WORD SENSE DISAMBIGUATION IN THAI USING DECISION LIST COLLOCATION

Miss Wipharuk Kanokrattananukul

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Arts in Linguistics

Department of Linguistics

Faculty of Arts

Chulalongkorn University

Academic Year 2001

ISBN 974-03-0553-9

การแก้ปัญหาความกำกวมของคำหลายความหมายในภาษาไทย โดยใช้รายการตัดสินของคำปรากฏร่วม

นางสาววิภารักษ์ กนกรัตนนุกูล

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาอักษรศาสตรมหาบัณฑิต
สาขาวิชาภาษาศาสตร์ ภาควิชาภาษาศาสตร์
คณะอักษรศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2544

ISBN 974-03-0553-9

ACKNOWLEDGEMENTS

Firstly, I would like to express my highest gratitude and indebtedness to Dr. Wirote Aroonmanakun, my thesis advisor who devoted his valuable time to teach, advise, and support me in all directions during my life as a linguistic student and during my work on the thesis. I would like to express my thanks to all committee members including, Assistant Professor Dr. Peansiri Vongvipanond, Assistant Professor Dr. Sudaporn Luksaneeyanawin, and Associate Professor Dr. Pranee Kulavanich, for their valuable comments and support needed to complete this thesis. I would like to express my thanks, again, to Assistant Professor Dr. Peansiri Vongvipanond for her kind help in creating sense labels for this thesis. I would like to give special thanks to Assistant Professor Dr. Krisadawan Hongladarom for her kind help in searching some books needed in this thesis when she went abroad. I would like to thank my brother and all my friends who always gave me enthusiasm and made me smile.

Finally, I am greatly indebted and grateful to my mother, Mrs. You-Kee Sae-Heng, for all of her great devotion to my life. Thank you.

Wipharuk Kanokrattananukul

วิภารักษ์ กนกรัตนนุกูล : การแก้ปัญหาความกำกวมของคำหลายความหมายในภาษาไทยโดยใช้รายการตัดสินของคำปรากฏร่วม. (WORD SENSE DISAMBIGUATION IN THAI USING DECISION LIST COLLOCATION) อ. ที่ปรึกษา : อ. ดร. วิโรจน์ อรุณมานะกุล, 332 หน้า. ISBN 974-03-0553-9.

วิทยานิพนธ์นี้มีวัตถุประสงค์เพื่อพัฒนาต้นแบบโปรแกรมแก้ปัญหาความกำกวมของคำหลายความหมายในภาษาไทย โดยใช้รายการตัดสินของคำปรากฏร่วม โดยศึกษาคำว่า หัว เป็นตัวแทนของคำนาม และคำว่า เก็บ เป็นตัวแทนของคำกริยา ผู้วิจัยได้ทำการวิเคราะห์ความหมายของ หัว และ เก็บ จากพจนานุกรมฉบับราชบัณฑิตยสถาน และ 1,800 ตัวอย่างของ หัว และ 1,800 ตัวอย่างของ เก็บ ซึ่งรวบรวมมาจากคลังข้อมูลของหนังสือพิมพ์กรุงเทพธุรกิจ จากการวิเคราะห์ ได้ความหมายของ หัว 20 ความหมาย และของ เก็บ 9 ความหมาย เพื่อที่จะค้นหาตำแหน่งและระยะทางของคำบ่งชี้ความหมาย ผู้วิจัยได้ใช้โปรแกรมตรวจหาคำปรากฏร่วมของ หัว และ เก็บ จากคลังข้อมูลฝึกสอน ในรูปแบบและขอบเขตต่างๆ จำนวน 20 แบบ ผลที่ได้คือรายการตัดสิน 20 รายการ เพื่อใช้ในการทดสอบแต่ละแบบ จากนั้นได้ทดสอบโปรแกรมโดยใช้รายการตัดสิน 20 รายการนี้ เพื่อหาระยะทางที่ดีที่สุดและตำแหน่งของคำบ่งชี้ความหมาย โปรแกรมทำการตัดสินโดยเลือกความหมายที่เกิดร่วมกับรูปคำที่มีค่าน้ำหนักการเกิดร่วมมากที่สุด ผลการทดลองปรากฏว่า ระยะทาง ± 2 เพียงพอในการแก้ปัญหาของทั้งสองคำ และคำที่อยู่ข้างขวาของทั้งสองคำเป็นตัวบ่งชี้ความหมาย ระยะทางที่ดีที่สุดในการแก้ปัญหาความกำกวมของ หัว คือ 1 คำทางขวาและทางซ้าย โดยมีค่าความแม่นยำเท่ากับ 87% ในขณะที่ระยะทางที่ดีที่สุดของ เก็บ คือ 2 คำทางขวา โดยมีค่าความแม่นยำเท่ากับ 80.25% ค่าความแม่นยำที่สูงนี้แสดงให้เห็นว่าโปรแกรมการแก้ปัญหาความกำกวมของคำหลายความหมายโดยใช้รายการตัดสินของคำปรากฏร่วมนี้ทำงานได้ดีในระดับหนึ่ง

ภาควิชา.....ภาษาศาสตร์..... ลายมือชื่อนิสิต.....
สาขาวิชา.....ภาษาศาสตร์.....ลายมือชื่ออาจารย์ที่ปรึกษา.....
ปีการศึกษา....2544.....ลายมือชื่ออาจารย์ที่ปรึกษาร่วม.....

4280184222 : MAJOR LINGUISTICS

KEY WORD: WORD SENSE DISAMBIGUATION / DECISION LIST ALGORITHM / COLLOCATION / SPAN

WIPHARUK KANOKRATTANANUKUL : WORD SENSE DISAMBIGUATION IN THAI USING DECISION LIST COLLOCATION. THESIS ADVISOR: WIROTE AROONMANAKUN, Ph.D., 332 pp. ISBN 974-03-0553-9.

This thesis aims to develop a prototype of word sense disambiguation program in Thai by using the decision list collocation algorithm. *หัว* /hua4/ is chosen as a representative of nouns and *เก็บ* /kep1/ as a representative of verbs. We analyzed the senses of *หัว* /hua4/ and *เก็บ* /kep1/ for preparing manually sense-tagged corpus based on the Thai dictionary of "The Royal Institute" and 1,800 samples of *หัว* /hua4/ and 1800 samples of *เก็บ* /kep1/ collected from a corpus of *Bangkok Business* newspaper. Twenty senses of *หัว* /hua4/ and nine senses of *เก็บ* /kep1/ were found. To test for the optimal span and the location of sense indicators, we trained the algorithm at twenty spans of collocation and obtained twenty decision lists and tested the algorithm with these decision lists. The algorithm made the decision by choosing the sense with the highest collocational weight. The result suggests that the span ± 2 is sufficient for the disambiguation of both words and the sense indicators of both words are mostly on the right side. The optimal span for disambiguating *หัว* /hua4/ is one-word-to-the-right-and-left with the precision rate of 87%, while *เก็บ* /kep1/ is two-words-to-the-right, with the precision rate of 80.25%. The high precision rate suggests that the decision list algorithm used in this study is applicable to the task.

Department……Linguistics……Student's signature.....

Field of study…….Linguistics……Advisor's signature.....

Academic year……2001…….Co-advisor's signature.....

Thesis Title	Word Sense Disambiguation in Thai Using Decision List Collocation
By	Miss Wipharuk Kanokrattananukul
Field of Study	Linguistics
Thesis Advisor	Wirote Aroonmanakun, Ph.D.

Accepted by the Faculty of Arts, Chulalongkorn University in Partial Fulfillment of the Requirements for the Master 's Degree

…… Dean of Faculty of Arts
(Assistant Professor M.R. Kalaya Tingsabadh, Ph.D.)

THESIS COMMITTEE

…….. Chairman
(Assistant Professor Peansiri Vongvipanond, Ph.D.)

……..……. Thesis Advisor
(Wirote Aroonmanakun, Ph.D.)

…….. Member
(Assistant Professor Sudaporn Luxsaneeyanawin, Ph.D.)

CONTENTS

LIST OF TABLES

Tables Pages

LIST OF TABLES

Tables Pages

LIST OF FIGURES

Figures Pages

Figures Pages

CHAPTER I

INTRODUCTION

1.1 Background

Natural language or human language possesses many types of ambiguities. One of them is **word sense ambiguity** — a phenomenon in which a word can have more than one meaning or sense. However, human usually neither notices this type of ambiguity nor has problems when facing with the ambiguity. He can determine the correct sense instantly almost every time the ambiguity occurs in a context. This language competence of human is supported by the work of Hirst (1987:84) in psycholinguistic research on word sense ambiguity resolution that "In general, people did not notice occurrences of word sense ambiguity, and seem to disambiguate without any conscious effort." Hirst (1987:85) stated further that "Many researches or even intuition suggested that ambiguities are always resolved by the end of the sentence, with a good guess being made if the information provided is insufficient." Thus, an important key to human's ability of ambiguity resolution, besides his language competence, is that he can make use of necessary information provided by a context (especially within a sentence) for disambiguation.

However, word sense ambiguity, like other types of ambiguities (categorial and structural ambiguities), is a crucial problem for **natural language processing (NLP) systems** as stated by Small, et al. (1988) that "ambiguity is a central problem in lexical semantics and its resolution determines progress in NLP in general." NLP systems for Thai language also face with the problem of word sense ambiguity. Thepkanjana (1993) explained that word sense ambiguity is one of the characteristics of Thai that causes a huge problem for Thai language processing systems. When an NLP system, for example, a Thai-English machine translation program translates a sentence like "เขา

เป็นเด็กที่หัวปานกลาง", suppose that the word *หัว* /hua4/ has 10 different senses, then the machine translation program will have at least 10 ways for translating this sentence. Thus, how to correctly select the right meaning of the word *หัว* /hua4/ is an important problem to be resolved for a machine translation system as well as for other NLP applications to a more or less degree as follows.

(1) Syntactic analysis: WSD is useful for syntactic analysis such as the analysis of grammatical gender, prepositional phrase attachment (Ravin, 1990). Ide and Véronis (1998) gave an example of WSD in helping the analysis of grammatical gender as in French, *livre* can mean "book" or "pound", knowing which sense is required can help tagging whether *livres* is a masculine noun or a feminine noun (in "book" sense, *livres* is masculine, in "pounds" sense, it is feminine).

(2) Text processing: WSD is necessary for text processing tasks because knowing the correct sense of a word can help editing the word correctly. For example, in spelling correction in French, the system can correctly change *comte* to *comté* when the correct sense of the word in a context is "county" (and not "count"). In accent restoration in textual medium where accents are missing, as in French *cote*, the system can put the correct accent to the word when it knows the correct sense of the word in a context (*côté* means "coast" and *coté* means "side") (Yarowsky, 1994). In changing case, as in *HE READ THE TIMES,* knowing that TIMES is the name of a newspaper, the system can correctly change this sentence to *He read the Times*.

(3) Speech processing: WSD is required for such speech processing tasks as correct phonetization of words in speech synthesis, word segmentation and homophone discrimination in speech recognition.

(4) Lexicography: Sense-annotated information from a corpus-based word sense disambiguation (WSD)[1] reduces the considerable overhead task for lexicographers in sorting large-scaled corpora according to word usage for the determination of different senses of words. (Kilgarriff; 1997)

(5) Information retrieval (IR): WSD is useful for IR in that they can supply the correct sense of the ambiguous words so that the IR system will return its finding which is relevant to a query. However, Kilgarriff (1997) argued that it is not clear whether WSD has the potential to significantly improve IR performance. There are two reasons. First, if WSD program is inaccurate, it will cause a huge trouble to the IR system more than the ambiguous word itself. Second, in longer queries, different words in the query will tend to be mutually disambiguating, so WSD is probably only relevant where the query is very short.

(6) Natural Language Understanding (NLU): According to Kilgarriff (1997), WSD is not much important to NLU because NLU applications such as message understanding and man-machine communication deal only with very specific text types, so only the sense of an ambiguous word that is relevant to this specific sublanguage will be likely to occur. However, there is a tendency that NLU systems will become more sophisticated, with richer domain models and less limitations in the varieties of text they can analyze. These will make WSD to be more relevant to NLU.

In conclusion, WSD is a necessary tool for most NLP tasks. It is very important for a machine translation system, very benefit to lexicography, required for some tasks of text and speech processing and helpful for syntactic analysis. For IR, WSD is necessary in some moderate degree because problems can be resolved by using longer

---

[1] See detailed explanation about a corpus-based WSD approach insection 2.5.2.

queries. NLU seems to require very little help from WSD because its applications are mostly domain specific. However, with the recent trend of NLU applications towards more general and unrestricted domain, WSD becomes more important.

1.2 Previous Researches on Word Sense Disambiguation and Focuses of this Study

NLP systems have difficulties when facing with ambiguities because they do not have an ability to exploit necessary information from a context for disambiguation like a human. Therefore, most of the works on WSD, like other works on ambiguity resolution for NLP systems, tried to imitate a model of human language processing of ambiguity by using information from a context as argued by Ide and Véronis(1998:18) that "context is the only means to identify the meaning of polysemous words."[2] There are two kinds of contexts[3] involved in the disambiguation. The first is **micro context or local context** (the open- and closed-class items that occur within a small window, usually a sentence, around a word). The second is **macro context or global context** which can be subdivided into **topical context** (the open-class words that co-occur with a particular sense, usually within a window of several sentences) and **domain** (context or script activated by the general topic of the discourse). Several issues regarding the use of context for WSD have been addressed in WSD researches as follows:

---

[2] For disambiguating homonymous words (another type of word sense ambiguity), beside context, information from a homonymous word itself is another useful information. See section 2.1 for the different between homonymous and polysemous words and section 2.4.1 for the explanation about information from an ambiguous word itself.

[3] More details about context are explained in section 2.4.2.

(1) Does the combination of various kinds of context yield a better result than using one of them alone?

Leacock, Chodorow and Miller (1998) tested a statistical classifier, TLC (Topical/Local Classifier) which is a **Bayesian classifier**[4] that uses topical context, local context, or a combination of them. The results suggested that local context is superior to topical context. Whether combining both local and topical contexts yields a better result, according to Leacock, Chodorow and Miller (1998), depends on syntactic categories. For example, there is a substantial benefit for a noun *line*, a slightly less for a verb *serve* and none for an adjective *hard*. This is because several senses of verb and adjective tend to occur in more general or unrestricted domain discourses or texts, while different senses of noun tend to occur in different or restricted domain discourses or texts. Beside syntactic categories, the existence of **nontopical senses** (senses that are not limited to a specific topic and appear freely in many different domains of discourse) also limits the use of topical context.

The results of these works as well as the current trend towards disambiguating senses of words with parts of speech other than noun in unrestricted domain texts are reasons why current WSD tasks make use of information from local context only (Ide and Véronis; 1998). Based on these findings, since we are interested in WSD of not only noun but also verb in unrestricted texts, we will use only local context for WSD in this study.

---

[4] See section 2.5.2.1 for more details about Bayesian classifier.

(2) Does the combination of different sources of information from local context[5] yield a better result than using only one source of information?

