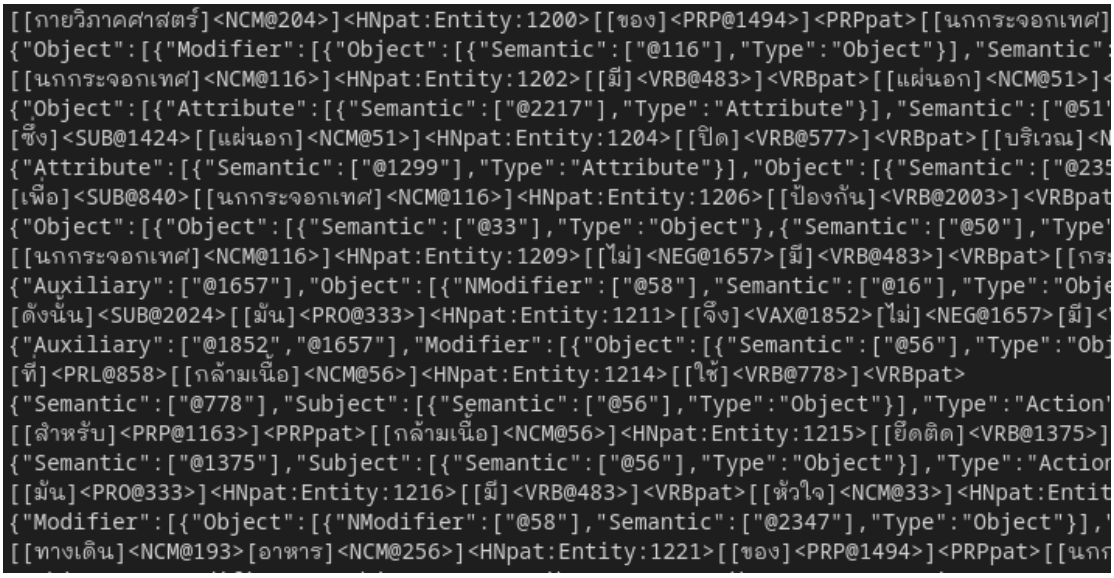


Figure 3.33. The example of knowledge base

### 3.6 Answer Extraction

Answer extraction is a process that extracts the precise answer from the knowledge base by using the question semantic structure that produces from a semantic parser the same process as the building of the knowledge base. The pattern matching technique is the technique that is used in the answer extraction process to identify the semantic ID that is the answer from the knowledge base. Figure 3.34 shows the process of answer extraction from question to answer.

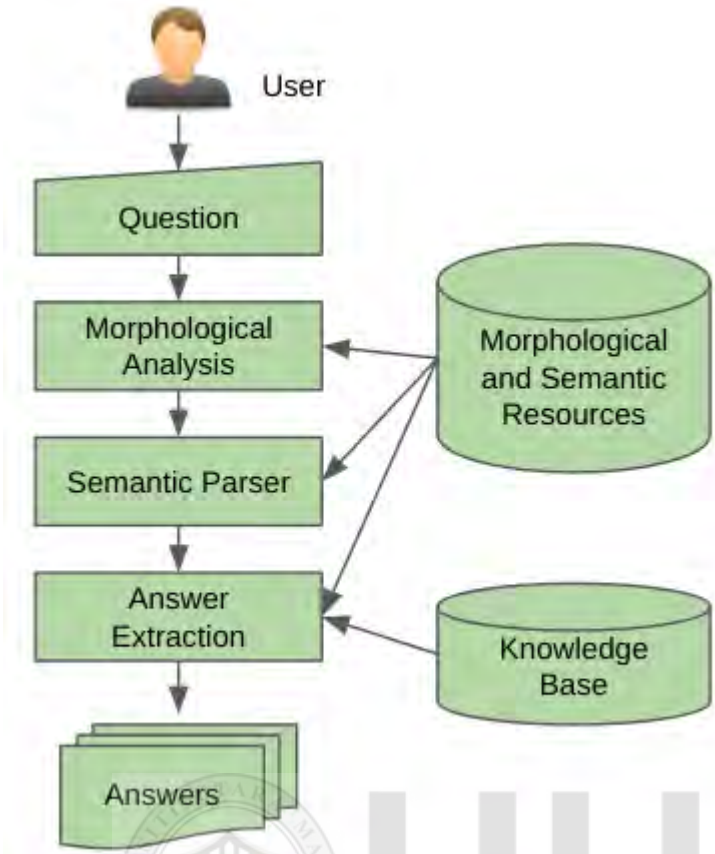


*Figure 3.34.* The process of answer extraction from question to answer



### 3.6.1 Frame Matching

The question sentence will transform to the semantic frame with the semantic ID of the question word for the query intention. The semantic ID of the question word will be located at the position of the frame slot so that the frame matching technique can query the answer from the knowledge base. Table 3.20 shows the question words and the semantic ID.

Table 3.20

*The question words and the semantic ID*

| Question Word | POS | Semantic ID |
|---|---|---|
| ใคร(who) | QOB | @2433 |
| เมื่อไหร่(when) | QAT | @2434 |
| ไหน(where) | QOB | @2435 |
| ใด(which) | QAR | @2436 |
| ใด(what) | QOB | @2437 |
| เท่าไหร่(how much) | QAT | @2438 |
| อะไร(what) | QOB | @2439 |
| อะไร(which) | QAR | @2440 |
| กี่(how many) | QAT | @2441 |



The question sentence transforms to the question semantic frame and applies to query the answer in the knowledge base. Each semantic frame in the knowledge base will retrieve to match with the question semantic frame one by one to identify the semantic ID that could be the answer. After identifying all the answers, the semantic ID will be converted to a Thai word using a semantic dictionary and replied to the user. Figure 3.35 shows the example of frame matching for what question and Figure 3.36 shows the example of frame matching for how many question

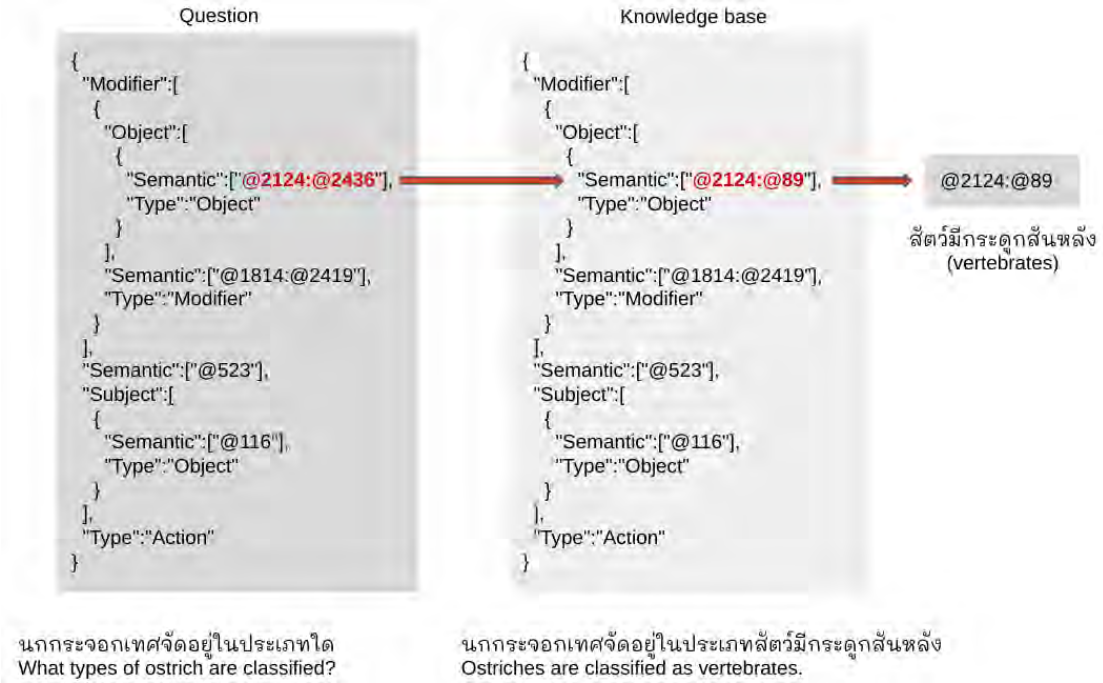


*Figure 3.35.* The example of frame matching for what question.

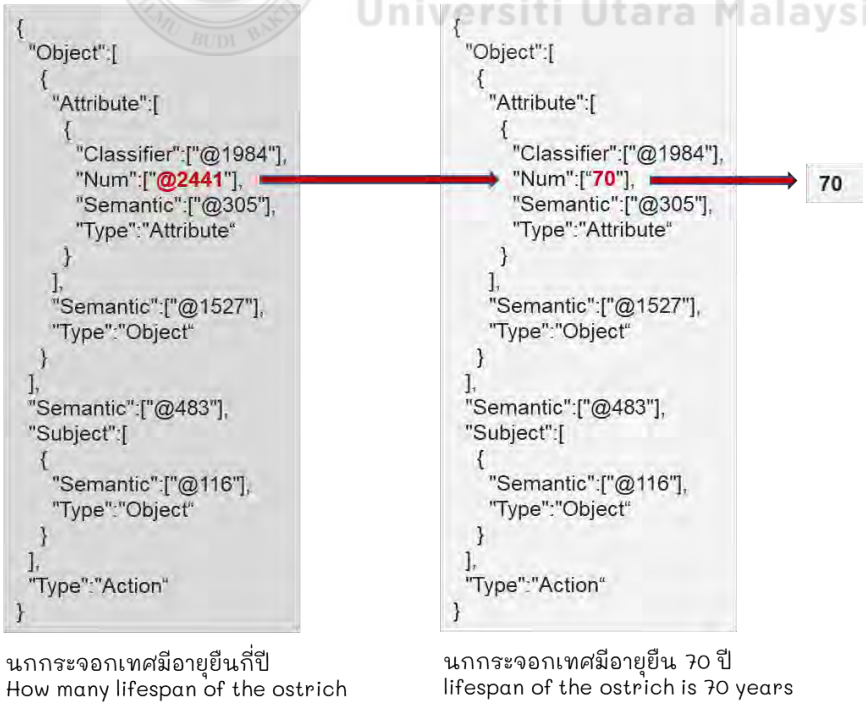


*Figure 3.36.* The example of frame matching for how many question.

### 3.6.2 Fuzzy Matching

Fuzzy logic in natural language processing is an interesting topic to integrate the fuzzy concept into the natural language processing research (Gupta et al., 2018; Novák, 1992, 2017a). In this work, the fuzzy theory is adapted to calculate the information in the attribute frame. The attribute frame is the frame that contains information about the attribute of the object such as color, size, and weight. The adjective in a noun phrase is the origin of the content of the attribute frame. The value in the attribute frame can be a series of semantic IDs and the range of a number of measurements. The membership function is defined for calculating the similarity of the value in the attribute frame between the question semantic frame and the knowledge base. There are 2 types of membership functions for the attribute value: membership function in measurement value and membership function in semantic ID. And then, the confidence value is the aggregate of all of the values from the membership function in the attribute frame.



#### 3.6.2.1 Membership Function in Measurement Value

The attribute of the object can be the measurement value such as weight and height. The value in the attribute mostly appears in range numbers. For example. the sentence ”เสือหนักประมาณ 180 - 245 กิโลกรัม (Tiger weighs about 180 – 245 kilograms.)” show the weight in range 180 to 245 kilograms. If there is the question ”What weighs about 170 kilograms?”, the tiger could be the answer due to its weight being very close, but it will be less confident. The trapezoidal membership function is considered to apply to estimate the membership value of the measurement value in the attribute frame. The trapezoidal membership function is defined in equation 3.11.

$$f(x;l,h) = max\left(min\left(\frac{x-(0.8*l)}{h-(0.8*l)},1,\frac{(1.2*h)-x}{(1.2*h)-h}\right),0\right) \quad (3.11)$$

In equation 3.11, ***x*** stands for the number in the question. ***l*** for the lower value and ***h*** for the higher value. The membership value of the lower value will decrease to 0 when the lower value is decreased to 80% and the membership value of the higher value will decrease to 0 when the higher value is increased to 120%.

#### 3.6.2.2 Membership Function in Semantic ID

The attribute of the object may be not the number but semantic ID instead. The color and taste are examples of the attribute of the object that is word value. The membership function for the word value attribute can be estimated by counting the joint semantics between both attribute frames. The membership function in semantic ID between the attribute frame can be defined in equation 3.12.

$$f(A_1,A_2) = \frac{count(A_1 \cap A_2)}{count(A_1 \cup A_2)} \quad (3.12)$$

In equation 3.12, A1 stands for the semantic ID member of the attribute frame 1 and A2 for the semantic ID member of the attribute frame 2.

### 3.6.2.3 Confidence Value

The membership value of each attribute frame will aggregate to one value for the whole semantic frame called the confident value. The confident value is the value that reflects the consistency matching between the semantic frame. The production of all the membership values of each attribute is considered to calculate the confident value of the matching.The confident function is defined in equation 3.13.

$$\Gamma(\theta_1, \theta_2) = \prod_i f_i(\theta_1, \theta_2) \tag{3.13}$$

$\boldsymbol{\theta_1}$ stands for semantic frame 1 and $\boldsymbol{\theta_2}$ for semantic frame 2. $\boldsymbol{i}$ for the number of attribute frames in both semantic frames and $\boldsymbol{f}$ for the membership function.



## 3.7 Evaluation

The criteria of evaluation of the question answering system that is proposed by Breck et al. (2000) will determine that the correctness and completeness are chosen to evaluate the question answering system with the precision, recall, and F1 measurement. Suppose $\boldsymbol{\alpha}$ is the number of facts that is retrieved from the system, $\boldsymbol{\beta}$ is the number of corrected fact that is retrieved from the system, and $\boldsymbol{\eta}$ is the number of all corrected fact in the system. Then we can define the precision, recall and F measurement that are shown in equation 3.14, equation 3.15, and equation 3.16 as below:

$$P = \frac{\beta}{\alpha} \tag{3.14}$$

$$R = \frac{\beta}{\eta} \tag{3.15}$$

$$F1 = \frac{2PR}{P+R} \tag{3.16}$$

### 3.8 Summary

After the whole process, the text is analyzed by the process of morphological analysis and then by a semantic parser to create the knowledge in frame-based representation. In the user view, the question is entered to generate the question semantic structure. The question semantic structure is used by answer extraction to extract the semantic ID from knowledge-based and then generate the answer for the user.

# CHAPTER FOUR

# EXPERIMENTAL RESULTS

This section describes the experimental results from each methodology. The results are stated in the word segmentation and POS tagging. Then the following are the EDU segmentation process, the ellipsis and anaphora resolution, the word sense disambiguation, and answer extraction.

## 4.1 Word Segmentation and POS Tagging

Our corpus for training contains a total of 18,248 words with POS tagging. Dictionary contains 2,171 words for word segmentation and 52 words for the shallow parser. There are 235 POS patterns for the word correction process. There are 58,750 features of the CRF model for word segmentation and 11,975 features for the shallow parser. There are 114 syntactic patterns and 12 EDU reconstruction rules in the EDU segmentation process. The precision, recall, and F1 score are used to evaluate the algorithm. The measures can be defined as follows.

$$Precision = \frac{\#\ of\ correct\ result\ by\ system}{\#\ of\ result\ determined\ by\ system} \quad (4.1)$$

$$Recall = \frac{\#\ of\ correct\ result\ by\ system}{\#\ of\ result\ in\ domain} \quad (4.2)$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \quad (4.3)$$

The results of word segmentation are shown in Figure 4.1 that shows the bar chart of the results of word segmentation. CRF shows a good performance for word segmentation. However, CRF tends to segment words that could be one word to be more chunks. Word segmentation by machine learning (Aroonmanakun, 2002; Kawtrakul & Thumkanon, 1997; Kongyoung et al., 2015; Kruengkrai et al., 2006) produces some errors due to segmented words into more chunks. The extra process is needed to gain more precision in the word segmentation process. To increase the precision of results, the word segmentation correction and the POS re-tagging process are applied. The fragment word is reconstructed into a correct word-by-word segmentation correction process that can boost the precision and recall significantly. The re-tagging process can increase a little of precision and recall.

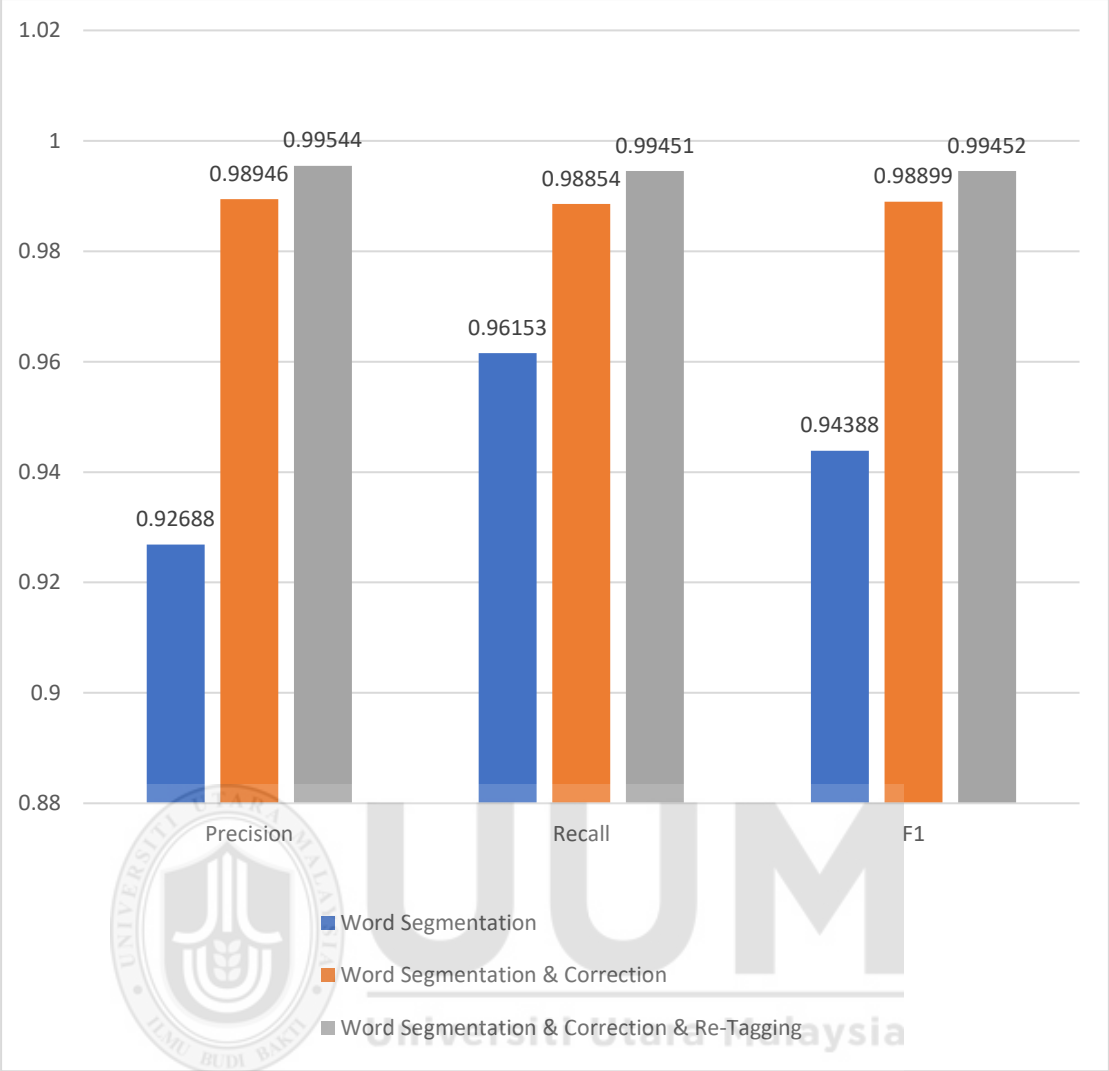


*Figure 4.1.* Bar chart of the results of word segmentation

## 4.2 EDU Segmentation

The clue marker is useful to partially segment EDU. Some parts of EDU need syntactic information to indicate the point of EDU Segmentation and then the EDU reconstruction by rule-base is applied to build the precise EDU structure. The evaluation of the EDU segmentation by precision, recall, and F1 measurement.

The evaluation will be measured by each state of EDU segmentation that is proposed. Figure 4.2 shows the bar chart of the results of EDU segmentation.

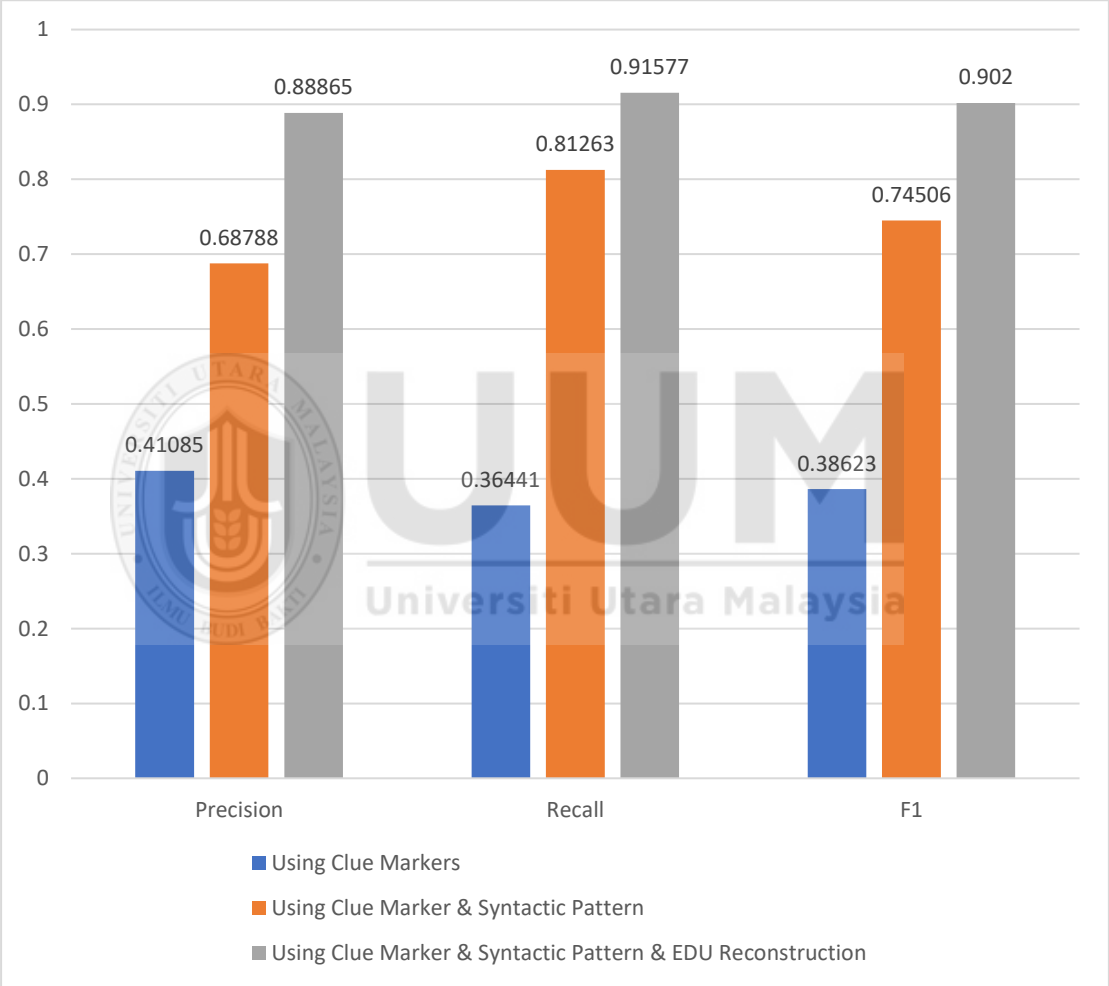


*Figure 4.2.* Bar chart of the results of EDU segmentation

The results show that the use of a clue marker alone produces the EDU segment with low precision to 0.41085 and recall to 0.36441. Many EDUs appear without a clue marker to indicate the boundary. However, a clue marker can be a starting tool to

indicate the EDU boundary (Chareonsuk et al., 2005). Syntactic information is used in various ways to indicate the boundary of EDU (Ketui et al., 2012; Sinthupoun & Sornil, 2010). Syntactic patterns from a shallow parser can be a good option to identify the point of EDU segmentation. The use of syntactic pattern significantly increase the precision and recall, however, noun list EDU and the use of space in some writing style still indicate the wrong EDU boundary. Finally, the EDU reconstruction process is used to combine noun list EDU and some fragment EDUs to produce a more precise EDU with 0.88865 precision and 0.91577 recall.