Ng and Lee (1996) considered multiple knowledge sources including parts of speech of nearby words, morphological forms, unordered set of surrounding words, local collocations, and verb-object syntactic relations, for WSD. They found that all of these information contribute to disambiguation, however, the result suggested further that local collocations yield the highest accuracy. Ng and Lee explained that this finding agrees with the past observation of Choueka and Lusignan (1985) that humans need a narrow window of only a few words to perform WSD. However, McRoy (1998) combined the strongest, most obvious sense preferences drawn from knowledge sources including, parts of speech, word frequencies, collocations, semantic context, role-related expectation, syntactic restrictions. The results suggested that the combination of all sources of information yields a better result than using only one source of information alone because each of them has its own limitation. For example, normally the collocation *wait on* means "serve" (as in "Mary waited on John."), but role-related expectation, such as that the BENEFICIARY be inanimate (as in "Mary waited on the steps.") indicates that *wait on* does not mean "serve".

Therefore, based on the findings from these previous researches, there still be no unified conclusion to this issue. However, since this study is the first step of research on WSD in Thai, we will follow Ng and Lee (1996) and Choueka and Lusignan (1985) in considering only local collocation. This study will reveal whether the use of local collocation alone is sufficient for WSD in Thai.

---

[5] Local context can be divided into two groups: (1) collocation and (2) restriction. See section 2.4.2.1.1 for more details.

(3) "What minimum value of N will, at least in a tolerable of cases, lead to the correct choice of meaning for the central word?", which is a question raised by Weaver (1955).

This question arises because collocation does not need to be immediately adjacent (Haliday, 1961; Ide and Véronis, 1998). Choueka and Lusignan (1985) found that 2-contexts (the context of two words to the left and to the right of the ambiguous words) are highly reliable for disambiguation, and even 1-contexts are reliable in 8 out of 10 cases. Leacock, Chodrow, and Miller (1998) used a local window of ±3 open-class words, arguing that this number showed the best performance in previous test. Yarowsky (1993, 1994a, 1994b) examined different windows of context, including 1-contexts, k-contexts, and words pairs at offsets -1 and -2, -1 and +1, and +1 and +2, and sorted them using a log-likelihood ratio to find the most reliable evidence for disambiguation. He found that the optimal span when considering local context is k = 3 or 4, and k = 20-50 words for global context. However, since Yarowsky also used other information (such as part of speech, global collocation), his result does not suggest the impact of window size alone.

In this study, we will focus only on the impact of window size by exploring different spans of local context and determining the **optimal span** (the distance from an ambiguous word to its sense indicator) for WSD in Thai without considering any other information like parts of speech or global contexts.

(4) Does the process of feature selection (considering the strongest feature of only one context word) yield a better result than the process of combining evidences from all features (considering features of all words surrounding the ambiguous word)?

WSD algorithms that use **Baysian classifier** or **dictionary-based approach**[6] are examples of statistical classifiers that perform no feature selection. Instead, they combine the evidence from all features, that is, they rely on the information from all words in the context as the sense indicators (Manning and Schütze; 1999). For example, in considering a window span of ±5, all features (of words within this span) will be combined together as the evidence for disambiguation. The advantage is that it has high efficiency and accuracy because its ability to combine evidence from a large number of features. However, Manning and Schütze (1999) argued that its strength lies its weakness. It ignores the structure and linear ordering of words within the context when the evidence are combined (this is referred to as a **bag-of-word model**). This bag-of-word model leads to the assumption that the presence of one word in the bag is independent of another which is opposite to the real piece of language, for example, *president* usually occurs in context that has *election* rather than in a context that has *poet*. So, the algorithm does not make use of this useful information for disambiguation.

Yarowsky's **decision list algorithm** (1994) and Brown et al's **information-theoretic approach**[7] (1991b) are examples of another approach that try to rely on just one reliable piece of evidence instead of combining all available pieces of evidences. For example, in considering a window span of ±5, only one feature (of a word within this span) that is the strongest sense indicator will be selected. The features that are

---

[6] See section 2.5.2.2 for details about dictionary-based WSD

[7] See section 2.5.2.3 for details about information theoretic approach and 2.5.2.4 for details about decision list algorithm

considered by Yarowsky are word forms, part of speech, and lemma (morphological root). The features that are considered by Brown et al. are syntactic relation such as object; grammatical category such as tense; co-occurrence such as word to the left. Yarowsky (1994) reported that "relying on only the strongest feature yields the same or even slightly better precision than the combination of evidence approach when trained on the same features."

This study will apply decision list algorithm (Yarowsky, 1994) for WSD in Thai with the assumption that relying on only the strongest feature, which in this study, is a word form can yield the high accuracy.

Following these arguments, we shall propose a WSD model for Thai -- a decision list algorithm using local collocations as clues for disambiguation -- which will be presented in the chapters that follow.

1.3 Hypotheses

(1) Local collocations provide necessary and sufficient information for indicating the correct sense of Thai noun and verb, which are *หัว* /hua4/ and *เก็บ* /kep1/ in this study.

(2) The span of ±5 is sufficient for sense disambiguation of *หัว* /hua4/ and *เก็บ* /kep1/.

(3) Sense indicators of *หัว* /hau4/ are words to the right and sense indicators of *เก็บ* /kep1/ are both words to the right and to the left.

(4) The optimal span for the disambiguation of each sense of *หัว* /hua4/ and *เก็บ* /kep1/ is ±1[8].

1.4 Objectives

(1) To develop a Thai WSD program using decision list collocations, using a noun, *หัว* /hua4/ and a verb, *เก็บ* /kep1/, as case studies.

(2) To find the span of collocation that is sufficient for sense disambiguation of *หัว* /hua4/ and *เก็บ* /kep1/.

(3) To find the optimal span or distance of sense indicators of each sense of *หัว* /hua4/ and *เก็บ* /kep1/.

(4) To analyze the possible meanings of *หัว* /hua4/ and *เก็บ* /kep1/ from Thai corpus.

1.5 Scope

(1) The program will disambiguate senses of *หัว* /hua4/ that is a noun[9] and disambiguate senses of *เก็บ* /kep1/ that is a verb. All other parts of speech of *หัว* /hua4/ and *เก็บ* /kep1/ are excluded from this study because they can be disambiguated

---

[8] In this fourth hypothesis, ±1 means the immediately adjacent word, whether to the right, to the left or both directions of an ambiguous word. Thus, following the hypothesis (3), the optimal span for the disambiguation of each sense of *หัว* /hua4/ is the immediately adjacent word to the right, and the optimal span for *เก็บ* /kep1/ is both the immediately adjacent word to the right and to the left of an ambiguous word.

[9] The noun includes the classifier -- a part of speech in Thai that is included in a noun.

by a part of speech (POS) tagger[10]. This is in according to the argument of Brill (1992), Cutting et al. (1992), cited in Ng and Lee (1996:2) that "POS taggers that can achieve accuracy of 96% are readily available to assign parts of speech to unrestricted English sentence." Thus, it is assumed in this study that POS tagger can help disambiguate senses of words with different parts of speech by disambiguating their parts of speech.

(2) The program will deal only with the senses that derived directly or metaphorically from the word form, which is in the scope of lexical semantics. Senses that must be inferred from the context are excluded. For example, *หัวหงอก* can be considered as consists of two adjacent lexical constituents *หัว* /hua4/ "hair" and *หงอก* /ŋɔɔk1/ "gray" or can be considered as one lexical unit *หัวหงอก* "old man" depending on its surrounding context. Only the first type, which is *หัว* /hua4/ "hair" and *หงอก* /ŋɔɔk1/ "gray" that sense of *หัว* as "hair" is considered.

(3) Only *หัว* /hua4/ and *เก็บ* /kep1/ will be tested as a representative of noun and verb in Thai. Noun and verb are chosen because they are content words which tend to occur with other content words such as adjective (which modifies noun) and adverb (which modifies verb) which are good indicators of senses than function words such as preposition, conjunction, etc. Thus, it is easier to begin the research with noun and verb than other parts of speech. The words *หัว* /hu4/ and *เก็บ* /kep1/ are chosen because of their high frequency of occurrence in a corpus and their highly ambiguous senses, thus they are good representatives for testing the ability of the algorithm in dealing with such difficult cases. Moreover, the reason that this study uses only two words is to lessen the problem of preparing manually sense tagged corpus which will require a lot of time and effort and seem impossible at all in this thesis if every word in the corpus has to be manually sense tagged.

---

[10] POS tagger will tell that *หัว* /hua4/ is a noun or a verb, if it tells that *หัว* /hua4/ is a verb, it also indicates that the meaning of *หัว* /hua4/ is "to laugh". Thus, by disambiguating parts of speech of *หัว* /hua4/, its meanings are also disambiguated.

(4) Different senses of the words หัว /hua4/ and เก็บ /kep1/ will be analyzed based on the Thai dictionary of "The Royal Institute" and a corpus of "Bangkok Business" newspapers. At least 1,000 examples of each word will be extracted from the corpus of about 132-MB from "Bangkok Business" newspapers for the semantic analysis. Then, these examples will be manually sense tagged for the word หัว /hua4/ and เก็บ /kep1/ and used as a training and testing data for this study.

1.6 Required Data

Sense-tagged corpus is required as training and testing data. The data for manually sense tagging must be word segmented.

1.7 Benefits

(1) This study gives benefit to the knowledge of Thai syntactic structure.

(2) This study is a prototype for the further development of word sense disambiguation program for Thai language.

1.8 Methodology: An outline

(1) Collect the data from Thai texts, which in this study are collected from the machine-readable corpus of "Bangkok Business" newspaper.

(2) Analyze data to establish all senses of หัว /hua4/ and เก็บ /kep1/ based on Thai dictionary of "The Royal Institute" and the additional evidence from the corpus for preparing the manually sense tagged training data.

(3) Create the sense-tagged training data by manually assigning the sense to the ambiguous word in a given context.

(4) Train the algorithm by applying "decision lists algorithm" proposed by Yarowsky (1994), which involves the following steps:

For 20 spans, which are the combinations of different numbers and location of context words trained including

- One-word-to-right (1WR) to five-words-to-right (5WR) of an ambiguous word
- One-word-to-left (1WL) to five-words-to-left (5WL)
- One-word-to-right-and-left (1WRL) to five-words-to-right-and-left (5WRL), giving priority to words to the right
- One-word-to-left-and-right (1WLR) to five-words-to-left-and-right (5WLR), giving priority to words to the left

perform feature selection by

Step 1: count the frequencies of co-occurrence of word forms, co-occurring with different senses of an ambiguous word ($C(S_i,W_k)$) and the frequencies of occurrence of word forms $C(W_k)$.

Step 2: compute the probabilities of co-occurrence of word forms and different senses of an ambiguous word to obtain discriminated weight or strength of each co-occurrence as follows:

$$P(S_i|W_k) = \frac{C(S_i,W_k)}{C(W_k)}$$

$$\text{Weight}(S_i, W_k) = \text{Log}\left(\frac{P(S_i|W_k)}{\sum_{j \neq i} P(S_j|W_k)}\right)$$

where, S_i is the senses of an ambiguous word, and W_k is the word forms co-occur with different senses of the ambiguous word.

After training the algorithm for 20 spans, there will be 20 decision lists for 20 spans. Each decision list consists of the co-occurrences between word forms and senses of the ambiguous word and their weights. Collocational patterns that receive higher weight are more statistically significant than those that have lower weight. Therefore, they should be better sense indicators.

(5) Test each decision list of each span separately with new (unseen) texts by, for each span, comparing whether word forms occur in the test data match the word forms in the decision list. If they match, the algorithm will choose the sense that co-occurred with a matched word form that has the maximum collocational weight.

(6) Compare the tested results with the result from manually disambiguation to determine the performance of each span. Then, compare the performance among these spans to obtain the optimal spans and the location of sense indicators for WSD of *หัว* /hua4/ and *เก็บ* /kep1/. Then, evaluate the optimal spans for WSD of *หัว* /hua4/ and *เก็บ* /kep1/ to know how best the algorithm can perform against the lower bound and upper bound performances.

1.9 Outlines of the Coming Chapters

Chapter 2 explains the tasks involved in WSD by, first exploring what word sense ambiguity is (section 2.1) and what WSD is (section 2.2). The second part of this

chapter elaborates the four steps involved in WSD tasks: (1) the analysis of word sense ambiguity (section 2.3), (2) the assignment of word with sense by exploring to useful cues to WSD (section 2.4), (3) exploiting these cues by using several WSD methods or algorithms (section 2.5) and (4) the evaluation of the performance of WSD (section 2.6).

Chapter 3 presents the details methodology of this study. The methodology consists of four processes. (1) The process before training and testing (section 3.1) which consists of data collection (section 3.1.1), word segmentation (section 3.1.2), word sense analysis (section 3.1.3) and word sense tagging (section 3.1.4). (2) The training process (section 3.2). (3) The testing or the disambiguation process (section 3.3). (4) The evaluation process, which involves the calculation of the performance of the algorithm for the evaluation against the lower bound and upper bound performances (section 3.4).

Chapter 4, in section 4.1 and 4.2, reports and explains the results of the disambiguation of *หัว* /hua4/ and *เก็บ* /kep1/ at different spans and the optimal spans for disambiguating *หัว* /hua4/ and *เก็บ* /kep1/ and each sense of *หัว* /hua4/ and *เก็บ* /kep1/.

Chapter 5 is the discussion, conclusion and further suggestions of this study. Section 5.1 discusses the important issues from this study. Section 5.2 summarizes the main points of this study. Section 5.3 suggests the way to increase the algorithm's performance and to further develop WSD program in Thai.

CHAPTER II

LITERATURE REVIEW

In this chapter, we explain the tasks involved in WSD by first, explaining the meanings of word sense ambiguity and WSD. Then, we elaborate four steps involved in WSD tasks. The first step is the analysis of word sense ambiguity. The second step is the assignment of word with sense by exploring to many useful cues to WSD. The third step is exploiting these cues by using several WSD methods or algorithms. The last step is the evaluation of the performance of WSD. The details are as follows.

The following two sections explain the terms word sense ambiguity and WSD. Section 2.1 explains what word sense ambiguity is, how it differs from lexical ambiguity in general, and what is focused in this study. Section 2.2 explains what the task of word sense disambiguation is all about.

2.1 Word Sense Ambiguity

When discussing about a phenomenon in which a word can have more than one meaning or sense, two related terms, namely **word sense ambiguity** and **lexical ambiguity** often cause a confusion on which term should be used to address this phenomenon as both of them often used interchangeably in many researches. This section discusses similarity and difference between these terms and points out that the relevant term for this study is word sense ambiguity.

According to Hirst (1987), there are three types of lexical ambiguity:

(1) Polysemy is a phenomenon in which a word has more than one related meanings (or fine-grained senses) which derived from the same word form and listed in the same lexical entry in a dictionary (Saeed, 1997).

(2) Homonymy is a phenomenon in which a word has completely different meanings or senses which accidentally has the same form and listed in different lexical entry in a dictionary (Saeed, 1997).

Both polysemy and homonymy can have the same or different parts of speech. The following examples will show the differences between homonymy and polysemy with the same and different parts of speech[1]. B*ear* (v) is an example of polysemy with the same part of speech. It can mean, "to carry" "to tolerate" or "to give birth" (EAGLE, 1996). *Plane* (n) is an example of homonymy with the same part of speech. It is ambiguous between "carpenter's tool' and "aeroplane" (Crystal, 1991). *Chair* is an example of polysemy with different parts of speech. It is ambiguous between "a piece of furniture for sitting" (n) and "to seat" (v) (Palmer, 1976). *Bear* (n) is an example of homonymy with different parts of speech. It can mean "a large furry animal" (n) or "to carry" (v) (Crystal, 1991).