### 4.3 Ellipsis and Anaphora Resolution

Our corpus for training contains a total of 18,248 words and 2,327 EDUs. There are 3,934 entities, 1,272 zero anaphora, 126 nominal anaphora, 64 pronominal anaphora, and 88 ellipses of the owner in the corpus. The precision, recall, and F1 score are used to evaluate the algorithm.



After finished all processes, The results will be evaluated from the anaphora determiner, resolution for non-referential anaphora, and resolution for referential anaphora. Each kind of anaphora is evaluated separately and also overall. The results of the anaphora resolution are shown in Table 4.1.

Table 4.1

*Results of the anaphora resolution*

| **Anaphora Types** | **Precision** | **Recall** | **F1** |
|---|---|---|---|
| Zero anaphora (non-referential) | 0.66475 | 0.91699 | 0.77075 |
| Zero anaphora (referential) | 0.78557 | 0.80176 | 0.79358 |
| Zero anaphora (overall) | 0.75521 | 0.82468 | 0.78841 |
| Pronominal anaphora (non-referential) | 1 | 1 | 1 |
| Pronominal anaphora (referential) | 1 | 1 | 1 |
| Pronominal anaphora (overall) | 1 | 1 | 1 |
| Nominal anaphora (non-referential) | 1 | 1 | 1 |
| Nominal anaphora (referential) | 0.96153 | 0.96153 | 0.96153 |
| Nominal anaphora (overall) | 0.99206 | 0.99206 | 0.999206 |
| Ellipsis of the owner (non-referential) | 0.70000 | 1 | 0.82352 |
| Ellipsis of the owner (referential) | 0.87837 | 0.87837 | 0.87837 |
| Ellipsis of the owner (overall) | 0.84042 | 0.89772 | 0.86812 |
| Overall | 0.77744 | 0.84967 | 0.81195 |

Zero anaphora is the kind of anaphora that mostly appears in EDUs. The results show a good precision of 0.75521 and a recall of 0.82468. The pronominal anaphora is finished with the amazing results that precision is 1 and recall is 1. These results are successful without using additional knowledge such as gender, and number. Because the use of pronominal anaphora in the corpus is not a complicated scenario. Then the only use of the surface word and syntactic information can produce good results in our

corpus. The nominal anaphora also recorded high precision of 0.99206 and a recall of 0.99206. The ontology that provides hyponymy knowledge is useful to resolve the nominal anaphora. The surrounding words in nominal anaphora and reference are also significant to resolving the ranking for nominal anaphora resolution. The ellipsis of the owner recorded high precision of 0.84042 and a recall of 0.89772. The ontology that provides the meronymy is a significant background knowledge that can be used to identify the entity that is a part of something, especially in the agriculture corpus. The overall results show that the precision is 077744, the recall is 0.84967, and the F1 is 0.81195.

### 4.4 Word Sense Disambiguation

The corpus with semantic tagging contains 14,820 semantics. The semantic dictionary contains 2,488 semantics. There is a total of 46,001 features for semantic disambiguation. The evaluation is measured by the correctness performance in the percentage of correct semantics resolved. The correct semantic resolved is a total of 14,799 semantics and the incorrect semantic is only 21 semantics. The correctness performance is 99.85%

The result of the word sense disambiguation is high due to the feature set is useful to determine the semantics of the surface word. There are a few incorrect semantics that mostly is the word of a preposition. Some preposition words can be many semantic and seem confusing to the system even if locate in a different context. The bigger corpus and context could provide better training that produces better correctness performance.

### 4.5 Answer Extraction

Answer extraction is an important part of the question-answering system. The question sentence is transformed into the question semantic frame and used to match with the knowledge base to query the precise answer. The question is predefined to evaluate the question answering system. There are a total of 55 questions are predefined to test the system. The anaphora is an important role to interpret the fact in a knowledge base. The knowledge that is generated without the anaphora resolution will interpret the incomplete fact. To find the influence of the anaphora in text, The system is evaluated with the condition that the knowledge base is resolved with the anaphora resolution and without the anaphora resolution. The fuzzy matching is also a factor in the answer extraction. The system is also evaluated with the condition that the answer is extracted with fuzzy matching and without fuzzy matching. Table 4.2 shows the examples of the predefined question.

Table 4.2

*The examples of the predefined question*

| |
|---|
| ชวน หลีกภัยเป็นใคร (Who is Chuan Leekpai?) |
| ชวน หลีกภัยเกิดเมื่อไหร่ (When was Chuan Leekpai born?) |
| ธนาคารไทยได้เปลี่ยนชื่อเป็นอะไร (What name was the Thai bank changed to?) |
| นกกระจอกเทศมีถิ่นกำเนิดในทวีปใด (Which continent does the ostrich originate from?) |
| เสือสืบสายเลือดมาจากอะไร (What is the lineage of tigers?) |
| ลูกแมวแรกเกิดหนักเท่าไหร่ (How much does a newborn kitten weigh?) |
| อะไรหนักประมาณ 150 กิโลกรัม (What weighs about 150 kilograms) |
| อะไรที่โตเต็มที่จะมีสีเทา (What is full-grown will be gray?) |



The answer for some questions can be more than one answer depending on the number of facts in the knowledge base. The anaphora resolution can affect the number of the answer that can extract from the knowledge base. Table 4.3 shows the example of the question and the answers that extract from answer extraction with anaphora resolution.

Table 4.3

*The example of the question and the answers that extract from answer extraction with and without the anaphora resolution*

| **Question** | **Answer with Anaphora Resolution** | **Answer without Anaphora Resolution** |
|---|---|---|
| ชวน หลีกภัยเป็นใคร<br>(Who is Chuan Leekpai?) | นักการเมืองชาวไทย<br>(Thai politician)<br><br>คนรูปร่างเล็ก<br>(small body person)<br><br>สมาชิกสภาผู้แทนราษฎร<br>จังหวัดตรัง<br>(Member of the House of Representatives of Trang Province)<br><br>หัวหน้าพรรคประชาธิปัตย์<br>(Democratic Party Leader) | นักการเมืองชาวไทย<br>(Thai politician) |



Fuzzy matching also affects the number of results. The attribute value such as color, taste, and flavor can be in the form of the blend semantic for example the word ”ฟ้าอมเทา (grayish blue)”. If the question is asked about something blue then the answer may come with a confidence value. The answer extraction without the fuzzy matching may produce fewer answers or none from the knowledge base in any case. Table 4.4 shows the example of the question and the answers that extract with and without fuzzy matching.

Table 4.4

*The example of the question and the answers that extract with and without fuzzy matching*

| **Question** | **Answer with Fuzzy Matching** | **Answer without Fuzzy Matching** |
|---|---|---|
| อะไรมีสีเทาดำ<br>(What is gray black?) | ลักษณะผิวหนังนกกระจอกเทศ<br>(Ostrich skin characteristics)<br><br>นกกระจอกเทศตัวเมียโตเต็มที<br>(Fully grown female ostrich) | ลักษณะผิวหนังนกกระจอกเทศ<br>(Ostrich skin characteristics) |

The answer extraction determines the correctness and the completeness by evaluating the precision, the recall, and the F1 measurement. Figure 4.3 shows the bar chart of the results of the answer extraction.

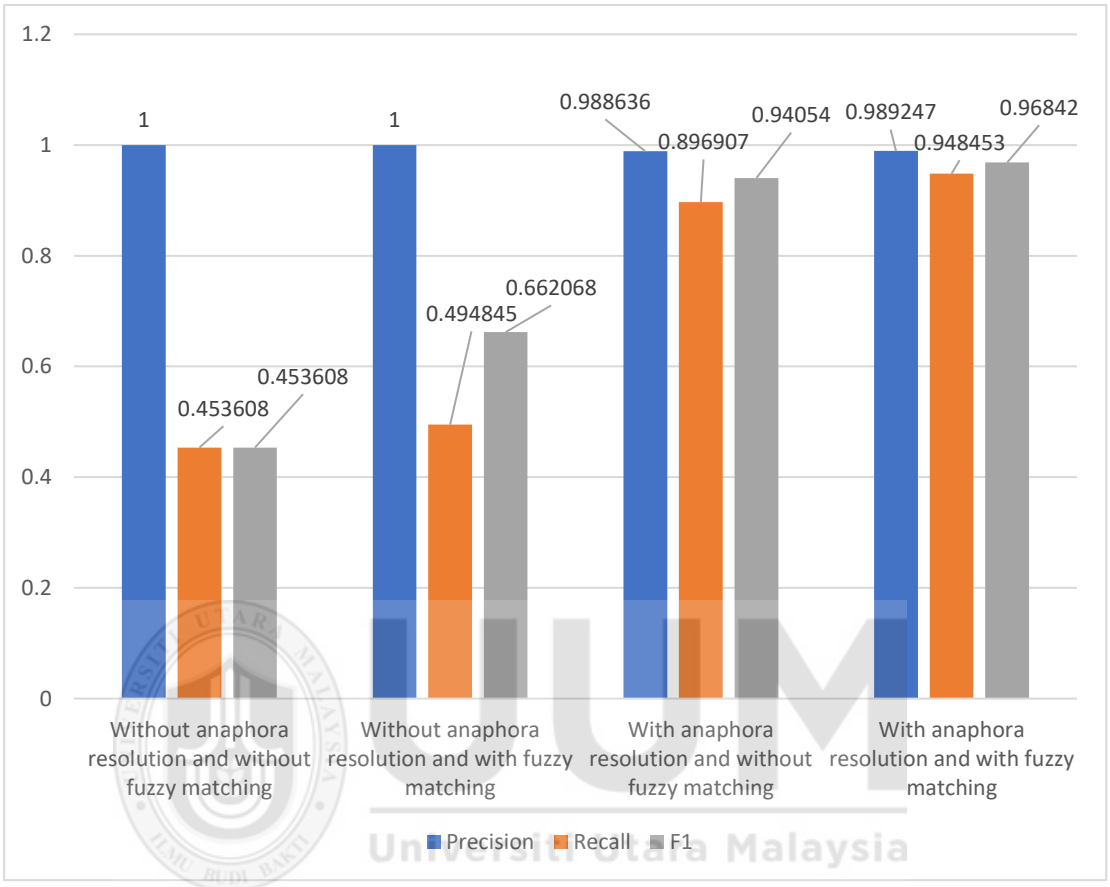


*Figure 4.3.* Bar chart of the results of the answer extraction

The results show that the answer extraction without anaphora resolution shows high precision but very low recall. The precision is high because the result is extracted directly from the sentence that the object is explicitly appear in the sentence. The recall is low because some answers that are in the form of anaphora cannot be retrieved from the knowledge base. If the object that is in the form of anaphora is not resolved then the facts in the knowledge base are incomplete due to the anaphora is not resolved making the object missing and causing the recall to be low. The fuzzy matching can

boost the recall a little because the fact that is the value can be retrieved more from the knowledge base.