(3) Categorial ambiguity is a phenomenon in which a word has more than one part of speech or syntactic category. For example, the word *sink* which can be a noun or a verb (Hirst, 1987).

---

[1] The general criteria for distinguishing betweenhomonymy and polysemy are (1) related or closeness of meanings (2) the historical evidence. See more detailed discussion about the different between homonymy and polysemy in Apresjan (1974), Buitelaar (1998), Lyon (1977), Kilgarriff(1992).

Words that fall into the first two phenomena are called **semantically ambiguous words** as they map more than one sense (Allen, 1995). Words that fall into the last phenomenon are called **categorial ambiguous words** which are words that have more than one syntactic category (Hirst, 1987), and the latter does not exclude the former.

However, the confusion occurs as some works (Allen, 1995, Jurafsky, 2000) used **lexical ambiguity** and **word sense ambiguity** interchangeably to refer to homonymy and polysemy, and **lexical category ambiguity** to refer to categorial ambiguity.

To avoid confusing, we will use the term **word sense ambiguity** (and not lexical ambiguity) to address a phenomenon in which a word can have more than one meaning or sense (which can range from coarse-grained sense in case of homonymy (as in the example of *plane (*n)) to fined-grained sense in case of polysemy (as in the example of *bear* (v)). In other words, our word sense ambiguity will be specific to homonymy and polysemy with the same part of speech because they are ambiguous only in their several meanings. This is to contrast with lexical ambiguity because lexical ambiguity includes homonymy and polysemy with different parts of speech, which are ambiguous in their several meanings as well as their several parts of speech, thus their ambiguities intersect with the categorial ambiguity. Besides, in many researches, lexical ambiguity often includes not only word sense ambiguity and categorial ambiguity but also other kinds of related ambiguities such as **accent ambiguity** (a word can have more than one accent, as in Spanish or French, in textual medium where accents are missing), **capitalization ambiguity** (in the medium of all-capitalized (or case-free) text such as news headlines (for example, *AIDS* is ambiguous between "disease" and "helpful tools")) (Yarowsky, 1994).

2.2 Word Sense Disambiguation

Many researches on word sense disambiguation (WSD) often define the task of WSD before the beginning of the task. For example, Ide and Véronis (1998:3) explained that "In general, WSD involves the associations of a given word in a text or a discourse with a definition or meaning (sense) which is distinguishable from other meanings potentially attributable to the word". In Karov and Edelman (1998:41), "WSD is the problem of assigning a sense to an ambiguous word, using its context." In Fuji (1998:7), "The task of a WSD system is to resolve the lexical ambiguity of a word in a given context. For Schütze (1998) , WSD is the task of assigning sense label to occurrences of an ambiguous word.

However, in specific, WSD task can be divided into two sub-problems: **sense discrimination** and **sense labeling**. (Schütze, 1998) Schütze (1998:97) explained that "sense discrimination divides the occurrences of a word into a number of classes by determining for any two occurrences whether they belong to the same class or not. Sense labeling assigns a sense to each class, and, in combination with sense discrimination, to each occurrence of the ambiguous word." Kilgarriff (1997:3) explained that "sense discrimination involves identifying distinct senses and classifying occurrences of the word as belonging to one of those senses. It does not involve labeling the senses or associating them with any external knowledge source such as a dictionary."

In conclusion, WSD involves the resolution of word sense ambiguity by assigning sense label to an ambiguous word in a given context (which involves semantic analysis), while sense discrimination may be considered a subtask of WSD involving the classification of senses of an ambiguous word by identifying whether the occurrences of a word in different contexts belong to the same class (sense) or not.

Thus, this study involves both distinguishing one sense from the others and assigning each occurrence of an ambiguous word with appropriate sense label.

The next section is the elaboration of steps involved in WSD. According to Ide and Véronis (1998), in general, WSD task takes two steps, namely the analysis of word sense ambiguity (section 2.3) and the assignment of words with senses (section 2.4 and 2.5). In this review, the evaluation of the algorithm performance is also included as it receives equal consideration as the first two steps in many researches such as Resnik and Yarowsky (n.d.) and Fuji (1998) (section 2.6). They can be explained in details as follows.

2.3 The Analysis of Word Sense Ambiguity

This step involves the determination of all possible senses of the ambiguous words relevant to the target text. Many questions arise at this step. First, "How can we know how many senses does a word have?" The lexicographers are persons who can answer these questions well and everyday dictionaries are the products of the lexicographers' researches that provide a huge information on word sense listings especially those of polysemous words. (Kilgarriff, 1992)

Second, then, "What are criteria used by lexicographers in determining which senses of words are worth listing in a dictionary?" According to Kilgarriff, 1992, lexicographers use **Sufficiently Frequent and Insufficiently Predictable (SFIP)** as criteria for determining which senses of words are worth listing in a dictionary. **Sufficiently frequent** criterion means only senses occurring frequently are listed in a dictionary entry due to the commercial constraints on the size of a dictionary.

As for **insufficiently predictable**, according to Kilgarriff (1992:52) "Sense is predictable if language learners or users familiar only with a core sense for the word in

question could, on hearing the word in a context demanding some other reading, correctly interpret it and draw appropriate inferences." To explain this argument, firstly, the different between **usage** and **sense** should be addressed. Usage is a particular meaning of a word occurs in a particular context. This means that if a word occurs in two different sentences or even occurs in the same sentence with two different occasions, it has two usages. For example, *brick red* and *pillar-box red* give the word *red* the two usages. For usages of words to become senses, all their members must have some aspect of the meaning in common to say that they all mean the same thing. Thus, usage is considered as a **token** while sense is considered as a **type** and only sense or type is to be listed in a dictionary. So, *brick red* and *pillar-box red* fall into the same cluster, and only "red" will be listed as a sense of the word *red* in a dictionary (and not "red as a brick" or "red as a pillar-box" sense). Then, a criterion to decide that two senses are so distinct that they are listed as different sense in a dictionary is that they must be unpredictable from each other. For example, for the senses of *newspaper* as "copy" or "corporation", even though they are dissimilar and seem to be listed as different senses in dictionaries, but they will not be listed separately because they can be predictable from each other. The distinction of sense using predictability criterion can be used to explain the determination of the senses of 3 related kinds of words as follows (Kilgarriff, 1992).

(1) Homonymy: The different senses of homonymy must be listed in a dictionary certainly because their different senses cannot be predictable from each other.

(2) Collocation: The sense of collocation will be listed as a whole because it is **opaque** -- each constituent of collocation cannot be divided into semantic constituents, and thus, considering the predictability of sense of each word for sense distinction is side step. However, it should be emphasized that, in this case,

collocation means **idiom**, which is a special kind of collocation as their senses are opaque[2]

In Thai, content or open-class words usually co-occur and seem like idioms or phrases. They are called **compound words**. However, the senses of these compound words can be opaque or transparent. If their senses are opaque, they will be like idioms, so their senses will be listed as senses of whole words (their senses are **non-compositional**). For example, *หัว* /hua4/ + *หน้า* /naa2/ means "leader" which its sense is changed totally from the original senses of both *หัว* /hua4/ "head[3]" and *หน้า* /naa2/ "face". If their senses are **transparent**, their constituents only simply co-occur and the senses of each constituent will be listed separately, then the predictability is considered. For example, the sense of *หัว* /hua4/ that co-occurs with such open-class items as *เก่า* /kao1/ "old", *ใหม่* /mai1/ "new", *โบราณ* /booraan/ "old-fashioned", as in *หัวเก่า*, *หัวใหม่*, *หัวโบราณ* is compositional. It can be divided into two semantic constituents, in which *หัว* /hua4/ means "viewpoint". Besides, this sense of *หัว* /hua4/ cannot be predictable from the existing sense, thus it should be considered as a new sense of *หัว* /hua4/.

(3) polysemy: the senses which are highly predictable from the existing senses will not be listed, as in the example of *newspaper*. But if the two senses are not predictable from each other, they will be listed as distinct sense. For example, the sense of dog as "a species" and as "the male of that species", even though they seem similar, but they are not predictable from each other. Thus, they are listed separately.

The third question is "Are senses listed in a dictionary workable for WSD task or other NLP system?" The criterion that only listing sense but not usage is questioned

---

[2] See the explanation about opaque meanings of idioms and phrases and transparent meanings of collocations in section2.4.2.1.1.1

[3] See section 3.2 for the analysis of senses of *หัว* /hua4/ and *เก็บ* /kep1/.

by many lexicographers (such as COBUILD project, 1980) as they turn to the **corpus-based lexicography** which consider not only sense but also usage. The predictability criterion is cited as only human can have the ability of predictability and not a machine. Kilgarriff (1992) suggested that the system need generalization for having predictability like human and dealing with this issue is one of the important tasks of computational lexicographers.

In this study, the dictionary is used as the main source of sense listings, even though its criteria of sense listings (only senses but not usages, and SFIP criteria) are questionable. Additional senses that do not exist in a dictionary will be added when it is found from the training corpus. For example, *หัว* /hua4/ when co-occurs with *ดำ* /dam/ "black", *แดง* /dɛɛŋ/ "red", *หงอก* /ŋɔɔk1/ "grey", as in *หัวดำ*, *หัวแดง*, *หัวหงอก* when substituting *หัว* /hua4/ with *ผม* /phom4/ "hair" the meaning remains the same. Thus, *หัว* /hua4/ refers to a new sense as "hair"[4] because this sense is not yet listed in the dictionary.

The next two sections involve the assignment of words with senses. How to correctly assign the words with the right senses requires knowledge about useful cues to WSD (section 2.4) and knowledge about the ways to use these cues (section 2.5).

2.4 Cues to Word Sense Disambiguation

Cues to WSD can be the information provided by an ambiguous word itself (section 2.4.1), the information from its surrounding linguistic contexts (section 2.4.2) and the information from non-linguistic contexts (section 2.4.3), which are explained in details as follows.

---

[4] See section 3.2 for the analysis of senses of *หัว* /hua4/ and *เก็บ* /kep1/.

2.4.1 Knowledge of an ambiguous word itself

As stated earlier in this study, context is the most important cues to WSD, however, the information from an ambiguous word itself also plays a role in WSD. It has been used by many researches as follows.

2.4.1.1 Morphology and syntactic tags

McRoy (1992) considered information from morphological analysis (an analysis of each word into its root and affixes) as the first indicator of sense of an ambiguous word. The information from morphological analysis will be used for tagging part of speech (determining the correct part of speech for the word) which in turn, help determining the correct meanings of an ambiguous word.

2.4.1.2 Frequency or dominance of meanings

Counting the frequencies of senses and considering a preference for the common interpretations of senses over the rarer senses is another useful information. Allen (1997) gave the following example.

Assume that there are 5845 uses of *bridge* in a corpus, in which there are

5651 uses of STURCTURE1

194 uses of DENTAL_DEV37

from this data, *bridge* will occur in the STRUCTURE1 sense almost every time. If a training data is representative, this information would give the right answer 97 percent of the time.

- **Knowledge of the ambiguous words and WSD**

Although these cues are helpful to disambiguation, they have never been used alone without any consideration of context because of their several limitations. First, morphology and part of speech are useful only when disambiguating senses of words, which have different part of speech. When disambiguating senses of words, which have the same part of speech, this information is not useful. Besides, information from morphological analysis is useful only for inflecting language, for isolating languages like Thai, this information is not applicable. For the information provided by the frequency of sense, even though it is simple and in according to the human storage and retrieval of senses (senses of an ambiguous word will be stored in human memory according to their frequencies, with the highest will be retrieved first (Simpson, 1981)), it has very low accuracy, only 70 percent of the time for a broad range of English (Allen, 1997). Thus, if we want more accuracy, we have to consider the effect of context.

2.4.2 Knowledge of the Context[5]

This section discusses the use of contexts for WSD, which can be divided into two types, namely linguistic and non-linguistic context.

In Palmer (1976), **non-linguistic context, which** he called **context of situation** conveys the meaning of a word in term of the context in which language is used, while

---

[5] In Lyon (1977), there are two kinds of lexical relations: (1) paradigmatic relation (2) syntagmatic relation. Paradigmatic relation is a relationship between a word and other words that can replace it (e.g. homonymy, meronymy,etc.) Syntagmatic relation is a relationship between a word and other words that occur in the same context. So, syntagmatic relation corresponds to context of words.

linguistic context conveys the meaning of a word in term of the context in which language occurs.

2.4.2.1 Linguistic context

"Interpretation of natural language is inherently context-sensitive. Most words in natural language are ambiguous and their meanings are heavily dependent on the linguistic context in which they are used. The study of lexical semantics cannot be separated from the notion of context. In different situations or contexts, the same sentence may be resolved in different ways." (Zhai, 1997:1)

Palmer (1976) gave the following example as an illustration why context is necessary in distinguishing between different meanings of an ambiguous word, especially polysemous word such as *chair*.

(i) sat in a *chair*

(ii) the baby's high *chair*

(iii) the *chair* of philosophy

(iv) has accepted a University *chair*

(v) the *chair* of the meeting

(vi) will *chair* the meeting

(vii) the electric *chair*

(viii) condemned to the *chair*

We may not notice that there is any ambiguity in these sentences because we are presented with contexts, so we can interpret the meaning of *chair* by knowing its contexts. The effect of context to the interpretation of sense, then, is the most important consideration in WSD.

In the early studies of contexts, namely Firth, 1953 (cited in Palmer, 1976); Palmer, 1976; Lyon, 1977; Cruse, 1986, only the term collocation, collocation and grammar (or syntactic restriction) have been proposed. However, in later works, namely Hirst, 1987; Ide and Véronis, 1998; Buitelaar, 2000, contexts included such notions as scripts, discourses, domains, etc. We have studied these works and found that, beside the different terms used by different authors, these terms can be arranged into 2 groups namely, local or micro context, and global or macro context. **Local context** refers to words occurring in the same sentence as the ambiguous word while **global context** refers to words occurring in other sentences. Both local and global contexts can be divided into sub-groups, which are explained in details as follows.

2.4.2.1.1 Local or micro context

"Local or micro context is generally considered to be some small window of words surrounding a word occurrence in a text or discourse, from a few words of context to the entire sentence in which the target word appears." (Ide and Véronis; 1998:19) For Ng and Lee (1996), "Local context is the open- and closed- class items that occur within a small window around a word." In conclusion, local context is the **open-** and **closed-class items**[6] that occur within a small window, usually a sentence, around a word.

Local or micro context can be subdivided into two groups:

---

[6] In English, open-class words consist of nouns, adjectives, verbs and adverbs. They are four main classes of words that are necessary to form sentences and also contribute to the meaning of sentences. Close-class words consist of articles, pronouns, prepositions, particles, quantifiers, conjunctions, etc. They are also necessary to form a sentence, however, contribute little to the meaning of sentences. Nouns, adjectives, verbs and adverbs, are called open-class words as new words in these classes are regularly introduced into the language, while close-class words are fixed as new words in these classes are rarely introduced. (Cruse, 1986; Allen, 1995)

2.4.2.1.1.1 Collocation

The term **collocation** has been explained from two different perspectives as follows.