### 4.6 Explanation with The Other Thai QAS

The question answering system in the Thai language is still rare research. QAST (Jitkrittum et al., 2009) is the question answering system that acquires Thai Wikipedia as the data source. This system translates the Thai Wikipedia articles to the RDF knowledge base and then uses the SPARQL to query the answer in 150 the RDF knowledge base. The anaphora resolution is not involved to process the article before generating the RDF knowledge. The performance of this system is 0.47 which is not high. Kongthon et al. (2011) presents the semantic-based question answering system for Thailand tourism information. This system works on a close domain and collected the example request from Pantip.com. The knowledge is created to the tour ontology and uses SPARQL to query the answer from this ontology. The question is translated into a semantic template to identify the answer type from the ontology. The system performance is high. However, this system operates in a close domain in tourism. WabiQA (Noraset et al., 2021) is the question answering system that works on data intensive using deep learning to identify the piece of answer from Thai Wikipedia. This system works on the open domain by ranking the document that is related to the question and reading the document to identify the answer. The performance is good, however, the anaphora resolution is still not involved to resolve the hidden entity in the document. Table 4.5 shows the explanation of the Thai question answering system.

Table 4.5

*The explanation of the Thai question answering system*

| QAS | Domain | Technique | Explanation |
|---|---|---|---|
| QAST | Open | Knowledge Intensive on Subgraph Matching | - Anaphora Resolution is not involved<br>- No Fuzzy Matching<br>- Performance is not high |
| QAS on Tourism | Close on Tourism | Knowledge Intensive on Template-Based | - Anaphora Resolution is not involved<br>- No Fuzzy Matching<br>- Performance is good |
| WabiQA | Open | Data Intensive on Deep Learning | - Anaphora Resolution is not involved<br>- No Fuzzy Matching<br>- Performance is good |
| Our System | Open | Knowledge Intensive on Template-Based | - Anaphora Resolution is involved<br>- Fuzzy Matching is applied<br>- Performance is high |



Our system produces a high performance to extract the answer because of the application of the anaphora & ellipsis resolution and also the fuzzy matching. The entities in the document that could be the answer can be hidden in the form of anaphora. We can see that the anaphora resolution is affected the performance of the system. Also, fuzzy matching can help to resolve the attribute which is the measured value and word value. However, the knowledge base that is produced by using the complex NLP technique could be expensive. The incremental knowledge base is a good resource of knowledge base for the semantic base question answering system. The data for the complex NLP task could be incremented from time to time.

## 4.7 Summary

The semantic-based question answering system is the system that needs the essential processes of the natural language processing from Thai word segmentation, Thai EDU segmentation, ellipsis and anaphora resolution, semantic parser, and the final is the answer extraction. Thai word segmentation is improved by the more complex process to gain more precision and recall. Thai EDU segmentation is the complicated process to segment the Thai EDU from the continual textual paragraph. A pipeline of processes is developed to identify the EDU segment and reconstruct the EDU to make the high precision of Thai EDU segmentation. Ellipsis and anaphora resolution is the essential component to make the completed fact in Thai EDU segmentation for high performance of the semantic base question answering system. The fuzzy matching also boosts the performance of the question-answering system. To validate the question-answer pair, 30 native Thai volunteers and 3 expert reviewers are the validators and reviewed.

# CHAPTER FIVE

# CONCLUSION

This section discusses the conclusion of the research. The conclusion includes the research summary, limitations, and research contribution.

## 5.1 Research Summary

### 5.1.1 Thai EDU Segmentation

A pipeline of the process for Thai EDU segmentation is done by using the syntactic information from a shallow parser. The first step, word segmentation, and POS tagging process are done by using CRF to identify the word boundary and its POS with 0.92688 precision and 0.96153 recall. The CRF shows good results in word segmentation, however, some composition words tend to segment into fragment words. To improve the precision of word segmentation and the POS tagging process, we implemented the word segmentation correction by using a POS pattern with a dictionary to merge some fragment words and improve the precision to 0.98946 and recall to 0.98854. POS re-tagging process is used to redetermine the POS label further improving the precision and recall of POS tagging to 0.99544 and 0.99451 respectively.

In the second step, the shallow parser is used to determine the POS sequence to phrase types such as head noun, amount, adjective, and determiner pattern. The EDU segmentation process is done by using a clue marker, syntactic pattern, and reconstruction rule-based. The clue marker indicates the EDU segment with 0.41085 precision and 0.36441 recall. The syntactic pattern is used to identify the EDU segment

with the improvement of precision and recall to 0.68788 and 0.81263 respectively. The fragment of noun list EDUs and some partial EDUs are reconstructed by rule-based and finally improve the precision and recall to 0.88865 and 0.91577 respectively.

From the corpus development, we found that some writing styles use the space to make it more readable in some words or some parts of the sentence. The non-formal use of space can cause some difficulties to identify the boundary of EDU and also increase ambiguity. A preposition and noun list are the most apparent non-formal use of space in our corpus. Moreover, POS label designing is also an important component to make a more reliable grammar structure in the corpus. POS labels with embedded semantics can be a useful resource to analyze the structure of EDU and sentences.



### 5.1.2 Ellipsis and Anaphora Resolution

The methodology to resolve the anaphora in Thai EDU segmentation is done by using the background knowledge to resolve the hyponymy and meronymy relation between the anaphora and the references. The algorithm contains three parts: anaphora determiner, resolution for non-referential anaphora, and resolution for referential anaphora. The first step is the algorithm to determine the kind of anaphora in each EDU. The algorithm searches each entity in EDU and analyzes the word and the surround together with the ontology to decide the kind of the anaphora. The second step is the resolution for non-referential anaphora. The resolution utilized the ranking model to identify whether anaphora is a non-referential or a referential anaphora. This resolution works on the only use of the surface word and the surround for learning the model. The final step is the resolution for referential anaphora. The candidate references are

generated from the entities in each EDU up to 10 prior EDUs. The ranking model computes the probabilistic value in each candidate and then chooses the candidate with the highest probabilistic value for the referential anaphora. The overall results are that the precision is 0.77744, the recall is 0.84967, and the F1 score is 0.81195. In addition, this paper mentions the anaphora types that could be concerned in Thai anaphora resolution especially the ellipsis of the owner. The non-referential anaphora is also significant and could not be overlooked. However, this work is based on the collected corpus that could not be comprehensive. Changing domain can be affected the results and could be needed additional features and also further background knowledge. To make reliable the anaphora resolution, the making of the comprehensive corpus on various domains and also the modification features could be the topic to be done.



### 5.1.3 Semantic Parser

The semantic parser is done by using the syntactic information from the shallow parser to identify the syntactic pattern and then construct the semantic frame. To build the high-precision semantic parser, the process of word sense disambiguation is developed to identify the correct semantic of the surface word. The semantic is identified by learning from the context that surrounds the surface word to decide the correct semantic for the surface word. The performance of the word sense disambiguation is high at 99.85%. The incorrect semantic is mostly the word of a preposition. Some prepositions in Thai text can be many semantic and cause confusion in the system.

### 5.1.4 Answer Extraction

The answer extraction is the main role component of the semantic base question answering system for the user. Frame matching is the technique that is used to identify the fact from the knowledge base and then give the answer to the user. The ellipsis and anaphora resolution are an important process to complete the EDU and make the complete fact to the knowledge base. The high recall of semantic base question answering system cannot develop by no ellipsis and anaphora resolution process. The fuzzy matching is also an additional process to help the question answering system to retrieve more answers with the confidence value to the user.

## 5.2 Limitations

The source of this work is a freestyle text in an open domain document. However, the accomplishment of this work is based on some limitations as follows.



**No misspelling word:** Misspelling words can occur in the text normally. The whole system expects accurate spelling text as a data source. The misspelling word correction is not in our interesting scope of work. Spelling word correction is done manually before feeding to the whole system.

**Ignore the additional information from parenthesis:** There is the use of parenthesis in the text to provide additional data such as abbreviations, synonyms, and foreign words. Text in parenthesis can be in free form depending on the aspect of the writer. In this work, text in parenthesis is ignored and pruned by the corpus cleaning process.

**Do not follow the structure information from a table:** Structure information, such as a table, is widely used in the text. A table is not a part of EDU in the text and is not in our interesting scope of work. Tables in the source corpus are pruned by the corpus cleaning process.

### 5.3 Research Contribution

This research is aim to develop a semantic base question answer system that acquires knowledge from Thai text. There is the research contribution as follows.

**Thai Word Segmentation:** The Thai word segmentation is improved by the pipeline of the additional process to gain more precision and recall. The high performance of Thai word segmentation will be a great impact on the research in the Thai natural language processing field.



**Thai EDU Segmentation:** The Thai EDU segmentation is a complicated process. Clue word is a useful resource to identify the simple point of the EDU segment in Thai text. Additional complicated processes are proposed to analyze the segmentation of Thai EDU. The rule-based technique is utilized to finish the high performance of the Thai EDU segmentation process.

**Ellipsis and Anaphora Resolution:** Semantic-base question answering system needs the resolution to solve the reference in the ellipsis and anaphora. The feature for the learning model of ellipsis and anaphora resolution is proposed.

Research in Thai anaphora resolution is still rare and the proposed model can impact the anaphora research field in Thai text.

**Semantic Parser:** Semantic parser for Thai text is proposed with the semantic frame construction rule. The semantic frame construction rule is a great tool to develop text understanding in the future work of the Thai natural language processing research field.

**Fuzzy Matching:** To retrieve more answers from the knowledge base, the fuzzy matching technique for the measurement value and semantic is proposed to match the semantic frame with the confidence value. The membership function is proposed to achieve the purpose of fuzzy matching to gain more recall for the semantic-based question answering system.



## 5.4 Future Work

This study accomplishes the answering of the "What" and "How many" Questions. The development of a corpus needs to be more comprehensive to cover the most words and semantics used in daily life and questions. To answer the "How" and "Why" questions, the study of the relation between sentence and paragraph is needed to resolve. The relation of sentence is important in other domain applications such as text summarization and best practice resolution. The annotation for resolving the sentence and paragraph relation has to be designed in the further corpus. Also, the knowledge frame should be extended to support the new semantics of the relation between frames.

# REFERENCES

Abbasiantaeb, Z., & Momtazi, S. (2021). Text-based question answering from information retrieval and deep neural network perspectives: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 11(6), e1412.

Ali, N. Y., Das, J. K., Al-Mamun, S. M. A., & Nurannabi, A. M. (2008). Morphological analysis of bangla words for universal networking language. *3rd International Conference on Digital Information Management, ICDIM 2008*, 532–537. https://doi.org/10.1109/ICDIM.2008.4746734

Andrade, R. A. E., Fernández, E., & González, E. (2014). Compensatory fuzzy logic: A frame for reasoning and modeling preference knowledge in intelligent systems. In *Soft computing for business intelligence* (pp. 3–23). Springer.

Antoniou, C., & Bassiliades, N. (2022). A survey on semantic question answering systems. *The Knowledge Engineering Review*, 37.

Arreerard, R., Mander, S., & Piao, S. S. (2022). Survey on Thai NLP language resources and tools. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 6495-6505).

Aroonmanakun, W. (1989). A dependency analysis of thai sentences for a computerized parsing system [Doctoral dissertation, Chulalongkorn University].

Aroonmanakun, W. (1997). Referent Resolution for Zero Pronouns in Thai. Southeast Asian Linguistics Studies In *Honor of Vichin Panupong*, (October), 11–24.

Aroonmanakun, W. (2000). Zero Pronoun Resolution in Thai : A Centering Approach. Burnham, Denis, et. al. Interdisciplinary Approaches to Language Processing: *The International Conference on Human and Machine Processing on Human and Machine Processing of Language and Speech*. NECTEC: Bangkok, 127–147.