From **linguistics perspective**, the degree of what is called collocation can be ranged from the strongest degree as **idiom** which is a special kind of collocation (Palmer, 1976; Cruse, 1986) to **bound collocation** (Cruse, 1986; Buitelaar, 2000) to **usual** or **habitual** (Firth, 1953; Palmer, 1976; Lyon, 1977; Cruse, 1986), and to the lowest degree as simple co-occurrence (Buitelaar, 2000). **Opaque meaning** -- non-transparent meaning where each constituent cannot be a semantic constituent (Cruse, 1986). -- and **mutually selective** -- the semantic integrity or cohesion where its constituent is highly restricted contextually (Cruse, 1986) -- are used as criteria to differentiate among these terms. They can be showed in table as follows.

	Mutually selective	Not mutually selective
Opaque meaning	Idioms, phrases	---
Transparent meaning	Usual, habitual co-occurrence, Bound collocation	Simple co-occurrence

Table 1: Different degrees of collocations with opaque meaning and mutually selective criteria.

Palmer explained **idiom** as a special kind of collocation because the meaning of its combination is **opaque**. This means that its combination gives a new meaning as if it is a new and single word and that meaning is not related to the old meaning of each individual word. Such idioms are *kick the bucket, fly off the handle "*to become suddenly or violently angry or excited"*, spill the beans "*to divulge secret information", *red herring* etc. So, in *kick the bucket,* the combination of *kick* and *the bucket* not only give a collocation but also give the new meaning of its collocation "to die". Besides,

idioms can be defined in terms of non-equivalence in other languages. For example, *kick the bucket* or *red herring* are idioms because they cannot be directly translated into French or German.

Cruse (1986) and Buitelaar (2000) has a concept of **bound collocation** which is similar to idiom such as *kick the bucket* in which two constituents cannot be separated as in *foot the bill*. Consider these sentences,

(i) I've just got the *bill* for the repairs.
(ii) ?I hope you don't expect me to *foot* it.

However, it is also un-idiom-like in that some of its constituents can be freely modifiable, thus we can have *to foot the electricity bill* but not *to kick the red bucket.*

For Firth (1953), Lyon (1977) Palmer (1976), collocation means **habitual** or **usual** co-occurrence. Firth's famous example is the word *ass* which occurred in "You silly *ass*", "Don't be such an *ass*" and with a limited set of adjectives such as *silly*, *obstinate*, *stupid*, *awful* and *egregious.* However, unlike idiom, the meanings of habitual co-occurrence are fully transparent as each lexical constituent is also a semantic constituent (Cruse, 1986).

Buitelaar (2000), collocation is **co-occurrence** of strings or sequences of words with simple structure. Thus involve neither mutually selection nor opaque meaning.

From **computational perspective**, the degree of collocation can be ranged from significant co-occurrence (Ide and Véronis, 1998; Yarowsky, 1994) to simple co-occurrence (Haliday, 1961 Yarowsky, 1994; Ng and Lee, 1996). The **probability of grater-than-chance co-occurrence** is being used as a criterion. From WSD perspective, distance is involved because collocation is not necessary limited to immediately adjacent words. Hirst (1987) use the concept of collocation by stating that nearby words are useful for WSD. These terms can be explained in details as follows.

Ide and Veronis, (1998:20), based on Haliday's, defined **significant collocation** as "a syntagmatic association among lexical items, where the probability of item *x* co-occurring with itemsd *a, b, c*… is greater than chance."

Haliday's (1961), cited in Ide and Veronis (1997:20) definition of collocation as "the syntagmatic association of lexical items, quantifiable, textually, as the probability that there will occur at *n* removes (a distance of *n* lexical items) from an item *x*, the items *a,b,c*…" imply simple co-occurrence and that co-occurrence is not necessary be immediately adjacent.

Yarowsky's (1994), Ng and Lee (1996) use the concept of collocation in their WSD researches in which collocation implies words frequently adjacent to or near each other (literally, in the same location) and does not imply idiomatic or non-compositional associations. This is also the definition of collocation used in this study.

According to Haliday's definition, which implies the distance between collocational words, the concept of **distance** or **span** is being considered in WSD. Weaver (1955) is the first person who raised a question concerning the optimal span or distance for WSD. If X is an ambiguous word, the optimal span is the distance from X to its sense indicator either on the left or on the right. For example, if a context of X is …$b_3\ b_2\ b_1\ X\ a_1\ a_2\ a_3$…, if the sense indicator of X is a_2, the distance from X to a_2 is the optimal span for the disambiguation.

Evidences from many researches (Yarowsky, 1994; Leacock, Chodorow and Miller, 1998) suggested that ± 3 open-class words is the optimal span for WSD in English. However, the optimal span for other languages may be different because of different structure of language. The optimal span or distance of collocation consists of two factors namely, the numbers of words and the sides of words.

In Thai, we usually find the structure like หัว /hua4/ "head" + ปลา /plaa/ "fish", in which it takes only 1 word to know the meaning of หัว /hua4/. This is

because in Thai content words, which can be used to disambiguate the word senses, are usually immediately adjacent to the ambiguous word[7]. The immediately adjacency also implies that a word and an ambiguous word are in the same syntactic construction, such as in the same compounded unite, phrase, or sentence. Thus, they are semantically related. This is the reason why the hypothesis of the optimal span for WSD of *หัว* /hua4/ and *เก็บ* /kep1/ in this study is set to be one.

When considering the location of collocated word for WSD, the indicator in the example *หัว* /hua4/ + *ปลา* /plaa/ is on the right. Thus, we expect the syntactic structure of head and modifier to play an important role in the disambiguation of *หัว* /hua4/. In Thai, the head noun is usually on the left of the modifier (adjective or noun, which are content words that tell something about its head). However, for a verb, we expect that both words to the right and to the left are useful indicator in both English and Thai. This is because both Thai and English are in the same language typology, that is, SUBJECT-VERB-OBJECT where both SUBJECT and OBJECT play an important role in disambiguation. This is the reason why the hypothesis of the sense indicator's location of *หัว* /hua4/ is set to be on the right , while the sense indicator's location of *เก็บ* /kep1/ is set to be both on the left and on the right.

Therefore, the optimal span for disambiguation of each sense of *หัว* /hua4/ should be one-word-to-right (1WR) and *เก็บ* /kep1/ should be one-word-to-right-and-left (1WRL) according to the reasons explained above.

---

[7] Many researches of WSD in English, which do not consider the close-class or function words in the disambiguation, suggested that ± 3 is the optimal span for the disambiguation. In this study, even though we include the close-class words in the disambiguation, we still expect that the optimal span for the disambiguation in Thai is one word. This is because we expect to find more usage of content word that immediately co-occurs with another content word, as in *หัว* /hua4/ + *ปลา* /plaa/ "fish" the use of content + function + content, as in *หัว* /hua4/ + *ของ* /khɔɔŋ 4/ "of" + *ปลา* /phaa/ "fish".

The concept of collocation is applied to WSD by Hirst (1976) which stated that a semantic association between one sense of the ambiguous word and nearby words gives rise to the determination of an appropriated meaning. For example,

(i) The dog's *bark* woke me up.

Just know the meaning of *dog* without considering the global context, we know that *bark* does not mean "surface of tree".

There may be a case when nearby word is itself ambiguous. For example, *deep pit*, *deep* can mean "profound" or "extending far down" and *pit* can mean "fruit stone" or "hole in the ground". However, there will be only one combination of meanings that fits together, this is called **mutually disambiguating**. So, when *deep* is near *pit*, the "extending far down" sense of *deep* and the "hole in the ground" sense of *pit* is selected.

- **Local collocation and WSD**

In sum, local collocation provides very useful information to WSD as explained above. Besides, local collocations can be easily captured, especially in Thai, in which we usually find content words (which are better sense indicators than function words (Allen; 1997)) to be immediately adjacent to another content word. For example, in หัว /hua4/ + ปลา /plaa/ "fish", we easily know that หัว /hua4/ means "head" because the immediately adjacent word is ปลา /plaa/.

2.4.2.1.1.2 Restriction

This type of local context refers to rules that restrict the combination of words. These restrictions can be used to disambiguate the meaning of word. Restriction can be divided into two subgroups as follows

2.4.2.1.1.2.1 Syntactic restriction

Syntactic restrictions are the rules that specify or restrict the combination or co-occurrence of syntactic elements or features. The syntactic elements can be **syntactic cues** such as grammatical SUBJECT, OBJECT, COMPLEMENT, **grammatical cases** like AGENT, PATIENT, INSTURMENT and **dependency structure** like head and modifier or argument. The applications of such rules with WSD are explained as follows.

Hirst (1987) gave the example of the restriction on syntactic cues, which are useful for selecting the correct meaning of an ambiguous word *kept*.

(i) Ross *kept* staring at Nadia's decolletage.
(ii) Nadia *kept* calm and made a cutting remark.
(iii) Ross wrote of his embarrassment in the diary that he *kept*.

Knowing that the word *kept* in the sense of "continue to do" requires its object to be a gerund, "continue to be" sense requires adjectival phrase and "maintain" sense requires a noun phrase, these three meaning of keep can be disambiguated by these syntactic cues.

According to Buitelaar (2000), dependency structure is an analysis of the semantic structure of phrases and sentences. For example, the verb *run* takes one argument as a direct object that is *business*. This can help in WSD, for example, in "He

run a private business", the head (*run*) that takes an argument as a direct object (*business*) gives the meaning of *run* as "to operate" (and not other meanings like "to go rapidly").

Hirst (1987) gave the following examples showing **case slot flags and restrictions** as cues for disambiguation.

(i) Ross *played* with his toys.
(ii) Ross *played* his guitar.
(iii) The baby *played* with the guitar.
(iv) Ross *played* football.

In (i) and (iii), we can easily know that the word *played* is used in the sense of "recreation" because its PATIENT is flagged with the word *with*. In (ii), *played* is used in the sense of "music-making" because its PATIENT is flagged with OBJECT *a musical instrument*. This is easily distinguishable from (iv), its PATIENT is flagged with OBJECT *football,* so *played* is used in the sense of "sport-playing".

2.4.2.1.1.2.2 Semantic restriction

Selectional restriction is a rule that restricts the combination of certain semantic categories (Dijk,1977). A sentence like "The table was laughing" is semantically deviant because it violates the selectional restriction rule, which indicates that the verb *laughing* requires a HUMAN subject. Selectional restriction provides information for WSD in the same way as in this example. For example, in "The dishwasher reads the article", from selectional restriction rules, we know that the *dishwasher* is a HUMAN sense, not a MACHINE sense, because the verb *read* requires its subject to be HUMAN (Allen; 1997). Semantic restriction differs from syntactic restrictions in that

it is the restriction about the combination semantic categories such as HUMAN, ANIMATE.

- **Restriction and WSD**

Though syntactic and semantic restriction can provide a lot of useful information for WSD, there are some limitations. First, semantic restriction cannot deal with the deviance (the violation of restriction) caused by meaning extensions like metaphor, metonymy etc. Second, before these syntactic and semantic features can be used for WSD task, they must be manually coded into the lexicon. This process is time-consuming and hardly developed. These are the main reasons why we exclude information provided by both restrictions from this study.

2.4.2.1.2 Global or macro context

Global contexts can be ranged from several sentences to several discourses to the whole document, which can be explained as follows.

2.4.2.1.2.1 Topical context

According to Ide and Véronis (1998), topical context includes substantive words that co-occur with a given sense of a word, usually within a window of several sentences. Ng and Lee (1996) defined topical context in a similar way as "the open-class words that co-occur with a particular sense." From both works, we can conclude that topical context are the open-class words that co-occur with a particular sense, usually within a window of several sentences. Topical context provides the topic or knowledge of several sentences or discourses.

Hirst (1987) gave the following example of topical context involved in WSD.

(i) The lawyer stopped at the *bar*, and turned to face the *court*.

Here, *Bar* could refers to "the railing in a courtroom" when *court* refers to "the judiciary assembled in the court room", or it could refer to a "drinking establishment" when *court* refers to "courthouse across the street", or a "tennis court". Inference on the preceding context (within a paragraph or the preceding paragraph) would be the last resource when other cues fail.

2.4.2.1.2.2 Domains or scripts

Although a word can refer to different senses, when it is occurs in a specific domain, it tends to have only one meaning. For example, in the context of restaurant setting

(i) The waiter *served* the lasagna.

In a restaurant script, other meanings of *serve* such as in tennis script will not be noticed.

Buitelaar (2000) gave an example of abbreviations such as AI, which can means "artificial intelligence" and "amnesty international". He stated that it is unlikely to find both meanings in the same corpus or document. So, in a science script or domain, only "artificial intelligence" is likely to occur.

However, Hirst (1987) explained there are cases where scripts cannot be used as a cue for disambiguation.

(1) When there seem to be more than one script in a sentence, as in the following sentences.

(i) The lawyer stopped at a *bar* for a drink.

(ii) The waiter *served* in the army.

In (i) there are two scripts, lawyering and restaurant scripts. In (ii) there are restaurant and army scripts. So the problem arises as to which scripts should be chosen in order to determine the right meaning of ambiguous words.

(2) Even when there is only one script, an ambiguous word, especially polyseme may not be disambiguated. For example, in the lawyering script, the word *bar* could mean "the physical bar of courtroom" or "the legal profession".

- **Global context and WSD**

Information from topical context or discourse seems to be useful for WSD with the assumption that the larger the context, the better the performance of WSD. However, many researches (Agirre and Rigau (1995, 1996) revealed the opposite results that too much context can reduce the performance of WSD. Domain is useful for WSD tasks, when disambiguation is carried out in a restricted domain text based on the assumption of one sense per one domain. However, with the current trend of NLP applications towards unrestricted domain text, the usefulness of information from domain is limited.

2.4.2.2 Non-linguistic context

Non-linguistic context refers to inference and world knowledge. It seems to be the last resort for WSD when the information from context is weak or not useful. Levow (1997) gave the following examples to explain why inference and world knowledge are important for WSD. In Hebrew, the word *hagira* is ambiguous between "immigration" and "emigration". In this sentence "According to the new *hagira* bill

every Soviet citizen will have the automatic right to receive a passport valid for five years", knowing the right meaning of *hagira* requires some reasoning that a bill about passports for Soviet citizens must be a soviet bill, so passport issuing should be related to leaving rather than entering the country. Another example is that, in Chinese, *Gou chi ji* can be translated variously as "Dogs/ Dogs eat/ate/eats/have eaten chicken/chickens." Chinese has no surface inflection related to singular/plural or tense distinctions and all of these combinations are valid. Only general inference from knowledge about the event can resolve this ambiguity.

- **Inference and WSD**

From the above examples, we can see that inference and world knowledge are very useful when contexts like surface co-occurrence and global context provide no cues. However, this sort of meanings is beyond our scope because the interpretation of them required some reasoning beyond what the machine can get from a linguistic context.

From the strengths and weaknesses of these cues to WSD discussed above, we choose local collocation as a cue to WSD in this study.

2.5 Previous Researches on Word Sense Disambiguation

After exposing to several useful cues to WSD, the next concern is how to make use of such information for disambiguation. This section presents several methods of disambiguation.

2.5.1 Word sense disambiguation methods

According to Wilks and Stevenson, 1997; Mihalcea and Moldovan, 1998, WSD methods can be divided into two types according to the processes and the lexical knowledge sources which the algorithms rely on. They are corpus-based method and knowledge-based method. The first method can be further subdivided into 2 types namely supervised learning and unsupervised learning

2.5.1.1 Corpus-based method

A method that involves training process and relies on information from a training corpus. This method can be further subdivided into two types as follows.