Aroonmanakun, W. (2002). Collocation and Thai Word Segmentation. *Proceedings of the Fifth Symposium on Natural Language Processing & The Fifth Oriental COCOSDA Workshop*.

Auer, S., & Lehmann, J. (2007). What Have Innsbruck and Leipzig in Common ? Extracting Semantics from Wiki Content. *European Semantic Web Conference*, 503–517. https://doi.org/10.1007/978-3-540-72667-8_36

Baker, C., Schneider, N., Petruck, M. R., & Ellsworth, M. (2015). Getting the Roles Right: Using FrameNet in NLP. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, 10–12.

Bakhtyar, M., Kawtrakul, A., & NECTEC, P. (2011). Integrating knowledge resources and shallow language processing for question classification. *The KRAQ11 Workshop: Knowledge and Reasoning for Answering Questions*, 22.

Ballesteros, M., Dyer, C., Goldberg, Y., & Smith, N. A. (2017). Greedy transition-based dependency parsing with stack lstms. *Computational Linguistics*, 43(2), 311–347.

Ballesteros, M., Goldberg, Y., Dyer, C., & Smith, N. A. (2016). Training with Exploration Improves a Greedy Stack-LSTM Parser. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2005–2010. http://arxiv.org/abs/1603.03793

Bao, J., Duan, N., Zhou, M., & Zhao, T. (2014). Knowledge-Based Question Answering as Machine Translation. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), 967–976. https://doi.org/10.3115/v1/P14-1091

Benajiba, Y., Rosso, P., & Benedíruiz, J. M. (2007). Anersys: An arabic named entity recognition system based on maximum entropy. In *Computational Linguistics and Intelligent Text Processing: 8th International Conference, CICLing 2007, Mexico City, Mexico, February 18-24, 2007. Proceedings 8* (pp. 143-153). Springer Berlin Heidelberg.

Blum, A. (1997). Empirical support for winnow and weighted-majority algorithms: Results on a calendar scheduling domain. *Machine Learning*, 26(1), 5–23.

Boguslavsky, I., Frid, N., Iomdin, L., Kreidlin, L., Sagalova, I., & Sizov, V. (2000). Creating a Universal Networking Language module within an advanced NLP system.

*Proceedings of the 18th conference on Computational linguistics*, 1(99), 83–89. https://doi.org/10.3115/990820.990833

Boonkwan, P., & Supnithi, T. (2017). Bidirectional Deep Learning of Context Representation for Joint Word Segmentation and POS Tagging. *International Conference on Computer Science, Applied Mathematics and Applications*, 184–196.

Boonkwan, P., Luantangsrisuk, V., Phaholphinyo, S., Kriengket, K., Leenoi, D., Phrombut, C., ... & Supnithi, T. (2020). The annotation guideline of lst20 corpus. *arXiv preprint arXiv:2008.05055.*

Bordes, A., Usunier, N., Chopra, S., & Weston, J. (2015). Large-scale Simple Question Answering with Memory Networks. arXiv preprint arXiv:1506.02075. https://doi.org/10.1016/j.geomphys.2016.04.013

Bos, J. (2015). Open-Domain Semantic Parsing with Boxer. *Proceedings of the 20th nordic conference of computational linguistics (NODALIDA 2015)*, 301–304.

Breck, E., Burger, J., Ferro, L., Hirschman, L., House, D., Light, M., & Mani, I. (2000). How to evaluate your question answering system every day and still get real work done. *Lrec, 6*. http://arxiv.org/abs/cs/0004008

Burger, J., Cardie, C., Chaudhri, V., Gaizauskas, R., Harabagiu, S., Israel, D., Jacquemin, C., Lin, C. Y., Maiorano, S., Miller, G., Moldovan, D., Ogden, B., Prager, J., Riloff, E., Singhal, A., Shrihari, R., Strzalkowski, T., Voorhees, E., & Weishedel, R. (2001).

Issues , Tasks and Program Structures to Roadmap Research in Question & Answering ( Q & A ). *Document Understanding Conferences Roadmapping Documents*, (March 2017), 1–35.

Carlson, A., Cumby, C., Rosen, J., & Roth, D. (1999). The SNoW learning architecture. *Technical report UIUCDCS.*

Chang, C.-C., & Lin, C.-J. (2001). LIB-SVM: a library for support vector machines. http://www.csie.ntu.edu.tw/%7B~%7Dcjlin/libsvm

Chanlekha, H., & Kawtrakul, A. (2004). Thai Named Entity Extraction by incorporating Maximum Entropy Model with Simple Heuristic Information. *Proceedings of IJCNLP*.

Chareonsuk, J., Sukvakree, T., & Kawtrakul, A. (2005). Elementary discourse unit segmentation for Thai using discourse cue and syntactic information. *Proceedings of SNLP.*

Charoenpornsawat, P., Kijsirikul, B., & Meknavin, S. (1998). Feature-based proper name identification in Thai. *Proceedings of National Computer Science and Engineering Conference: NCSEC'98.*

Chen, D., Fisch, A., Weston, J., & Bordes, A. (2017). Reading wikipedia to answer open-domain questions. arXiv preprint arXiv:1704.00051.

Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on Information Theory*, 2(3), 113–124.

Choudhary, B., & Bhattacharyya, P. (2002). Text clustering using universal. networking language representation. *The proceedings of Eleventh International World Wide Web Conference.*

Cocke, J. (1970). Programming languages and their compilers: Preliminary notes.

Cristianini, N., & Shawe-Taylor, J. (2000). An introduction to support vector machines. Cambridge University Press.

Das, D., Schneider, N., Chen, D., & Smith, N. A. (2010). Probabilistic frame-semantic parsing. HLT '10 Human Language Technologies: *The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 3(June), 948–956.

Day, M. Y., Ong, C. S., & Hsu, W. L. (2007). Question classification in English-Chinese cross-language question answering: An integrated Genetic Algorithm and Machine Learning approach. *2007 IEEE International Conference on Information Reuse and Integration, IEEE IRI-2007*, 203–208. https://doi.org/10.1109/IRI.2007.4296621

Denis, P., & Baldridge, J. (2008). Specialized models and ranking for coreference resolution. *Proceedings of the 2008 conference on empirical methods in natural language processing*, 660–669.

Dhindsa, B. K., & Sharma, D. V. (2016). Translation Challenges and Universal Networking Language. *International Journal of Computer Applications*, 133(15), 36–40.

Di Eugenio, B. (1990). Centering theory and the Italian pronominal system. *Proceedings of the 13th conference on Computational linguistics*- Volume 2, 270–275.

Di Eugenio, B. (1996). The discourse functions of Italian subjects: a Centering approach. *Proceedings of the 16th Conference on Computational Linguistics (COLING'96)*, 352–357.

Dyer, C., Ballesteros, M., Ling, W., Matthews, A., & Smith, N. A. (2015). Transition-Based Dependency Parsing with Stack Long Short-Term Memory. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, 334–343. http://arxiv.org/abs/1505.08075

Earley, J. (1970). An efficient context-free parsing algorithm. *Communications of the ACM*, 13(2), 94–102. https://doi.org/10.1145/362007.362035

Edmonds, J. (1967). Optimum branchings. *Journal of Research of the National Bureau of Standards Section B Mathematics and Mathematical Physics*, 71B(4), 233. https://doi.org/10.6028/jres.071B.032

Eisner, J. M. (1996). Three new probabilistic models for dependency parsing: An exploration. *Proceedings of the 16th conference on Computational linguistics*-Volume 1, 340–345.

Fader, A., Zettlemoyer, L., & Etzioni, O. (2014). Open question answering over curated and extracted knowledge bases. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '14*, 1156–1165. https://doi.org/10.1145/2623330.2623677

Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A. a., Lally, A., Murdock, J. W., Nyberg, E., Prager, J., Schlaefer, N., & Welty, C. (2010). Building Watson: An Overview of the DeepQA Project. *AI Magazine*, 31(3), 59–79. https://doi.org/10.1609/aimag.v31i3.2303

Fillmore, C. J., & Baker, C. F. (2001). Frame semantics for text understanding. *Proceedings of WordNet and Other Lexical Resources Workshop*, 59–64. https://doi.org/10.1.1.469.9423

Fodor, O., & Werthner, H. (2005). Harmonise: A step toward an interoperable e-tourism marketplace. *International Journal of Electronic Commerce*, 9(2), 11–39.

Freitas, A., & Curry, E. (2014). Natural language queries over heterogeneous linked data graphs. *Proceedings of the 19th international conference on Intelligent User Interfaces - IUI '14*, 279–288. https://doi.org/10.1145/2557500.2557534

Frost, R. A., Donais, J., Mathews, E., Agboola, W., & Stewart, R. (2014). A demonstration of a natural language query interface to an event-based semantic web triplestore. *European Semantic Web Conference*, 343–348.

Gibbins, N. (2016). Resource Description Framework.

Green, B. F., Wolf, A. K., Chomsky, C., & Laughery, K. (1961). Baseball: an automatic question-answerer. *Proceeding IRE-AIEE-ACM '61 (Western) Papers presented at the May 9-11, 1961, western joint IRE-AIEE-ACM computer conference*, 219–224.

Grosz, B. J., Joshi, A. K., & Weinstein, S. (1983). Providing a unified account of definite noun phrases in discourse. *Proceedings of the 21st annual meeting on Association for Computational Linguistics*, 44–50.

Grosz, B. J., Weinstein, S., & Joshi, A. K. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational linguistics*, 21(2), 203–225.

Gupta, C., Jain, A., & Joshi, N. (2018). Fuzzy logic in natural language processing–a closer view. *Procedia computer science*, 132, 1375–1384.

Höffner, K., & Lehmann, J. (2014). Towards question answering on statistical linked data. *Proceedings of the 10th International Conference on Semantic Systems - SEM '14*, 61–64. https://doi.org/10.1145/2660517.2660521

Höffner, K., Walter, S., Marx, E., Usbeck, R., Lehmann, J., & Ngonga Ngomo, A. C. (2017). Survey on challenges of Question Answering in the Semantic Web. *Semantic Web*, 8(6), 895–920. https://doi.org/10.3233/SW-160247

Htay, H. H., & Murthy, K. N. (2008). Myanmar word segmentation using syllable level longest matching. In *Proceedings of the 6th Workshop on Asian Language Resources*.

Isbell, J., & Butler, M. H. (2007). Extracting and Re-using Structured Data from Wikis. Digital Media Systems Laboratory of Hewlett-Packard Development Company, Bristol, HPL-2007-182, 14th November.

Jitkrittum, W., Haruechaiyasak, C., & Theeramunkong, T. (2009). QAST : Question Answering System for Thai Wikipedia. *Proceedings of the 2009 Workshop on Knowledge and Reasoning for Answering Questions*, 11–14.

Jurafsky, D., & Martin, J. H. (2009). Speech and Language Processing: An Introduction to Natural Language Processing, *Computational Linguistics, and Speech Recognition*. Pearson Prentice Hall.

Jurafsky, Daniel and Martin, J. H. (2000). Speech and Language Processing: An Introduction to Natural Language Processing, *Computational Linguistics, and Speech Recognition*. MIT Press.

Kaisser, M., & Webber, B. (2007). Question answering based on semantic roles. *Proceedings of the workshop on deep linguistic processing*, 41–48. https://doi.org/10.3115/1608912.1608920

Kasami, T. (1965). An Efficient Recognition and Syntax-Analysis Algorithm for Context-Free Languages (tech. rep.). Air Force Cambridge Research Laboratory. Bedford, Massachusetts.

Kate, R. J., Wong, Y. W., & Mooney, R. J. (2005). Learning to transform natural to formal languages. *Proceedings of the 20th national conference on Artificial intelligence*-Volume 3, 1062–1068.

Katz, B. (1997). Annotating the World Wide Web Using Natural Language. *Proceedings of the 5th RIAO Conference on Computer Assisted Information Searching on the Internet (RIAO '97)*, 136–159.

Katz, B., Borchardt, G., & Felshin, S. (2006). Natural Language Annotations for Question Answering. *Proceedings of 19th International FLAIRS Conference*, 303–306.

Katz, B., & Lin, J. (2002). Annotating the Semantic Web Using Natural Language. *Proceedings of the 2nd workshop on NLP and XML*, 17, 1–8.

Katz, B., Lin, J., & Quan, D. (2002). Natural Language Annotations for the Semantic Web. *On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE*, 1317–1331.

Kawtrakul, A., & Thumkanon, C. (1997). A statistical approach to thai morphological analyzer. *Proceedings of of the 5th Workshop on Very Large Corpora*.

Ketui, N., Theeramunkong, T., & Onsuwan, C. (2012). A rule-based method for thai elementary discourse unit segmentation (ted-seg). *2012 Seventh International Conference on Knowledge, Information and Creativity Support Systems*, 195–202.

Knuth, D. E. (1965). On the translation of languages from left to right. *Information and Control*, 8(6), 607–639.

Kohonen, T. (2001). Self-organizing maps (S. Berlin, Ed.; Vol. 30).

Kongthon, A., Kongyoung, S., Haruechaiyasak, C., & Palingoon, P. (2011). A semantic based question answering system for Thailand tourism information. *The KRAQ11 Workshop: Knowledge and Reasoning for Answering Questions*.

Kongwan, A., & Kawtrakul, A. (2005). Know-what : A Development of Object-Property Extraction from Thai Texts and Query System. *The Sixth Symposium on Natural Language Processing*, 157–162.

Kongyoung, S., Rugchatjaroen, A., & Kosawat, K. (2015). TLex+: A hybrid method using conditional random fields and dictionaries for Thai word segmentation. *International Conference on Knowledge, Information, and Creativity Support Systems*, 112–125.

Konrad, H., Lehmann, J., & Usbeck, R. (2016). CubeQA — Question Answering on RDF Data Cubes. *International Semantic Web Conference (ISWC)*, 325–340. https://doi.org/10.1007/978-3-319-46523-4_20

Kruengkrai, C., Sornlertlamvanich, V., & Isahara, H. (2006). A Conditional Random Field Framework for Thai Morphological Analysis. *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC-2006)*, 2419–2424.

Kumar, P., & Goel, K. (2016). Universal Networking Language : A framework for emerging NLP applications. *Information Processing (IICIP)*, 1–6.

Kwok, C., Etzioni, O., & Weld, D. S. (2001). Scaling question answering to the web. *ACM Transactions on Information Systems*, 19(3), 242–262. https://doi.org/10.1145/502115.502117

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of the eighteenth international conference on machine learning, ICML*, 282–289.

Lange, M., & Leiß, H. (2009). To CNF or not to CNF - An Efficient Yet Presentable Version of the CYK Algorithm. *Informatica Didactica*, 8.

Lassila, O., & Swick, R. R. (1999). Resource description framework (RDF) model and syntax specification. W3C Recommendation, World Wide Web Consortium.

Lebedeva, O., & Zaitseva, L. (2014). Question Answering Systems in Education and their Classifications. *Joint International Conference on Engineering Education & International Conference on Information Technology*, 359–366.

Lertcheva, N. (2010). Thai Named Entity Recognition: A Study of Product Names in Economic News. *Master of Arts Thesis. Bangkok: Chulalongkorn University*.

Lestari, D. P., & Nugraha, R. R. (2017). A Spoken-Based Question Answering System for Train Route Service using the Frame-Based Approach. *Electrical Engineering and Informatics (ICEEI)*, 1–6.

Li, X., & Roth, D. (2002). Learning question classifiers. *Proceedings of the 19th international conference on Computational linguistics*-Volume 1, 1–7. https://doi.org/10.3115/1072228.1072378

Li, X., & Roth, D. (2006). Learning question classifiers: the role of semantic information. *Natural Language Engineering*, 12(03), 229. https://doi.org/10.1017/S1351324905003955



Lin, J., Fernandes, A., Katz, B., Marton, G., & Tellex, S. (2002). Extracting Answers from the Web Using Knowledge Annotation and Knowledge Mining Techniques. *Proceedings of the Eleventh Text Retrieval Conference*.

Lowe, J. B., Baker, C. F., & Fillmore, C. J. (1997). A frame-semantic approach to semantic annotation. *Proceedings 1997 Siglex Workshop/ANLP97*, 18–24.

McDonald, R., Pereira, F., Ribarov, K., & Hajic, J. (2005). Non-projective dependency parsing using spanning tree algorithms. *Proceedings of the conference on Human Language*

*Technology and Empirical Methods in Natural Language Processing*, 523–530. https://doi.org/10.3115/1220575.1220641

Mesgari, M., Okoli, C., Mehdi, M., Nielsen, F. Å., & Lanamäki, A. (2015). "The Sum of All Human Knowledge": A Systematic Review of Scholarly Research on the Content of Wikipedia. *Journal of the Association for Information Science and Technology*, 66(2), 219–245. https://doi.org/10.1002/asi.23172

Miller, S., Bobrow, R., Ingria, R., & Schwartz, R. (1993). Hidden understanding models of natural language. *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, 25–32.

Miller, S., Fox, H., Ramshaw, L., & Weischedel, R. (2000). A novel use of statistical parsing to extract information from text. *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, 226–233. http://portal.acm.org/citation.cfm?id=974335

Miller, S., Stallard, D., Bobrow, R., & Schwartz, R. (1996). A Fully Statistical Approach to Natural Language Interfaces. *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL-96)*, 55–61. https://doi.org/10.3115/981863.981871

Mishra, A., & Jain, S. K. (2016). A survey on question answering systems with classification. *Journal of King Saud University - Computer and Information Sciences*, 28(3), 345–361. https://doi.org/10.1016/j.jksuci.2014.10.007

Mridha, M. F., Saha, A. K., & Das, J. K. (2014). Solving Semantic Problem of Phrases in NLP Using Universal Networking Language (UNL). *International Journal of Advanced Computer Science and Applications (IJACSA), Special Issue on Natural Language Processing (NLP)*.

Nararatwong, R., Kertkeidkachorn, N., Cooharojananone, N., & Okada, H. (2018). Improving Thai word and sentence segmentation using linguistic knowledge. *IEICE Transactions on Information and Systems*, 101(12), 3218–3225. https://doi.org/10.1587/transinf.2018EDP7016

Narayanan, S., & Harabagiu, S. (2004). Question answering based on semantic structures. *Proceedings of the 20th international conference on Computational Linguistics*, 693.

Nomoto, H., Okano, K., Wittayapanyanon, S., & Nomura, J. (2019). Interpersonal meaning annotation for asian language corpora: The case of TUFS asian language parallel corpus (TALPCo). In *Proceedings of the Twenty-Fifth Annual Meeting of the Association for Natural Language Processing* (pp. 846-849).

Noraset, T., Lowphansirikul, L., & Tuarob, S. (2021). Wabiqa: A wikipedia-based thai question-answering system. *Information Processing & Management*, 58(1), 102431.

Novák, V. (1992). Fuzzy sets in natural language processing. In *An introduction to fuzzy logic applications in intelligent systems* (pp. 185–200). Springer.

Novák, V. (2017a). Fuzzy logic in natural language processing. *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 1–6.

Novák, V. (2017b). Fuzzy logic in natural language processing. *IEEE International Conference on Fuzzy Systems*. https://doi.org/10.1109/FUZZ-IEEE.2017.8015405

Pa, W. P., Thu, Y. K., Finch, A., & Sumita, E. (2016). Word boundary identification for Myanmar text using conditional random fields. In *Genetic and Evolutionary Computing: Proceedings of the Ninth International Conference on Genetic and Evolutionary Computing, August 26-28, 2015, Yangon, Myanmar-Volume II 9* (pp. 447-456). Springer International Publishing.

Pathanasin, S. (2018). Coherence in thai students' essays: An analysis using centering theory. *Manusya: Journal of Humanities*, 21(2), 112–130.

Peng, F., Feng, F., & McCallum, A. (2004). Chinese segmentation and new word detection using conditional random fields. *Proceedings of the 20th international conference on Computational Linguistics*, 562.

Pereira, A., Trifan, A., Lopes, R. P., & Oliveira, J. L. (2022). Systematic review of question answering over knowledge bases. *IET Software*, 16(1), 1–13.

Pingali, K., & Bilardi, G. (2015). A graphical model for context-free grammar parsing. *International Conference on Compiler Construction*, 3–27. https://doi.org/10.1007/978-3-662-46663-6_1

Poowarawan, Y. (1986). Dictionary-based thai syllable separation. *Proceedings of the Ninth Electronics Engineering Conference*, 409–418.

Pota, M., Esposito, M., & De Pietro, G. (2017). Learning to rank answers to closed-domain questions by using fuzzy logic. *IEEE International Conference on Fuzzy Systems*, (1).

Pranter, K., Ding, Y., Luger, M., Yan, Z., & Herzog, C. (2007). Tourism Ontology and Semantic Management System: State of the Art Analysis. In *Proceedings of the IADIS International Conference WWW/Internet*, 111–115.

Prud'Hommeaux, E., & Seaborne, A. (2008). SPARQL query language for RDF. W3C recommendation 15.

Rani, M., Muyeba, M. K., & Vyas, O. P. (2014). A Hybrid Approach using Ontology Similarity and Fuzzy Logic for Semantic Question Answering. Advanced Computing, *Networking and Informatics*, 1, 601–609.

Ranjan, P., & Balabantaray, R. C. (2016). Question Answering System for Factoid Based Question. *Contemporary Computing and Informatics* (IC3I), 221–224.

Ratnaparkhi, A. (1999). Learning to parse natural language with maximum entropy models. *Machine learning*, 34(1-3), 151–175.

Riaz, F., Anwar, M. W., & Muqades, H. (2020, February). Maximum entropy based urdu named entity recognition. In *2020 International Conference on Engineering and Emerging Technologies (ICEET)* (pp. 1-5). IEEE.

Ripon, S., Barua, A., & Uddin, M. S. (2014). Analysis Tool for UNL-Based Knowledge Representation. arXiv.

Robertson, S., Zaragoza, H., et al. (2009). The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4), 333–389.

Satayamas, V., Thumkanon, C., & Kawtrakul, A. (2005). Bootstrap cleaning and quality control for Thai tree bank construction. *The 9th National Computer Science and Engineering Conference*.

Sha, F., & Pereira, F. (2003). Shallow parsing with conditional random fields. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*-Volume 1, 134–141. https://doi.org/10.3115/1073445.1073473

Shen, D., & Lapata, M. (2007). Using Semantic Roles to Improve Question Answering. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 12–21.

Sinthupoun, S., & Sornil, O. (2010). Thai rhetorical structure analysis. *Journal of Computer Science and Information Security*, 7(1).

Sornlertlamvanich, V. (1993). Word segmentation for Thai in machine translation system. *Machine Translation*, National Electronics and Computer Technology Center, Bangkok, 50–56.

Sornlertlamvanich, V., Takahashi, N., & Isahara, H. (1999). Building a Thai part-of-speech tagged corpus (ORCHID). *Journal of the Acoustical Society of Japan (E)*, 20(3), 189-198.

Sowa, J. F. (2014). Principles of semantic networks: Explorations in the representation of knowledge. Morgan Kaufmann.

Srithirath, A., & Seresangtakul, P. (2013, May). A hybrid approach to lao word segmentation using longest syllable level matching with named entities recognition. In *2013 10th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology* (pp. 1-5). IEEE.

Steedman, M. (1996). Surface Structure and Interpretation. MIT press.

Steedman, M. (2000). The Syntactic Process. MIT press.

Sudprasert, S., & Kawtrakul, A. (2003). Thai word segmentation based on global and local unsupervised learning. *Proceedings of National Computer Science and Engineering Conference*.