2.5.1.1.1 Supervised training

This method needs to be trained on sense-tagged (disambiguated) corpus. Supervised training is a classification task in that there is a training set of exemplars where each occurrence of the ambiguous word w is annotated with a semantic label (usually its contextually appropriate sense s_k). The task is to build a classifier which correctly classifies new case (sense) based on their context of use c_i (Manning and Schütze, 1999). Yarowsky's decision list algorithm (1994, 1994a, 1994b), Gale et. al. 's Bayesian classification (1992b) and Brown et al. 's information theoretic approach (1991) are examples of researches that used this method.

This method has the prominent advantage in that it yields high accuracy because its decision is based on choosing the sense with the highest conditional probability. However, the need for training with sense-tagged corpus which is usually done manually lessen its advantage. This is because creating disambiguated corpus is time-consuming and costly. This leads to the problem of knowledge acquisition

bottleneck due to the lack of sense tagged corpus. Besides, it usually tested with restricted domain text and with other kinds of ambiguity resolution such as homograph disambiguation (Yarowsky, 1994), accent-restoration (Yarowsky, 1994) rather than polysemy.

2.5.1.1.2 Unsupervised training

This method does not need to be trained on sense-tagged corpus. This is an attractive method proposed by Yarowsky (1995) that can solve the problem of creating manually sense tagged corpus. The basic idea is that instead of training with the whole evidence of senses from manually sense-tagged corpus, it trains with seed collocations (which tend to occur in a multiple times in a corpus) representative (indicative) of each sense. For example, the examples of seed collocation for the word *plant* are *plant life* (occurs 82 times in a corpus or equal 1%) which indicates sense A and *manufacturing plant* (occurs 106 times or equal 1%) which indicates sense B. These small set of seed examples, then, be incrementally augmented with additional examples of each sense, using a combination of two properties of human language, that are, **one sense per collocation** -- "nearby words provide strong and consistent clues to the sense of a target word" (Yarowsky; 1995:1) -- and **one sense per discourse** -- "the sense of a target word is highly consistent within any given document" (Yarowsky, 1995:1).

This method receives much attention recently. However, its major disadvantage is that its senses are not well defined -- "sense disambiguation is not carried out relative to any well defined set of senses, but rather an ad hoc set" (Wilks and Stevenson, 1997). This is because it uses only small seeds of example not the whole evidence form sense-tagged corpus. Besides, if the seed collocation is wrongly chosen in the first place, the rest will be effected.

2.5.1.2 Knowledge-based method

This method involves no training process from a large corpus but employs information form external large-scaled lexical knowledge sources which usually are in the form of machine readable dictionaries (MRDs) such as WordNet[8] (Miller, 1990), Longman Dictionary of Contemporary English (LDOCE) (Procter, 1978).

2.5.1.2.1 WordNet and WSD

The examples of researches using WordNet are Agirre and Rigau (1996), Mihalcea and Moldovan (1998). The basic idea of using the information from WordNet is that the words that fall into the same semantic class and have the same concept will have closely related relationship. So, one meaning of an ambiguous word will be chosen over others because it has semantic closeness (which can be determined by measuring conceptual distance among concepts[9]) with its contextual word. By this way, the system needs to know how words are clustered in semantic classes and how semantic classes are hierarchically organized. The lexical knowledge that provides this information is WordNet, which is a broad semantic taxonomy for English.

---

[8] "WordNet is an on-line lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory. English nouns, verbs and adjectives are organized into synonym sets, each representing one underlying lexical concept. Different relations link the synonym sets." (Miller ,et al., 1993:1)

[9] The factors that has to be considered when measuring conceptual distance are : (1)The length of the shortest path that connects the concepts involved : the shorter the path, the closer the relationship among concepts between that path (2) The depth in the hierarchy : the deeper the hierarchy, the closer the relationship among concepts in that hierarchy (3) The density of concepts in the hierarchy : the denser the hierarchy, the closer the relationship among concepts in that hierarchy (4) the measure should be independent of the number of concepts that are measuring (Agirre and Rigau, 1995)

2.5.1.2.2 LDOCE and WSD

Subject-filed codes provided in LDOCE can be used for WSD. The LDOCE MRD contains **subject-field codes** that indicate the semantic field (network or taxonomy) to which the senses of a lexical item belong. For example, from the definition below, the relationship of *COUP* ISA *CAR* ISA *VEHICLE* can be known.

coup (n.) "an enclosed **car** with two doors and a sloping back."

car (n.) "a road **vehicle** with usually four wheels which is driven by a motor."

The examples of 'subject-field codes' are ZOOLOGY, BOTANY, SPORTS, RELIGION, etc. The main subject fields also contain subfields. For example, SUBSTANCE has the subfileds LIQUID and GAS. These information are useful for sense disambiguation.

Knowledge-based method attracts several researchers because of its advantage that it needs no large manually sense-tagged corpus, instead it relies on the lexical knowledge sources that already exist. Besides, it can be tested with unrestricted domain text and fined-grain sense like polysemy. However, the results from many researches suggest very low accuracy (55% accuracy (SensEval-1[10]).

In this study, we choose supervised training, however, not because it is the most attractive but because it is the best suit with this study for the following reasons.

---

[10] SensEval-1 is a project held by Association of Computational Linguistics with the purpose of evaluating the strengths and weaknesses of WSD programs.

First, MRDs are not publicly available in Thai[11]. Though, sense tagged corpora are also not available in Thai, creating our own sense tagged corpus on two words is easier and more possible than creating MRDs. Second, even though unsupervised method seems to be the most attractive because of no sense-tagged corpus is required, we choose to follow the traditional corpus-based supervise training method for this first step of WSD in Thai because it is more understandable and easier for implementation.

2.5.2 Corpus-based WSD: supervised training

This section discusses four theoretical supervised learning algorithms namely, (1) Bayesian Classification (Gale et. al., 1992b) (2) Dictionary-Based Approaches: Disambiguation Based on Sense Definitions (Lesk, 1986) (3) Information Theoretic Approach (Brown et al, 1991b) and (4) Decision List Algorithm (Yarowsky, 1994, 1994a, 1994b), which is the algorithm applied in this study. The first two are similar approaches in that they perform features combination, which consider all possible features of all words surrounding an ambiguous word as cues to WSD. The last two perform feature selection, which considers all possible feature of context words in dictionary definition as cues to WSD.

2.5.2.1 Baysian classification

The basic idea of applying Bayesian classification to WSD is that Bayes classifier will explore every content word surrounding an ambiguous word (which has already been sense tagged) in a large context. Each content word will provide features useful for disambiguation. The classifier will not choose only one feature, instead, it

---

[11] NECTEC and KMIT developed MRDs in Thai for the purpose of NLP researches, however, they are not publicly available.

will combine evidence from several features for decision making. The Bayes' decision rule is as follow:

Bayesian decision rule: decide s' if $P(s'|c) > P(s_k|c)$ for s_k ne s' [12]

$P(s_k|c)$ is the probability of being s_k after knowing c, which can be determined by the following formula:

$$P(s_k \mid c) = [P(c \mid s_k) / P(c)]\, P(s_k)$$

where, $P(s_k)$ is **prior probability** of sense s_k (the probability of being s_k before knowing c). $P(s_k)$ will be updated by $P(c \mid s_k) / P(c)$ and results in $P(s_k \mid c)$ which is the **posterior probability** (the probability of being s_k after exposing or knowing the evidence provided by c). The value of $P(c \mid s_k)$ can be estimated as

$$P(c \mid s_k) = \frac{C(c, s_k)}{C(s_k)} = \frac{\text{total occurrence of c with } s_k}{\text{total occurrence of } s_k}$$

$$P(s_k) = \frac{C(s_k)}{C(w)} = \frac{\text{total occurrence of } s_k}{\text{total occurrence of w}}$$

To simplify the task, the classifier P(c) will be eliminated because it is constant for all senses, thus does not influence the answer. The log is added to make the computation simpler.

---

[12] The followings are symbols and their meanings which are used in this section: w is an ambiguous word , $s_1,\ldots,s_k,\ldots,s_K$ are senses of the ambiguous word (w), $c_1,\ldots,c_i,\ldots c_I$ are contexts of w in a corpus, $v_1,\ldots,v_j,\ldots v_J$ are words used as contextual features for disambiguation.

derivation of formula: $s' = \arg_{sk} \max P(s_k|c)$

$= \arg_{sk} \max [P(c \mid s_k) / P(c)] P(s_k)$

$= \arg_{sk} \max P(c \mid s_k) P(s_k)$

$= \arg_{sk} \max [\log P(c \mid s_k) + \log P(s_k)]$

This classifier (Gale et al. 1992b) is an example of Bayes classifier called **Naïve Bayes Classifier**. Naïve Bayes assumption explains c in term of v_j in c, where v is features of context words. This means that the context of w is the sum or the combination of features of context words (v_j) in the context.

Naïve Bayes assumption: $P(c \mid s_k) = P(\{v_j \mid v_j \text{ in } c\} \mid s_k) = \prod_{vj \text{ inc}} P(v_j \mid s_k)$

Naïve Bayes assumption leads to two consequences. (1) All the structure and linear ordering of words within the context are ignored. This is why this model is often referred to as a **bag of words model**. (2) The presence of one word in the bag is independent of the others.

Naïve Bayes decision rule: decide s' if $s' = \arg\max_{sk} [\log P(s_k) + \prod_{vj \text{ in } c} \log P(v_j \mid s_k)]$

where, $P(v_j \mid s_k) = \frac{C(v_j, s_k)}{C(s_k)} = \frac{\text{total occurrence of } v_j \text{ with } s_k}{\text{total occurrence of } s_k}$

$P(s_k) = \frac{C(s_k)}{C(w)} = \frac{\text{total occurrence of } s_k}{\text{total occurrence of w}}$

- **Strengths and weaknesses of Naïve Bayes classification**

Strengths:

1) It yields high accuracy because its decision is based on choosing the sense with the highest conditional probability.

2) It can deal with longer distance or wider context as it can catch cues from topical context.

3) It is efficient because of its ability to combine evidence from a large number of features for decision making.

Weaknesses:

It ignores the structure and linear ordering of words within the context when the evidence are combined. This leads to the consideration that the presence of one word in the bag is independent of another which is often opposite to the real piece of language. Thus, it cannot exploit the powerful information from local sequence and sentence for disambiguation.

2.5.2.2 Dictionary-based disambiguation

Lesk (1986) proposes this method with the basic idea that the definition in a dictionary can be exploited for choosing the correct sense of an ambiguous word. The algorithm will explore whether in a context, there is a word form that matches with word form in the dictionary's definition. If there is, that definition will be chosen as the definition of the correct sense. For example, assume that in a dictionary, there are 2 definitions for the word *ash*

1. A Tree of the olive family.
2. The solid residue left when combustible material is burned.

The first definition suggests sense1, which is "tree" and the second suggests sense2, which is "burned stuff". If there is a sentence "This cigar burns slowly and creates a stiff ash." Sense 2 "burned stuff" will be chosen because the word form *burn* (only lemma is considered) in this sentence matches with the same word form *burn* in the second definition. While, in this sentence "The ash is one of the last trees to come into leaf." Sense 1 "tree" will be chosen because the word form *tree* in this sentence is matched with the same word form in the first definition.

- **Strengths and weaknesses of Lesk's algorithm**

Strengths:

Its basic idea is simple and easily understandable.

Weaknesses:

1) Since the algorithm is a bag-of-word model, it cannot make use of powerful local sequence for disambiguation.

2) The definition in a dictionary may not provide enough information that is word form in the dictionary's definition may not match with word form in a context of an ambiguous word. So, there will be no decision. This is why the algorithm's accuracy rate is only about 50 - 70%, which is very low (Manning and Schütze, 1999).

2.5.2.3 Information-theoretic approach

Instead of trying to use information from all words in the context window in the disambiguation decision, the information theoretic approach finds a single contextual feature that reliably indicates which sense of the ambiguous word is being used. Features considered are syntactic role relation such as object; grammatical category such as tense; co-occurrence such as word to the left.

Table 2 is an example of the highly informative indicators for three ambiguous French words. The basic idea can be illustrated by using the following example. Assume that *prendre* has 2 sense, that are "to take" and "to make" , the best indicator of the correct sense of *prendre* is its object. If its object is *mesure*, the sense "to take" will be chosen as a correct meaning of *prendre*. If its object is *décision*, the sense "to make" will be chosen.

- **Strengths and weaknesses of information theoretic approach**

Strengths:

This method avoids the independent assumption when features are combined by using only the best contextual feature.

Weaknesses:

In considering only a single contextual feature of one word in a context (which usually near or occur in the same location with an ambiguous word), the algorithm cannot efficiently deal with the indicator that is in a wider context or a context that does not have any single best indicator.

Ambiguous word	Indicator	Examples: value → sense
prendre	object	*mesure* → *to take* *décision* → *to make*
vouloir	tense	present → *to want* conditional → *to like*
cent	word to the left	*per* → *%* number → *c.* [money]

Table 2: Examples of single best features selected as sense indicators.

2.5.2.4 Decision list algorithm

Decision list algorithm proposed by Yarowsky (1994, 1994a, 1994b) is based on Rivest (1987) but narrow its complexity by restricting only to word and class trigrams. The algorithm combined the advantages of corpus-based approaches, namely decision trees, N-gram tagger and Bayesian classifier in that it can deal with both local syntactic patterns (part of speech) (which is the advantage of N-gram tagger) and more distance collocational evidence (which is the advantage of Baysian classifier). The features that are considered by the algorithm are part-of-speech, lemma (morphological roots) and word class. However, the algorithm does not combine all the features but select only one single best feature to perform WSD. For example, in French accent restoration (Yarowsky, 1994a), in context of *cote* containing *poisson*, *ports*, and *atlantique*, if the adjacent feminine article *la* is present, only this best evidence (which is *la*) is used as a single best sense indicator (indicating that *cote* means "the coast" and will be assigned the accent as *côté*) and the supporting semantic information (which are *poisson*, *ports*, and *atlantique*) are ignored. If no gender agreement constraint were present in that context (if *la* is not presented), the first matching semantic evidence would be used (which may be *poisson*, *ports* or *atlantique*).

- **Strengths and weaknesses of decision list algorithm**

Strengths:

1) It combines the advantages from corpus-based approaches so that it can make use of information from both local syntactic patterns (part of speech) (which is

the advantage of **N-gram tagger**[13]) and more distance collocational evidence (which is the advantage of Baysian classifier)

2) The algorithm is significant simplicity and ease of implementation

3) The result of the training -- a decision list -- is clearly understandable.

4) The algorithm is easily adaptable to new domains or tasks like lexical ambiguity resolution such as accent restoration, capital restoration, recovering vowels in Hebrew text, etc.

5) The algorithm achieves high accuracy (In Yarowsky 1994a, the algorithm achieve the accuracy about 96% on the average.)

Weaknesses:

The decision list algorithm when testing whether it works for fine-grained sense distinctions (such as WordNet senses (Miller et al., 1990)) is less accuracy (70% vs. 99% reported earlier) (Martinez and Agirre, 2000)

From the strengths and weaknesses of these four theoretical methods discussed above, we choose to apply Yarowsky's decision list algorithm (1994), however, without taking the advantage of the algorithm's ability to exploit information from wider context and part of speech to help disambiguation. This is because we want to know to effect of local context alone, and will use word form as the only feature for disambiguation. However, we take the advantages of performing feature selection, easily implementation, easily understandable decision list and high accuracy result.

---

[13] N-gram taggers are used to tag each word in a sentence with its correct part of speech, thus, help resolving categorial ambiguity and also ambiguity with different parts of speech.

The decision list algorithm proposed by Yarowsky's (1994) research on lexical ambiguity resolution, which is the homograph disambiguation in text-to speech synthesis, involved the following steps.

- **The Decision list algorithm**

Step1: Collect and label training data

Collect all samples of the target homographs observed in a large text corpus. Then, label each sample with its correct pronunciation in that context.

Step 2: Measure collocational distributions

Count the co-occurrences of features and the target ambiguous word. Then, measures the collocational distribution of each co-occurrence, which is the probability of the co-occurrences of features and the target ambiguous word. The features that used by Yarowsky are word form (in the form of lemmas), trigrams and (optionally) verb-object pairs.

Steps 3: Compute likelihood ratios

Measure the discriminating strength of each co-occurrence by using log-likelihood ratio:

$$\text{Weight} = \text{Abs}\left(\text{Log}\ \left(\frac{P(\text{Pronunciation}_1 \mid \text{collocation}_i)}{P(\text{Pronunciation}_2 \mid \text{collcoation}_i)}\right)\right) \quad [14]$$

The collocation patterns most strongly indicative of a particular pronunciation will have the most extreme log-likelihood ratio.

Step 4: Sort by likelihood ratio into decision lists

The weight computed will be sorted in decision list according to log likelihood ratios from the highest to the lowest weight. The collocation pattern that has the highest weight will be the most reliable indicator of the particular sense.

Step 5: Using the decision lists

The decision list algorithm is tested with new (unseen) text. During the test, the algorithm will look up for the target ambiguous words, if found, then check for the contextual feature that matches with the feature in the decision list by looking from the highest to the lowest weight. If the match found, the sense that indicated by the feature that has the highest weight would be assigned to the ambiguous words.

---

[14] Computing the ratios may arise the problem when the denominator, $P(\text{sense}_j \mid \text{feature}_k)$, is equal 0. This problem occurs when there is no such collocation probability observed in a corpus while it is clearly that it should not be so, for example, the probability of seeing *cote* in the context of *poisson* is not 0, but no such collocation is observed in a corpus (Yarowsky, 1994a). Many factors such as the size of the training sample, the noise in the training corpus lead to such problem. Many smoothing techniques are proposed for solving this problem.

Since the decision list algorithm proposed by Yarowsky was used for the lexical ambiguity resolution such as homograph disambiguation, accent restoration rather than WSD, in this study, we follow these steps with some adaptations to suit our case, which is WSD in Thai. The detailed explanation about the decision list algorithm used in this study will be presented in Chapter 3.

2.6 The Evaluation of the Performance

The last step in WSD task, like other tasks, is the evaluation of its performance in order to know degree of accuracy or achievement. The evaluation should be done against both (1) the human performance (the upper bound performance), which will be discussed in section 2.6.1 and (2) the performance of the simplest WSD algorithm (the lower bound performance), which will be discussed in section 2.6.2 below. It should be noted here that the evaluation of the algorithm's performance should consider the degree of difficulty of the task that an algorithm performs. For example, POS tagging program for English can easily achieve 90% accuracy while machine translation system nowadays can not achieve this level.

2.6.1 An upper bound performance

An upper bound performance is the disambiguation performed by a human. In case of WSD, if a human cannot disambiguate correctly, it is expected that a machine cannot too. The case that human cannot perform correctly is where there is not enough information in the context. Gale, et al. (1992a) found that in disambiguating words that have no related meanings (homonyms such as bank) the upper bound is 95% or higher whereas in disambiguating words that have highly related meanings (polysemes such as *title*, *side*, *way*) upper bound is only 65-70%.

For the evaluation against the upper bound performance in this study, since there is a lack of unified agreement among judges (Ahlswede, 1995)[15] and there is no basic research about word sense disambiguation by human informants in Thai, the algorithm will be evaluated against the disambiguation manually done by the author.

2.6.2 A lower bound performance

A lower bound performance is the performance of the simplest algorithm usually where there is strong contextual cues and dominant meaning assigned. For example, assuming that an ambiguous word occurs 1,000 times in a corpus, with 600 times of "sense1", 200 times of "sense2", and 200 times of "sense3". If choosing the most dominant meaning in all cases, the algorithm will achieve the accuracy rate of 60%.

$$\frac{\text{Number of times the sense is correctly disambiguated}}{\text{Total number of answered senses}} = \frac{600}{1000} = 60\%$$

Thus, if the performance of the optimal algorithm (span) is above this based line (if it exceeds such value) it will pass the evaluation against the lower bound performance.

[15] There is a problem in evaluating the performance against human judgements due to the lack of agreement among judges. Ahlswede (1995)'s ambiguity questionnaire reported the large gap between 63.3% and 90.2% agreement among human judges.

CHAPTER III

METHODOLOGY

This chapter consists of three main sections. Section 3.1 describes the data used in this study. Section 3.2 is word sense analysis. Both the data and the senses from the analysis are used in the WSD processes, which are explained in details in section 3.3. The details of these three main sections are as follows.

3.1 The Data

This section discusses three main concerns about the data used in this study, which are source (section 3.1.1), scope (section 3.1.2) and size of the data (section 3.1.3).

3.1.1 Source of the data

The corpus of "Bangkok Business" newspaper during November 1, 1999 to October 31, 2000 is used in this study. The corpus is kept in the form of files -- one file per one day. So, there are total 365 files for one year, which has the total size of 132 MB. The data containing the ambiguous words and their context are randomly extracted from this corpus. Only 2,200 examples of *หัว* /hua4/ and 2,200 examples of *เก็บ* /kep1/ are extracted from the corpus (see section 3.1.3).

3.1.2 Scope of the data

All occurrences of *หัว* /hua4/ and *เก็บ* /kep1/ as an individual word are the data of this study. In addition, since we would like to have all possible meanings of *หัว* /hua4/ and *เก็บ* /kep1/, we include *หัว* /hua4/ and *เก็บ* /kep1/ that immediately co-occur with other words, even though they could be viewed as compound words, reduplicative words or repetitive words. These words are included in the scope of data if the meanings of these compound words, reduplicative words or repetitive words are transparent, or in other words, the meaning of each unit does not change from its original meaning. However, some occurrences of *หัว* /hau4/ and *เก็บ* /kep1/ are not included in the scope of data. The followings are the data, which are beyond our scope.

(1) *หัว* /hua4/ and *เก็บ* /kep1/ co-occurs with other lexical units, which are

(1.1) Idiom, or idiom-like unites.

Example 1

(i) …สโมสรดังพากันรุมจีบ จน**หัวบันไดบ้านไม่แห้ง**ทีเดียว…

(ii)...เปิดเผยรายชื่อทีม รองผู้ว่าฯกทม. ให้ผู้สื่อข่าวฟังต่อที่ตีนบันไดในขณะที่บริเวณ **หัว**บันได เริ่มมีผู้มา…

In example 1, in (i), we can see that *หัว* /hua4/ co-occurs with another lexical unit, namely *บันได* /bandai/ meaning "stairs", but when consider their context and other lexical units, namely *บ้าน* /baan2/ meaning "home", *ไม่* /mai2/ meaning "not" and *แห้ง* /hɛɛŋ2/ meaning "dry", they are combined to be an idiom meaning "(a place) that always have visitors". It has opaque meaning such that each unit does not have its original meaning. Thus *หัว* /hau4/ in (i) is excluded from our data. However, in (ii), *หัว* /hua4/, which co-occurs with *บันได* /bandai/, is included in our data because the

meaning of each word is transparent. The meanings of *หัว* /hua4/ and *บันได* /bandai/ remain the same as "top" [1] and "stairs" respectively. More examples of the data excluded by this criterion are shown in appendix A.

(1.2) Compound, repetitive and reduplicative words that have opaque meaning, that is, they have the new meanings, or the meaning of each part is totally changed from its original meaning.

Example 2

(i) …อบายมุขมอมเมาเยาวชนผู้ใหญ่และผู้ชรา**หัวอสรพิษ**ที่มักนิยมบริโภคหญ้าอ่อน…

(ii) …พระนางพรหมจารี มารีย์ถึงแม้จะเหยียบ**หัว**อสรพิษไว้ใต้เท้าแล้ว พระแม่ยัง…

In example 2, in (i), we can see that *หัว* /hua4/ co-occurs with *อสรพิษ* /ʔa1sɔɔ 4ra3pit3/ meaning "snake", but from the context, *หัวอสรพิษ* is excluded from our data, because it is a compound word meaning "an old womanizer". However, in (ii) *หัว* /hua4/ co-occurs with *อสรพิษ* /ʔa1sɔɔ4ra3pit3/ meaning "snake", from the context, is included in the data because the meanings of *หัวอสรพิษ* is the combination of the meanings of *หัว* /hua4/ - "head" and *อสรพิษ* /ʔa1sɔɔ4ra3pit3/ - "snake". More examples of the data excluded by this criterion are shown in appendix A.

(1.3) Proper names.

Example 3

(i) …มีน้ำป่าจากเขาพุราง เขา**หัวล้าน** และเขาตอง ได้ไหลบ่าเข้าท่วมพื้นที่ทำการเกษตร…

(ii) …ถ้าเกิดผมร่วงเป็นกระจุกแล้วหัวล้านขึ้นมา ก็เลยปรึกษากับ…

---

[1] This meaning of *หัว* /hua4/ is extended from its original meaning of "head".

In example 3, in (i), *หัว* /hua4/ co-occurs with *ล้าน* /laan3/ meaning "bald", but in this context, *หัวล้าน* is a single word because it is the name of a mountain. Thus, it is not included in the data. However, in (ii), *หัว* /hua4/ co-occurs with *ล้าน* /laan3/ meaning "bald" is included in the data as it has transparent meaning. The meaning of *หัวล้าน* is the combination of the meanings of *หัว* /hua4/ - "head" and *ล้าน* /laan3/ - "bald". More examples of this type of data are shown in appendix A.

(2) *หัว* /hua4/ which has parts of speech other than noun and *เก็บ* /kep1/ which has parts of speech other than verb will be excluded. For example, *หัว* /hua4/ in (i) and (ii) are excluded because they are verb and *เก็บ* /kep1/ in (iii) is excluded because it is adjective.

(i) …หนำซ้ำอาจจะแอบยิ้ม**หัว**อยู่ในใจว่าเดี๋ยวก็รู้เมื่อไทยเตรียมเสิร์ฟเมนูอาหารชามเก่ง

(ii) …เมื่อสามพันปีมาแล้วหรือเรื่องชวน**หัว**ของการพัฒนาในภาคอีสาน

(iii) …อะไรไว้ ไม่มีทรัพย์สิน มรดก มีเงิน**เก็บ**นิดๆหน่อยๆ ก็หมดไปกับการ…

We would like to note here some advantages of including compound, repetitive and reduplicative words with transparent meaning in our data. First, by doing this, we would have a large variety of meanings of *หัว* /hua4/ and *เก็บ* /kep1/, and have a large number of data for testing collocational words in this study. Second, there will be smaller lexical units in the lexicon. For example, instead of having at least 7 lexical units (*หัวอ่าน, หัวฟัง, หัวลาก, หัว, อ่าน, ฟัง and ลาก*), we have only 4 units (*หัว, อ่าน, ฟัง,* and *ลาก*). However, there is also a disadvantage that some of the senses are not applicable to some tasks such as human or machine translation. For example, *หัวลาก* would better be considered as one word and translated as "trailer" than considered as two words with *หัว* /hua4/ means "machine part" and *ลาก* /laak2/ means "to trail" and *หัวลาก* is translated as "part which trail the rest of a machine".

3.1.3 Size of the data

As we define our scope to be at least 1,000 samples of each word, we collected 600 samples of *หัว* /hua4/ and 600 samples of *เก็บ* /kep1/ as our training data and 400 samples of *หัว* /hau4/ and 400 samples of *เก็บ* /kep1/ as our testing data. Then, we carried out the pilot study on these samples. But the precision rate is not so high. Thus, we collected additional 600 samples of *หัว* /hua4/ and 600 samples of *เก็บ* /kep1/ as our training data. Then, we tested at these sample sizes (1,200 samples for each word). The precision rate is significantly increase. However, we would like to know whether the precision rate would significantly increase with the increasing in training sample size. Thus, we collected additional 600 training samples and got 1,800 samples of *หัว* /hua4/ and 1,800 samples of *เก็บ* /kept1/. Then we tested at this sample size and found that the precision rate is increased but not significantly (see the precision rates of testing at 600, 1,200, and 1,800 training samples in figure 35 and figure 36, section 5.1.1). Thus the training data in this study is set as 1,800 samples.

The number of token classified by senses of the ambiguous words is shown in table 3 for *หัว* /hua4/ and table 4 for *เก็บ* /kep1/.

In the training corpus of *หัว* /hua4/, most of the senses [2] found are "head", "entity", "viewpoint", "bulb" and "brain" respectively. In the training corpus of *เก็บ* /kep1/, most of the senses found are "to keep", "to charge", "to take", "to gather" and "to hide" respectively.

---

[2] See the definitions of these senses and all other senses of*หัว* /hua4/ and *เก็บ* /kep1/ in section 3.2

Sense of หัว /hua4/	No. of training data	No. of testing data
Head	410	96
Entity	388	72
Viewpoint	202	36
Bulb	145	14
Brain	109	29
Front	102	31
Intelligence	72	16
Top	66	11
Titles or names	43	13
Concentrate	42	13
Topics	42	18
Machine part	36	14
Headline	33	8
Hair	32	5
Early hours	30	11
Chief	25	5
Emotion	9	4
Heading	6	1
Talent	4	1
Head of coin	4	2
Total	1800	400

Table 3: Sizes of the training and testing data of หัว /hua4/ classified by senses and sorted by the number of occurrences of each sense in a training data from the highest to the lowest.

Sense of เก็บ /kep1/	No. of training data	No. of testing data
To keep	672	160
To charge	520	107
To take	276	46
To gather	233	62
To hide	52	9
To arrange	35	6
To buy	15	5
To kill	7	3
To pick up	5	2
Total	1800	400

Table 4: Sizes of the training and testing data of เก็บ /kep1/ classified by senses and sorted by the number of occurrences of each sense in a training data from the highest to the lowest

3.2 Word Sense Analysis

The senses used for tagging the training corpus in this study comes from two analyses, namely the analysis of word sense based on Thai dictionary of "The Royal Institute" (section 3.2.1) and the analysis of additional word senses based on the training corpus (section 3.2.2). The reason that we have to analyze additional senses from the training corpus is because the senses provided by the Thai dictionary of "The Royal Institute" may not fit in some contexts of a training corpus -- some senses may not occur anywhere in the data, or the senses provided may not suitable to some context. Thus, it is necessary to analyze all the senses of หัว /hua4/ and เก็บ /kep1/ before tagging. Figure 1 and figure 2 are flow charts illustrating word sense analysis of หัว /hua4/ and เก็บ /kep1/ respectively.

The steps involved in word sense analysis are as follows.