Sun, H., Ma, H., Yih, W.-t., Tsai, C.-T., Liu, J., & Chang, M.-W. (2015). Open Domain Question Answering via Semantic Enrichment. *Proceedings of the 24th International Conference on World Wide Web - WWW '15*, 1045–1055. https://doi.org/10.1145/2736277.2741651

Sutantayawalee, V., & Supnithi, T. (2016). A Comparative Study of Approaches of Incorporating Out-of-Vocabulary Word into Existing Word Segmentation Model. *Proceedings of the 11th International Symposium on Natural Language Processing (SNLP2016)*.

Sutheebanjard, P., & Premchaiswadi, W. (2009). Thai personal named entity extraction without using word segmentation or POS tagging. *2009 Eighth International Symposium on Natural Language Processing*, 221–226. https://doi.org/10.1109/SNLP.2009.5340914

Tapsai, C., Meesad, P., & Unger, H. (2019). An overview on the development of thai natural language processing. *Information Technology Journal*, 15(2), 45–52.

Tettamanzi, A. G. B. (2003). A Fuzzy Frame-Based Knowledge Representation Formalism. *International Workshop on Fuzzy Logic and Applications*, 55–62. https://doi.org/10.1007/10983652_8

Theeramunkong, T., Boriboon, M., Haruechaiyasak, C., Kittiphattanabawon, N., Kosawat, K., Onsuwan, C., ... & Tongtep, N. (2010). Thai-nest: A framework for thai named entity tagging specification and tools. In *Language Windowing through Corpora* (pp. 895-908). Servizo de Publicacións.

Tongchim, S., Altmeyer, R., Sornlertlamvanich, V., & Isahara, H. (2008). A Dependency Parser for Thai. *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, (October 2016).

Uchida, H., Zhu, M., & Della Senta, T. (2005). Universal Networking Language. UNDL foundation. Urrutià, A. T., López, M. D. J., & Blache, P. (2017). Fuzziness and variability in natural language processing. *IEEE International Conference on Fuzzy Systems*. https://doi.org/10.1109/FUZZ-IEEE.2017.8015692

Usbeck, R., Ngomo, A. C. N., Bühmann, L., & Unger, C. (2015). HAWK – hybrid question answering using linked data. *European Semantic Web Conference*, 353–368. https://doi.org/10.1007/978-3-319-18818-8_22

Vanthanavong, S., & Haruechaiyasak, C. (2011). LaoWS: Lao word segmentation based on conditional random fields. In *Conference on Human Language Technology for Development* (pp. 21-26).

Voorhees, E. M. (2001). The TREC question answering track. *Natural Language Engineering*, 7(04), 361–378.

Voorhees, E. M. (2003). Overview of the TREC 2003 Question Answering Track. *TREC*, 2003, 54–68.

Walker, M., Iida, M., & Cote, S. (1990). Centering in Japanese Discourse. *Proceedings of the 13th COLING*, 1–6.

Walker, M., Iida, M., & Cote, S. (1996). Japanese Discourse and the Process of Centering. *Computational Linguistics*, 20(2), 193–232. http://arxiv.org/abs/cmp-lg/9609006

Wang, H., Bansal, M., Gimpel, K., & McAllester, D. (2015). Machine Comprehension with Syntax , Frames , and Semantics. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, 700–706.

Wikipedia. (2005). Universal Networking Language. Retrieved May 10, 2017, from https://en.wikipedia.org/wiki/Universal%7B%5C_%7DNetworking%7B%5C_%7DLanguage

Wikipedia. (2007). CYK algorithm. Retrieved August 1, 2017, from https://en.wikipedia.org/wiki/CYK%7B%5C_%7Dalgorithm

Woods, W. A. (1973). Progress in natural language understanding: an application to lunar geology. *Proceedings of the National Computer Conference and Exposition on AFIPS '73*, 441–450. https://doi.org/10.1145/1499586.1499695

Yamada, H., & Matsumoto, Y. (2003). Statistical dependency analysis with support vector machines. *Proceedings of IWPT*, 195–206.

Yang, W., Xie, Y., Lin, A., Li, X., Tan, L., Xiong, K., Li, M., & Lin, J. (2019). End-to-end open-domain question answering with bertserini. arXiv preprint arXiv:1902.01718.

Yao, X. (2015). Lean Question Answering over Freebase from Scratch. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, 66–70. https://doi.org/10.3115/v1/N15-3014

Yih, W.-t., Chang, M.-W., He, X., & Gao, J. (2015). Semantic Parsing via Staged Query Graph Generation: Question Answering with Knowledge Base. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (Volume 1: Long Papers), 1321–1331. https://doi.org/10.3115/v1/P15-1128

Younger, D. H. (1967). Recognition and parsing of context-free languages in time n3. *Information and control*, 10(2), 189–208. Zadeh, L. A. (1965). Fuzzy sets. Information and Control, 8(3), 338–353. https://doi.org/10.1016/S0019-9958(65)90241-X

Zadeh, L. A. (1988). Fuzzy logic. *Computer*, 21(4), 83–93. https://doi.org/10.1109/2.53

Zadeh, L. A. (1989). Knowledge representation in fuzzy logic. *IEEE Transactions on Knowledge and Data Engineering*, 1(1), 89–100. https://doi.org/10.1109/69.43406

Zadeh, L. A. (1999). Fuzzy sets as a basis for a theory of possibility. *Fuzzy sets and systems*, 100, 9–34.

Zettlemoyer, L. S., & Collins, M. (2005). Learning to Map Sentences to Logical Form : Structured Classification with Probabilistic Categorial Grammars. *21st Conference on Uncertainty in Artificial Intelligence*, 658–666. http://arxiv.org/abs/1207.1420

Zhang, D., & Lee, W. (2003). Question classification using support vector machines. *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, 26–32. https://doi.org/10.1145/860435.860443

Zhou, G., Zhou, Y., He, T., & Wu, W. (2016). Learning semantic representation with neural networks for community question answering retrieval. *Knowledge-Based Systems*, 93, 75–83. https://doi.org/10.1016/j.knosys.2015.11.002

Zhu, F., Lei, W., Wang, C., Zheng, J., Poria, S., & Chua, T.-S. (2021). Retrieving and reading: A comprehensive survey on open-domain question answering. arXiv preprint arXiv:2101.00774.

# Appendix A

# Thai Part of Speech

## A.1 Noun and Classifier

NCM : Common Noun

For example : นก ลิง สัตว์ สำนวน เศรษฐกิจ

NPN : Proper Noun

For example : เคนยา โลก อังกฤษ ไทย อยุธยา

NPP : Noun for Part and Position

For example : ปลาย โคน ใต้ บน ต้น ทั้ง

NTT : Title Noun

For example : นาย นาง นางสาว

NCA : Noun and Classifier for Attributed, Kind and Group

For example : ขนาด สี อายุ ราคา พวก เหล่า กลุ่ม ประเภท ชนิด จำพวก

NCT : Noun and Classifier for Time and Moment

For example : ชวง สมัย ยุค วัน ที ครั้ง นาที

CLS : Classifier for Measurement, Object and Concrete Thing

For example : ข้าง ตัว อัน กิโลกรัม บาท

NUM : Number

For example : 1 20 299

NNA : Noun for Amount

For example : สิบ ร้อย พัน หมื่น แสน ล้าน

**A.2 Pronoun**

PRO : Pronoun

For example : คุณ มัน อัน ฉัน เธอ

PRL : Relative Pronoun

For example : ที่ ซึ่ง อัน ผู้



**A.3 Verb**

VRB : Transitive Verb

For example : กิน หิว มอง รัก

VRI : Intransitive Verb

For example : ยิ้ม วิ่ง บิน

VAT : Attribute Verb

For example : อ้วน ใหญ่ หนัก ยาว สูง

VAX : Auxiliary Verb

For example : จะ จึง จึงจะ ก็ ก็จะ ยัง ยังจะ คง คงจะ ยังคง จะยังคง มัก มักจะ เพิ่ง

เพิงจะ ได้ ก็ได้ จะได้ อาจ อาจจะ ควร ควรจะ เริ่ม เริ่มจะ เคย ต้อง จะต้อง ค่อนข้าง

ค่อนข้างจะ กำลัง กำลังจะ ที่จะ น่า น่าจะ สุด สุดจะ แสน แสนจะ กลับ กลับจะ สามารถ

จำเป็น

VAV : Auxiliary Verb for Passive Voice

For example : ถูก เป็นที่ โดน

VPR : Preverb

For example : ชอบ กล้า ช่วย ห้าม โปรด พร้อม ใช้ นิยม ออก สังเกตุ ให้ ต้องการ แสร้ง

พยายาม มา เริ่มต้น ไป

VPT : Postverb for Direct Verb

For example : ไป ไว้ มา ได้ ขึ้น ลง เข้า ออก

VPO : Postverb for Indirect Object

For example : ไป ไว้ มา ได้ เป็น พร้อม ให้ เหมือน แทน คล้าย จาก ถึง แทบ โดน ว่า

VPA : Postverb for Connected with Attribute Verb or Adverb

For example : ไป ไว้ มา ได้ ให้ เป็น แทบ

VSA : Stated Attribute Verb

For example : ถูก ผิด

## A.4 Adverb

ADV : Adverb

For example : พอสมควร เร็ว มาก ดี นาน อีกประการหนึ่ง แล้ว เกินไป เพียงนี้ เท่านี้

ขนาดนี้ ถึงเพียงนี้ เอง อีก ก็ตาม กว่า ที่สุด ไป มาก

## A.5 Determiner

DSO : Determiner for Specific Object

For example : นี้ นั้น ทั้งหมด เชนนี้ อย่างนี้ อย่างหนึ่ง ส่วนใหญ่ ดังกล่าว อื่น เอง

DHN : Determiner at Head Noun

For example : บรรดา อดีต รอง นาน อีก สารพัด

DQE : Determiner for Quantity Expression

For example : ประมาณ บาง ทุก หลาย แต่ละ ตลอด หลายๆ อีก

DRF : Reference Determiner

For example : เช่น ตัวอย่างเช่น ได้แก่

DRE : Reference Determiner Ending

For example : เป็นต้น

## A.6 Preposition

PRP : Preposition

For example : ของ ที่ ใน โดย ต่อ ระหว่าง ต่อจาก สู่ เชง ทั่ว กว่า เท่ากับ เพียง ทั้ง แต่ ิ

## A.7 Conjunction

CON : Conjuction

For example : และ หรือ แล้ว พร้อม อม ปน ผสม แซม

SUB : Subordinate

For example : ซึ่ง สำหรับ แต่ แต่ถึงอย่างไร ถ้า ให้ นอกจากนั้น จากนั้น นอกจากนี้

นอกจากนี้แล้ว โดย โดยที่ โดยเฉพาะ โดยเฉพาะอย่างยิง หาก จน จนกว่า เมื่อ เมื่อเวลา

เพื่อ ก็เพื่อ เนื่องจาก ทั้งนี้ อย่างไรก็ตาม ดังนั้น ทำให้ จะทำให้ ที่ทำให้ ต่อมา หลังจาก

หลังจากที่ ขณะ ขณะที่ ขณะเดียวกัน ในขณะที่ ก็ย่อมที่จะ ย่อมที่จะ ส่วน แล้ว แล้วก็ยัง

อีกทั้งยัง ยกเว้น ทั้งนี้ เพราะฉะนั้น เหมือน แม้ แม้ว่า ถึงแม้ว่า กระทั่ง ตาม รวมทั้ง

ก็ต่อเมื่อ เว้นเสียแต่ว่า

## A.8 Negation

NEG : Negation

For example : ไม่ มิ ผิด ห้าม

### A.9 Prefix and Suffix

FVN : Prefix to transform Noun and Verb to be Noun

For example : การ

FDN : Prefix to transform Attribute Verb, Verb and Adverb to be Noun

For example : ความ

FNH : Prefix to transfrom Noun to be Humanize Noun

For example : ชาว นัก นักการ

FVH : Prefix to transform Verb to be Humanize Noun

For example : นัก

FAV : Prefix for Adverb Modifier to Attribute Verb, Verb and Adverb

For example : อย่าง อัน ที่ โดย ได้

FXO : Prefix for Ordinal Number

For example : ที่ (วันที่ ตัวที่ อันที่)