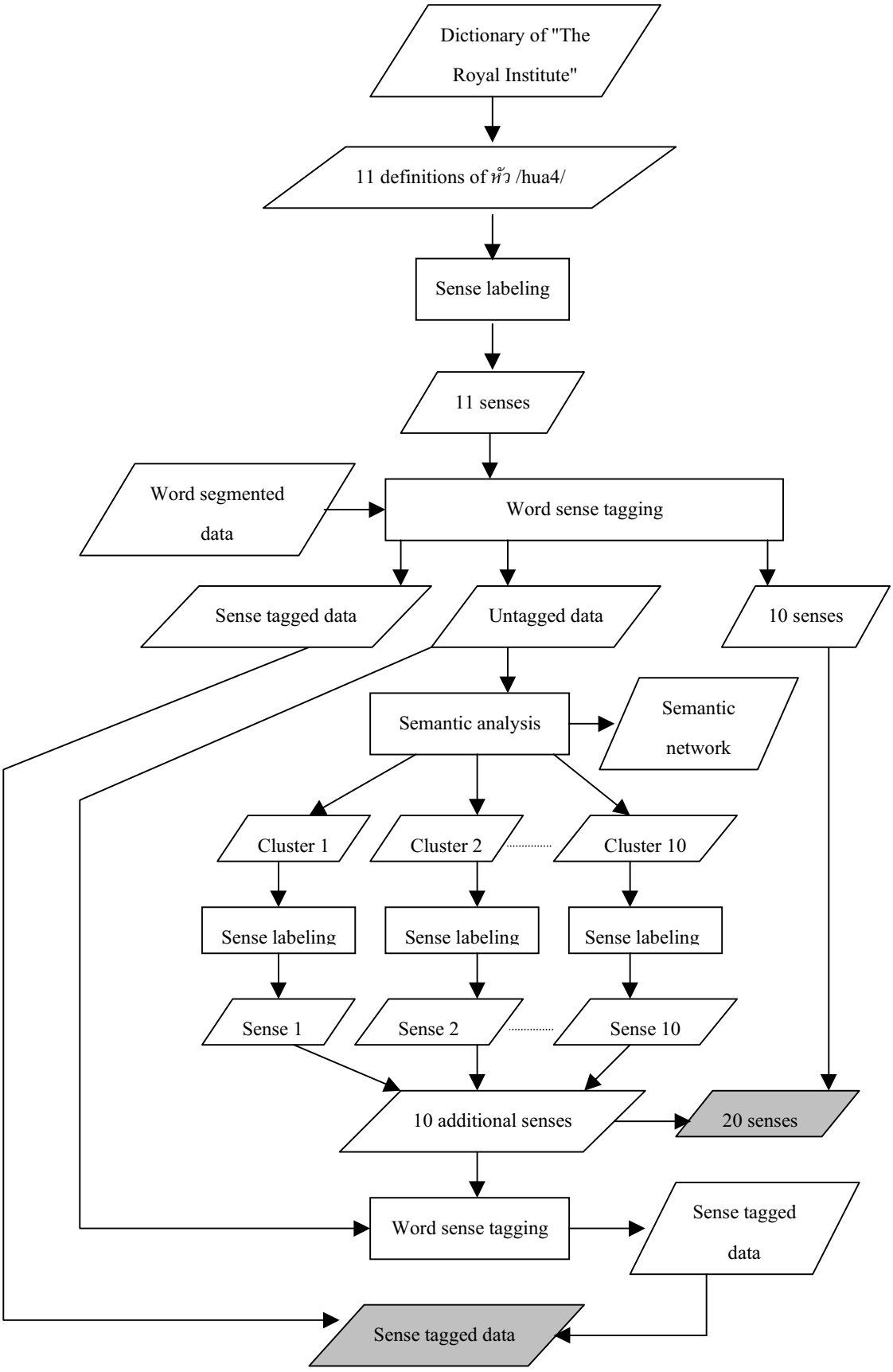


Figure 1: Word sense analysis of หัว /hua4/.

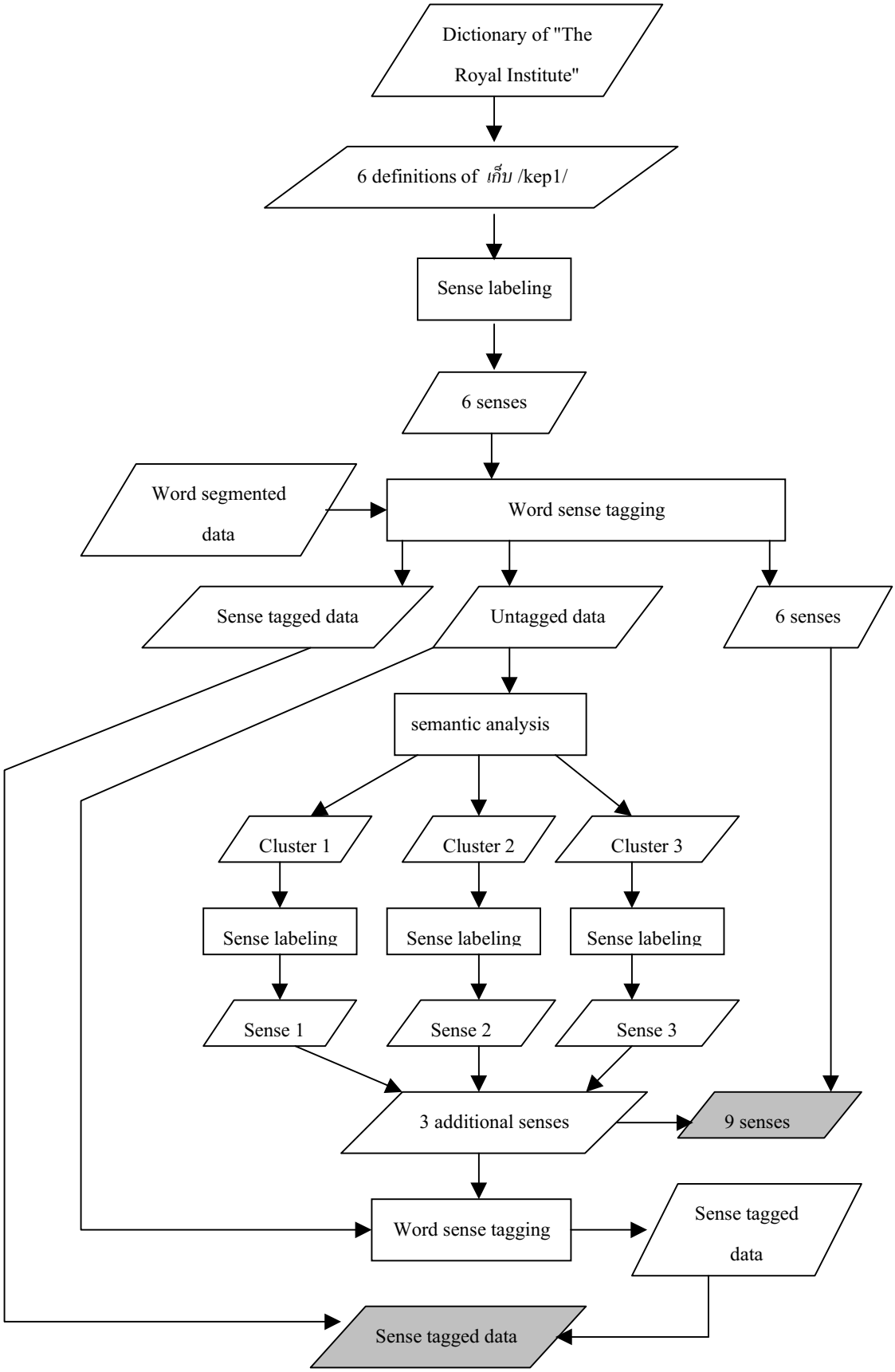


Figure 2: Word sense analysis of เก็บ /kep1/.

3.2.1 Analysis of word sense based on Thai dictionary of "The Royal Institute"

Step I: Extract the definitions of *หัว* /hua4/ and *เก็บ* /kep1/ as stated in the dictionary. There are eleven definitions of *หัว* /hua4/ and six definitions of *เก็บ* /kep1/. Then derives the senses based on these definitions. Table 5 and table 6 contain the definitions, derived senses and examples of *หัว* /hua4/ and *เก็บ* /kep1/ respectively.

SENSE	DEFINITION	EXAMPLE
Head	Body part which contains the brain. Also used as a classifier.	…กลุ่มชายฉกรรจ์ใช้ปืนจี้**หัว**และนำตัวส่งให้กับตำรวจ… …และประดับประดาด้วย**หัว**ปลานั้นน่าจะเป็นที่ถูกอกถูกใจของนักเขียน… …เสริมด้วยท่าทางจากการขยับมือ **หัว** และเท้า และ เป็นที่แน่นอนด้วยว่า… …5 ชิ้นขึ้นราคาจะตกอยู่ที่ **หัว**ละ 11 บาท **หัว**สัตว์ ขนาดใหญ่ 170 บาท…
Head of coin	Side of coin where a person's profile is represented, opposite of tails.	…อย่างไรก็ตามเมื่อมีด้าน**หัว**ของเหรียญแล้วด้านก้อยก็ต้องมีเช่นกัน… …เปรียบเสมือนเหรียญที่มีทั้งด้าน**หัว**และก้อย แต่เป็นเหรียญเดียวกัน…
Intelligence	Ability of a person's brain; intelligence	…บุช ในช่วงวัยเรียน เขาเป็นเด็กที่**หัว**ปานกลาง ในช่วงวัยรุ่นเขา… …ตัวจะดีก็ขึ้นอยู่กับ**หัว**ครับ**หัว**ดีคิดสะอาดก็แล้วไป **หัว**ที่มคิดไม่เป็นก็เตรียมใจกัน … …หลังอย.ระบุมีพ่อค้าเร่**หัว**ใส นำยาออกเร่ขาย… …ผู้ฝึกฝนชี่กงจะมีโอกาสได้มากกว่าคนทั่วไป จะมี **หัว**ไวขึ้นคิดเก่งขึ้น…

Table 5: Definitions, derived senses of *หัว* /hua4/ provided by the dictionary and examples of these senses found in a training corpus.

SENSE	DEFINITION	EXAMPLE
View point	The way a person views or thinks about an issue.	…ศัพท์ใหม่นี้ เขาเป็นพวก**หัว**รุนแรง คนหนึ่งซึ่งมีแนวโน้มจะ… …แม้ในขณะนี้กลุ่มนักเล่น**หัว**อนุรักษ์จะยังคงเอ่ยปากปฏิเสธ… …อาจเพราะผมเป็นคนประเภท**หัว**โบราณ ที่นิยมความร่วมสมัยอยู่… …ทุกวันนี้ ต้องการคนที่มี**หัว**สมัยใหม่ในการจัดการ นอกจากนี้…
Talent	Talent or special ability to do something.	…กล่าวถึงศิษย์ผู้นี้ว่า เขาเป็นคนที่มี**หัว**การค้ามาตั้งแต่เด็ก เพราะตลอดระยะเวลา… …อาจจะฉลาดแกมโกงหรือว่ามี**หัว**ทางธุรกิจก็ตาม แต่ ตั้งแต่ครั้งที่เขาคุมทีมท็อตแนม… …ธุรกิจครอบครัว มีไม่พอให้ทำสมาชิกรุ่นเยาว์ที่มี**หัว**เป็นผู้ประกอบการที่มีประสบการณ์ในเรื่อง…
Top	Top part or pointed end of an object.	…โดยสังเกตด้วยวิธีการง่ายๆ ให้ใช้**หัว**ไม้ขีดไฟใส่ลงไปในร่อง… …มองเห็นดอกดวงของ**หัว**ฝี ที่เริ่มแตกบนใบหน้าของเขาเปล่งปลั่ง… …เข้าปอด แม้แพทย์จะผ่าตัดเอา**หัว**กระสุนออกแล้ว แต่ไม่สามารถช่วยชีวิต… …ผิดปกติ เลือดหรือน้ำเหลืองบริเวณ**หัว**นมเพื่อเพิ่มความมั่นใจ…
Front	Front or pointing part of an object.	…ที่ออกแบบเป็นสายรัดและ**หัว**เข็มขัดเข้าชุดกับกระเป๋าถือ… …ออกจากบ้าน ตรวจตราสับปะรดตั้งแต่**หัว**ไร่จนถึงท้ายไร่เขาเห็นหัวหนึ่ง… …โดยผู้ผลิตรถยนต์บางรายจะสร้างรถยนต์สูตรหนึ่งตั้งแต่**หัว**จรดท้าย จากนั้นจึงขายพื้นที่… …ลักษณะคล้ายรูปเรือบินหัน**หัว**ไปทางดอนเมือง และโครงการปรับปรุงภูมิทัศน์คลอง…

Table 5: Definitions, derived senses of *หัว* /hua4/ provided by the dictionary and examples of these senses found in a training corpus.

SENSE	DEFINITION	EXAMPLE
Early hours	The early hours or part of a time period.	…ทยอยไปแสดงความยินดีกันอย่างมืดฟ้ามัวดินตั้งแต่**หัว**ค่ำเรื่อยไปจนถึงเที่ยงคืน… …เป็นการเลี้ยงอำลาตั้งแต่ยัง**หัว**วัน ทั้งๆที่ยังอยู่ในตำแหน่งนี้อีกร่วมปี… …แต่ไอ้แจ้ก็ต้องตื่นแต่**หัว**รุ่งด้วยความประหลาดใจลงจากคอน…
Bulb	Globular base of stem of some plants sending roots downward and leaves upwards; Bulb. Also used as a classifier.	…ด้านการตลาดขณะนี้มีปริมาณ**หัว**มันสดออกสู่ตลาดเป็น… …ในการผลิตเพิ่มขึ้น เนื่องจากหอมจะลง**หัว**ช้าเพราะต้องหาอาหารไปเลี้ยงดอกหอม… …ของพวกนั้นเป็นอย่างยิ่ง หารู้ไม่ว่าพืช**หัว**เล็กๆนี้อุดมไปด้วยคุณค่ามากมาย… …เที่ยวเดินพลิก**หัว**สัปปะรดเลือกเอาแต่**หัว**ที่ยังไม่เน่าแม้ไม่มีที่จะ…
Concentrate	Concentrated substance, usually to be diluted when used.	…อาทิการกันค่าใช้จ่ายเพื่อใช้ในการลดจำนวนสต็อก**หัว**เชื้อน้ำอัดลมของบริษัทบรรจุน้ำอัดลม… …ปัจจัยการผลิตต่างๆ เช่น**หัว**พันธุ์ปุ๋ย รวมทั้งจัดหาพื้นที่ให้ด้วย… …มีพืชผักกินน้ำพริกผักจิ้มราด**หัว**กะทิก็ได้แร่ธาตุวิตามินจากผัก… …นิยมมากกว่าการปรุง**หัว**น้ำซุปที่ได้จากเนื้อหมูแดง…
Head of Thai alphabet	The small circle marking the beginning of an alphabet in Thai script.	There is no example because this sense occurs nowhere in the corpus.

Table 5: Definitions, derived senses of *หัว* /hua4/ provided by the dictionary and examples of these senses found in a training corpus.