COC : Connector Suffix for Classifier

For example : ละ (ข้างละ ตัวละ)

**A.10 Other**

EAF : Affirmative Sentence Ending

For example : จ๊ะ ครับ เถอะ นะ แหละ

EIT : Interrogative Sentence Ending

For example : หรือ เหรอ ไหม มั้ย

INT : Interjection

For example : โอ้ย โอ้ เออ เอ๋ อ๋อ

NBM : Number Modifier

For example : ครึ่ง เศษ ถ้วน เดียว กว่า

NBO : Ordinal Number Word

For example : แรก สุดท้าย ต่อไป

UNK : Unknown

For example :

CHR : Single Charactor

For example : ป พ ศ

QOB : Question for Object

For example : ใคร อะไร ไหน ใด

QAT : Question for Amount ot Time

For example : เท่าไหร่ เท่าใด กี่ เมื่อไหร่ เมื่อใด

QAR : Question for Attribute

For example : ใค ไหน อะไร

### A.11 Shallow Parser Tags

(*) Head Noun Pattern

(&) Verbal Noun Pattern

(+) Adjective Pattern

(#) Transitive Verb Pattern

(%) Intransitive Verb Pattern

($) Adverb Pattern and Postverb that separate from Verb Pattern

(!) Preposition Pattern

(@) Amount Pattern

(=) Time Pattern

(?) Determiner Pattern

(-) Classifier Pattern

# Appendix B

## Question Answer Pair Validation Form

**Question Answer Pair Validation Form**

**Validator Information**

Name……

Education……

Position……

Work Place ……

……

……

……

……



Signed

**Question Answer Pair Validation**

Please / if agree and put × if disagree on agreement column

| No. | Question | Answer | Correction | Agreement |
|---|---|---|---|---|
| 1 | ชวน หลีกภัยเป็นใคร | นักการเมืองชาวไทย | / | |
| | | บุตรคนที่ 3 ในจำนวน 9 คน ของ นิยม กับ ถ้วน หลีกภัย | / | |
| | | คนรูปร่างเล็ก | / | |
| | | สมาชิกสภาผู้แทนราษฎร จังหวัดตรัง | / | |
| | | หัวหน้าพรรคประชาธิปัตย์ | / | |
| | | พลเรือนคนสอง | × | |
| | | คอมมิวนิสต์ | × | |
| 2 | ใครเป็นนักการเมืองชาวไทย | ชวน หลีกภัย | / | |
| 3 | ชวน หลีกภัยเกิดเมื่อไหร่ | วันที่ 28 กรกฎาคม พ.ศ. 2481 | / | |
| 4 | ชวนเกิดที่ไหน | ตำบลท้ายพรุ | / | |
| 5 | ใครเกิดเมื่อวันที่ 28 กรกฎาคม พ.ศ. 2481 | ชวน หลีกภัย | / | |
| 6 | ชวน หลีกภัยดำรงตำแหน่งใด | ตำแหน่งประธานรัฐสภา | / | |
| | | ตำแหน่งนายกรัฐมนตรี | / | |
| | | ตำแหน่งรัฐมนตรีว่าการหลายกระทรวง | / | |

| | | | | |
|---|---|---|---|---|
| | | ตำแหน่งรองนายกรัฐมนตรี | / | |
| | | ตำแหน่ง | × | |
| | | ตำแหน่งหัวหน้าพรรค 3 สมัย | / | |
| | | ตำแหน่งประธานสภาที่ปรึกษาพรรคประชาธิปัตย์ | / | |
| | | ตำแหน่งประธานสภาผู้แทนราษฎร | / | |
| 7 | ชวน หลีกภัยสังกัดพรรคใด | พรรคประชาธิปัตย์ | / | |
| 8 | ธนาคารไทยได้เปลี่ยนชื่อเป็นอะไร | ธนาคารมณฑล | / | |
| 9 | บริษัท ข้าวไทย จำกัดมีสถานะเป็นอะไร | รัฐวิสาหกิจ | / | |
| 10 | กลุ่มซอยราชครูแต่งตั้งใครขึ้นเป็นประธานคณะกรรมการ | พล.ต.ต. ละม้าย อุทยานานนท์ หนึ่งภายในกลุ่มซอยราชครู | / | |
| 11 | ธนาคารเกษตรได้รับการก่อตั้งโดยใคร | กลุ่มข้าราชการ | / | |
| 12 | ธนาคารเกษตรมีทุนจดทะเบียนในชั้นแรกเท่าไหร่ | - | × | |
| 13 | ธนาคารเกษตรได้กลายเป็นฐานเศรษฐกิจของใคร | กลุ่มซอยราชครู | / | |
| 14 | ธนาคารเกษตร จำกัดก่อตั้งโดยใคร | สุริยน ไรวา | / | |
| 15 | ธนาคารกรุงไทยใช้ตราสัญลักษณ์เป็นอะไร | ภาพนกวายุภักษ์ | / | |

| | | | | |
|---|---|---|---|---|
| 16 | กระทรวงการคลังเป็นอะไร | ผู้ | × | |
| | | หน่วยงานรัฐ | / | |
| 17 | ธนาคารเปลี่ยนสีพื้นหลังของนก วายุภักษ์จากสีขาวเป็นสีอะไร | ฟ้าแทน | / | |
| 18 | อะไรจัดอยู่ในประเภทสัตว์มี กระดูกสันหลัง | นกกระจอกเทศ | / | |
| 19 | นกกระจอกเทศจัดอยู่ในประเภท ใด | ประเภทสัตว์มีกระดูกสันหลัง | / | |
| 20 | อะไรมีถิ่นกำเนิดในทวีปแอฟริกา | นกกระจอกเทศ | / | |
| 21 | นกกระจอกเทศมีถิ่นกำเนิดใน ทวีปใด | ทวีปแอฟริกา | / | |
| 22 | เท้าของม้ามีกี่นิ้ว | - | × | |
| 23 | เท้าของม้าเรียกว่าอะไร | กีบเท้าม้า | / | |
| 24 | นกกระจอกเทศมีอายุยืนได้ถึงกี่ปี | ยืน 65 - 75 ปี | / | |
| 25 | อะไรมีอายุยืนได้ถึง 70 ปี | นกกระจอกเทศ | / | |
| 26 | อะไรมีอายุยืนได้ถึง 60 ปี | นกกระจอกเทศ | / | |
| 27 | ไข่ของอะไรเป็นไข่นกที่ใหญ่ที่สุด ในโลก | นกกระจอกเทศ | / | |
| 28 | นกกระจอกเทศกินอะไร | พืช | / | |
| | | สิ่งของแปลกปลอม | × | |
| 29 | นกกระจอกเทศถูกนำเข้ามา เมื่อไหร่ | ครั้งแรก | × | |
| 30 | อะไรเป็นสัตว์กินพืช | นกกระจอกเทศ | / | |

| | | | | |
|---|---|---|---|---|
| | | อาหารหลัก | × | |
| | | เหยื่อของเสือภายในสกุลแพนเทอรา | × | |
| 31 | นกกระจอกเทศเป็นอะไร | นก | / | |
| | | สัตว์กินพืช | / | |
| | | นกกระจอกเทศ | × | |
| | | นกกระจอกเทศป่า | × | |
| | | สัตว์ | / | |
| 32 | อะไรอาศัยอยู่ในมอรอคโค | นกกระจอกเทศพันธุ์คอดำ | / | |
| 33 | อะไรมีสีเทา | ลักษณะผิวหนังนกกระจอกเทศ | / | |
| | | นกกระจอกเทศตัวเมียโตเต็มที่ | / | |
| | | ขนของตัวเมีย | / | |
| 34 | อะไรมีสีดำ | - | × | |
| 35 | อะไรมีสีเทาดำ | ลักษณะผิวหนังนกกระจอกเทศ | / | |
| | | นกกระจอกเทศตัวเมียโตเต็มที่ | / | |
| | | ขนของตัวเมีย | / | |
| 36 | อะไรมาจากแอฟริกาตะวันออก | นกกระจอกเทศพันธุ์คอแดง | / | |
| 37 | นกกระจอกเทศพันธุ์อะไรเป็นนกกระจอกเทศป่า | พันธุ์คอแดง | / | |
| | | พันธุ์คอน้ำเงิน | / | |
| 38 | นกกระจอกเทศพันธุ์คอน้ำเงินมีถิ่นกำเนิดอยู่ทางใด | - | × | |

| | | | | |
|---|---|---|---|---|
| 39 | อะไรที่โตเต็มที่จะมีสีฟ้า | นกกระจอกเทศตัวเมียโตเต็มที่ | / | |
| 40 | อะไรที่โตเต็มที่จะมีสีเทา | นกกระจอกเทศตัวเมียโตเต็มที่ | / | |
| 41 | อะไรที่โตเต็มที่จะมีสีฟ้าอมเทา | นกกระจอกเทศตัวเมียโตเต็มที่ | / | |
| 42 | แมวเป็นอะไร | สัตว์เลี้ยงลูกด้วยนม | / | |
| | | แมว | × | |
| | | สัตว์กินเนื้อ | / | |
| | | สัตว์ | / | |
| | | สัตว์ภายใน กลุ่มเสือ และ แมว ขนาดเล็กที่สุด | / | |
| | | เสือ | / | |
| 43 | อะไรเป็นสัตว์เลี้ยงลูกด้วยนม | แมว | / | |
| | | เสือ | / | |
| | | สัตว์ภายในกลุ่มเสือ | / | |
| | | สัตว์ | × | |
| 44 | เสือสืบสายเลือดมาจากอะไร | แมวป่า | / | |
| 45 | อะไรสืบสายเลือดมาจากแมวป่า | เสือ | / | |
| 46 | อะไรเป็นสัตว์กินเนื้อ | แมว | / | |
| | | เสือ | / | |
| 47 | แมวบ้านสืบเชื้อสายมาจากอะไร | แมวป่าอาฟริกา | / | |
| 48 | อะไรสืบเชื้อสายมาจากแมวป่าอาฟริกา | แมวบ้าน | / | |
| 49 | ลูกแมวแรกเกิดหนักเท่าไหร่ | แรกเกิดประมาณ 85 - 110 กรัม | / | |

| | | | | |
|---|---|---|---|---|
| 50 | อะไรหนักประมาณ 85 กรัม | บุตรแมวแรกเกิดประมาณ 85 - 110 กรัม | / | |
| 51 | อะไรหนักประมาณ 80 กรัม | บุตรแมวแรกเกิดประมาณ 85 - 110 กรัม | / | |
| 52 | ลูกแมวหย่านมเมื่ออายุเท่าไหร่ | 8 - 10 สัปดาห์ | / | |
| 53 | เสือหนักเท่าไหร่ | หนักประมาณ 180 - 245 กิโลกรัม | / | |
| 54 | อะไรหนักประมาณ 190 กิโลกรัม | เสือหนักประมาณ 180 - 245 กิโลกรัม | / | |
| 55 | อะไรหนักประมาณ 150 กิโลกรัม | เสือหนักประมาณ 180 - 245 กิโลกรัม | / | |

# Appendix C

## List of Expert Validator

1. Name : Parinyapon Muangrak

   Position : Senior Professional Level Teachers

   Education : Master Degree in Education Administration

   Workplace : Wichainchom School, Songkhla, Thailand

2. Name : Pawich Wattanpraphan

   Position : Senior Professional Level Teachers

   Education : Bachelor Degree in Educational Technology

   Workplace : Bannongwa School (Chamainukul), Nakhon Si Thammarat, Thailand

3. Name : Kamonthip Sasutham

   Position : Senior Professional Level Teachers

   Education : Bachelor Degree in Educational Technology

   Workplace : Bangkhuntiensuksa School, Bangkok, Thailand