SENSE	DEFINITION	EXAMPLE
To pick up	To pick something up from the ground or the floor.	…กุญแจหลุดจากมือหล่อนรีบก้มลง**เก็บ** แต่เสียหลักล้มลงก้นจ้ำเบ้า… …คนโชคดีมากคนหนึ่ง ตั้งแต่อายุ 3 ขวบ เดินตาม**เก็บ**ลูกมะม่วงที่หล่นจากต้น… …ตามสายน้ำ และเขาง่วนอยู่กับการ**เก็บ**ฟืนขึ้น จึงไม่แน่ใจว่าเป็นสัตว์หรือท่อนไม้… …ความเรียบร้อยของสถานที่ อีเลียสรีบก้มลง**เก็บ** ท่านพัศดีเอ่ยถามทุกสิ่งทุกอย่างพร้อมมั้ย"…
To arrange	To put away; to arrange objects in a cabinet, a closet or a box.	…สลัดผ้าห่มทำตัวให้เหมือนปกติ จัดแจง**เก็บ**ที่นอนเรียบร้อย แล้วเดินออกมาข้างนอก… …ซิดนีย์ เกมส์ หนุ่มสาวยูโดเตรียม**เก็บ**กระเป๋าบุกแดนปลาดิบ ลุยศึกชิงแชมป์… …โคมทำด้วยกระดาษพับ**เก็บ**ไว้ได้ แต่ละปีก็คลี่นำออกมาใช้ใหม่… …นายเพี้ยนตะโกนไล่หลังเพื่อน เมื่อเห็นฝ่ายแรกรีบ**เก็บ**แผงหนังสือ เข็นเข้าบ้าน…
To take	To collect; to harvest, to take under one's care.	…Home for dying ในตอนแรก แม่ชีได้**เก็บ**ผู้หญิงที่กำลังนอนรอความตายอยู่ข้างถนนเข้ามา… …ไว้เพื่อประดับบ้านมากกว่า ที่จะได้**เก็บ**ฝรั่งบ้างก็ในตอนที่มีญาติหรือเพื่อน… …มีการระบายน้ำที่ดี การ**เก็บ**เกี่ยวให้**เก็บ**ตอนใบมันเหี่ยวและแห้งตาย ปัจจุบัน…
To keep	To keep or to store, to prevent loss or damage.	…เพราะเป็นสิ่งมีค่าหลาย ไม่**เก็บ**อาจจะ หายใครขวนขวายหาให้เรา… …แน่นอนแล้วการรับซื้อจะนำ**เก็บ**ในโกดังที่จะต้องมีวิธีการ**เก็บ**เนื้อมะพร้าวแห้งที่ดีได้… …ลูกชิ้นที่ขายไม่หมด ถ้า**เก็บ**ในตู้เย็นจะ**เก็บ**ได้นานถึง 2 อาทิตย์… … ใส่ถัง 200 ลิตรแล้วนำไป**เก็บ**ไว้ในที่ปลอดภัยห่างไกลจากชุมชน…

Table 6: Definitions, derived senses of *เก็บ* /kep1/ provided by the dictionary and examples of these senses found in a training corpus.

SENSE	DEFINITION	EXAMPLE
To gather	To gather; to save	…ละเอียดข้อเท็จจริงบางอย่างที่การ**เก็บ**ข้อมูล ของคนในวงการวรรณกรรมไม่สามารถทำได้… …ได้จากปั๊มหลอดขนาดเล็กแห่งหนึ่ง และได้**เก็บ**ตัวอย่างส่งให้ทางกรมสรรพากร… …ผมตอบไปว่าขยันทำงานและ**เก็บ**เงินให้ได้มากพอที่จะเดินทางกลับไป… … ผู้หญิงและผู้ชายบางคนก็ชอบ**เก็บ**สะสม บางคนชอบการรีไซเคิลและบางคนเรียกตัวเองว่า…
To charge	To collect or to charge a fee	…เทศบาลเมืองเลยเตรียม**เก็บ**เงินค่าทิ้งขยะมูลฝอยจากเทศบาลรอบๆที่เอาขยะมาทิ้ง… …ที่จะจ่ายค่าธรรมเนียม ทางเวบไซต์จึงเปลี่ยนจากการ**เก็บ**ค่าธรรมเนียมมาเป็นการขอ… …ต่อให้เปิดบ่อนเสรีแทงกันตามใจชอบ **เก็บ**ภาษีเข้ารัฐด้วยก็เหอะ… …และเมื่อใช้ 400,000 นาที **เก็บ**ค่าบริการ 0.15 บาทต่อนาที…

Table 6: Definitions, derived senses of *เก็บ* /kep1/ provided by the dictionary and examples of these senses found in a training corpus.

Step II: Tag the corpus using the senses provided by the Thai dictionary of "The Royal Institute" by putting these senses to their suitable contexts. During the tagging process, we found that the senses derived from the dictionary are not applicable to all occurrences in the data. There is one sense of *หัว* /hua4/, namely "head of Thai alphabet", that is not found in a corpus. Since this study is a corpus-based WSD, we discard this sense of *หัว* /hua4/ as it is not found in our training data. Besides, in some contexts, no senses of *หัว* /hua4/ and *เก็บ* /kep1/ provided by the dictionary are suitable, thus we have to analyze some additional senses that fit these contexts.

3.2.2 Analysis of additional word senses based on the training corpus

The analysis of the additional senses of *หัว* /hua4/ and *เก็บ* /kep1/ are as follows.

Step I: List all the data that cannot be tagged with the senses provided by the dictionary.

Step II: Categorize them into clusters according to the context in which they occur. We get ten clusters for *หัว* /hua4/ and three clusters for *เก็บ* /kep1/. The examples of the context of *หัว* /hua4/ that fall into the same clusters are as follows.

Cluster 1:

มี ตัว กรอง ที่ ปั๊ม ก่อน เข้า **หัว** จ่าย ไส้กรอง น้ำมัน เบนซิน ใน รถ คุณ ก็ จะ อยู่
ปอนด์ ที่ มี ประกอบด้วย สาย คล้อง ไหล่ **หัว** ฟัง และ รีโมท ได้รับ เสียง
เป็น ระบบป้องกัน การ ลักลอบ จูน และ **หัว** อ่าน บัตร ใน เครื่อง ออก
บริษัท ที่ รพ. สั่งซื้อ ไม่ ตกลง จะ นำ **หัว** เก่า เครื่อง กำเนิด รังสี ไป เก็บ
ได้ มี การ นำ โคบอลต์ 60 **หัว** ใหม่ มา แทน โดย มี ข้อ สัญญา ระบุ ว่า
การ บริการ ขนส่ง ทาง เรือ ธุรกิจ **หัว** ลาก และ แอร์ คาโก้ รวมถึง การ ร่วม ทุน

In this cluster, we find that this sense of *หัว* /hua4/ often co-occurs with some action or operational verbs such as *อ่าน* /ʔaan1/ meaning "read", *ฟัง* /faŋ/ , meaning "to listen, to hear", *จ่าย* /caai1/ meaning "to distribute" and *ลาก* /laak2/ meaning "to trail". We substituted this sense co-occurred with these contexts with sense "machine part", which is the suitable sense for these context.

Cluster 2:

มัน เริ่ม ทำให้ ผม อารมณ์ พลุ่ง พล่าน **หัว** เสีย มากขึ้น ผม อยากจะ ตะโกน ด่า

ชน และ ข้าราชการ ไป แล้ว ต่าง พา กัน **หัว** เสีย กัน เป็น แถว เพราะ ไม่ คุ้มค่า ซึ่ง เป็น วันขึ้นปีใหม่ ของ จีน เจ้าคุณ ปัจจนึก **หัว** เสีย เพราะ ไม่ มี ขนมปัง รับประทาน

In this cluster, we find that *หัว* /hua4/ always co-occurs with เสีย /siia4/ meaning "bad". We substituted this sense co-occurred with these contexts with a suitable sense for these context, which is "emotion".

Cluster 3:

ส.จ. เมือง ย่าโม ร้อน แข่ง ดุ **หัว** ละ 500 ขณะที่ การ เลือก ส.จ.
หนี้ ประเทศชาติ เมื่อ คิด เป็น ต่อ **หัว** ก็ เฉลี่ย ออกมา เท่ากับ คน รวย
มี ราคา เฉลี่ย ต่อ **หัว** ตั้งแต่ 600 - 700 บาท ขึ้นไป ส่วนใหญ่ เป็น ทัวร์
ประมาณ แบบ จ่าย ตาม ราย **หัว** ประชากร ภายหลังจาก เริ่ม ปฏิรูป ผู้สมัคร ส.ว. กับ ลาว
โดย ซีเรีย นั้น จะ เก็บ **หัว** ละ 2 เหรียญ ต่อ วัน ส่วน ลาว จะ คิด แค่ 10

In this cluster, *หัว* /hua4/ usually co-occurs with the amount of money or the number. Thus we substituted this sense of *หัว* /hua4/ as "entity".

Cluster 4:

นอกจาก กลุ่ม ยังเติร์ก แล้ว ใน ระดับ **หัว** มี เพียง พลเอก จิ๋ว คน เดียว ที่
โดย มี นาย นำชัย กฤษณาสกุล **หัว** ทีม คณะ สร้างสรรค์ สตูล และ หัวหน้า ฝ่ายค้าน
ของ เจ้าหน้าที่ตำรวจ ว่า เขา คือ **หัว** โจก ใน การ ถล่ม คุก เมือง หมอ

In this cluster, *หัว* /hua4/ usually co-occurs with the word *โจก* /cook1/ "leader". When substituting this sense occurred in these contexts with the sense "chief", we got the suitable sense for this cluster.

Cluster 5:

นี้ ต่อไป เธอ จะ ต้อง โกน **หัว** ปฏิบัติธรรม อยู่ ใน วัด นี้ ผม ต้อง
ขี้เกียจ ซ้อม จน ถึงกับ ต้อง โกน **หัว** ประชด ตัวเอง เส้นทาง ของ สมรักษ์ ค่อนข้าง
ยุโรป ที่ เป็น ของ พวก ฝรั่ง **หัว** แดง ผิว ขาว คน เอเชีย คือ พลเมือง ชั้นสอง ไป
ไม่ ใช่ เด็ก ใส่ ชุด ไทย **หัว** จุก ร้อย มาลัย ที่ ตาม ป้าย โปสเตอร์ รณรงค์

In this cluster, *หัว* /hua4/ usually co-occurs with *โกน* /koon/ "to shave" and other color terms such as *ดำ* /dam/ "black". When substituting the sense in these contexts with "hair", we got the suitable sense for this cluster.

Cluster 6:

เสียง เป็น สื่อ ที่ แปร ความ คิด จาก **หัว** สู่ ร่าง ก็ ขอให้ ช่วยกัน ถกเถียง หา
พูด พร่ำ ทำ เพลง ขอ เพียง มี **หัว** เป็น ไอเดีย มี ปาก เป็น ภาษาอังกฤษ มี
นี้ ว่า คณะกรรมการ บริษัท คือ ส่วน **หัว** หรือ มันสมอง ของ บริษัท ที่ จะ ชี้ ทิศ ให้
จน เขา เครียด ช็อค และ เกร็ง ปวด **หัว** ข้างเดียว แต่ ตอนนี้ หลังจาก มี ลูก เขา ดี
กำเริบ ด้วย จึง จะ ทำให้ อาการ ปวด **หัว** จาก ไมเกรน ดี ขึ้น ได้ ครับ

In this cluster, *หัว* /hua4/ usually co-occurs with *สมอง* /sa1mɔɔŋ4/ "brain", *คิด* /kit3/ "idea" and some feeling verbs like *ปวด* /puuat1/ "to be ached". The suitable sense for this cluster is “brain”.

Cluster 7:

วันที่ 24 กุมภาพันธ์ นี้ เอง ขณะที่ **หัว** หนังสือ ลง วันที่ ไว้ ว่า เป็น วันที่ 14
นำ แฟกซ์ มา ยืนยัน และ ตรวจสอบ ว่า **หัว** กระดาษ แฟกซ์ ส่ง จาก หมายเลข อะไร
ส่ง มา เมื่อ คืน นั้น **หัว** กระดาษ ระบุ ชื่อ โรงเรียน ลง เดือน และ พ.ศ. เรียน ผู้อำนวยการ

In this cluster, *หัว* /hua4/ usually co-occurs with some printed matters like *กระดาษ* /kra1daat1/ "paper", and some printed materials like *หนังสือ* /nan4sɯɯ 4/ "book". The sense "heading" then is the suitable sense in these contexts.

Cluster 8:

นสพ. แทบ ทุก ฉบับ พา กัน พาด **หัว** ตัว โต กัน เกรียวกราว เพราะ ไม่
เยาวชน ทั้งหมด ตาม ข่าว ที่ พาด **หัว** ตาม หน้า หนังสือพิมพ์ หนู เชื่อ ว่า
สเปน 3 เยอรมนี 1 เป็น การ พาด **หัว** โคตร กวน Teen ใคร ที่ ไม่ ได้ ติด
หนังสือพิมพ์ แนว ธุรกิจ และ เศรษฐกิจ พาด **หัว** ไป ทิศทาง เดียวกัน ว่า สิ้น

In this cluster, *หัว* /hua4/ usually co-occurs with *พาด* /paat2/ "to headline". Thus, we substituted the sense of *หัว* /hua4/ in these contexts as "headline".

Cluster 9:

ถึง 556 บทความ และ ยัง ยกตัวอย่าง พาด **หัว** ของ บทความ เกี่ยวกับ โทษ ของ
เข้ามา เปิด อ่าน โดย เมื่อ คลิก ยัง **หัว** เรื่อง ใดๆ ที่ ต้องการ ทาง หน้าต่าง ด้าน
มิ หน่ำซ้ำ แทบจะ ทุก เล่ม ก็ โปรย **หัว** โปรย ท้าย อวดอ้าง กัน มากมาย ว่า
มา เข้า เรื่อง ตามที่ จั่ว **หัว** กัน บ้าง คำถาม ที่ เกิดขึ้น ใน ใจ ของ ผู้คน หลาย

In this cluster, *หัว* /hua4/ usually co-occurs with *จั่ว* /cuua1/ "to introduce" *โปรย* /plooy/ "to introduce" *เรื่อง* /rɯɯan2/ "subject". When substituting the sense occurred in these contexts with the sense "topics", we got the suitable sense for these contexts.

Cluster 10:

แม้ จะ มี การ ฟอร์ม หนังสือ **หัว** เดียวกัน ขึ้นมา ใหม่ ช่วง สั้นๆ แต่ ก็ เป็น การ ดู

พลิก ดู นิตยสาร ลิซา แม้ จะ เป็น **หัว** นอก แต่ เมื่อ เข้ามา ใน ประเทศไทย แล้ว นิตยสาร น้องใหม่ แถม ยัง **หัว** นอก อีกด้วย อย่างไร ก็ แจ้ง เกิด ใน ประเทศไทย มี อักษร L HUAN HUA BAO เดา ว่า เป็น **หัว** หนังสือ แต่ อ่าน ว่า กระไร ไม่

In this cluster, *หัว* /hua4/ usually co-occurs with printed materials like *หนังสือ* /naŋ4sɯɯ4/ "book",*นิตยสาร* /nit4ta1ya1saan4/ "magazine",*หนังสือพิมพ์* /naŋ 4sɯɯ4pim/ "newspaper". Thus, we analyze this sense as "titles or names".

The followings are examples of *เก็บ* /kep1/ that do not have the same meanings as listed in the dictionary. They can be grouped into three clusters as follows.

Cluster 1:

ยโส หยิ่ง ทะนง ใน ตนเอง ค่อนข้าง **เก็บ** ตัว และ ห่างเหิน จาก คน อื่นๆ เมื่อใด ที่ ไม่ กล้า ทำความเข้าใจ ใน เรื่อง เพศ ศึกษา **เก็บ** กด ความ คิด และ ความ รู้สึก บางครั้ง คน พวก นี้ ต้อง **เก็บ** ตัวตน ไว้ ลึกๆ ข้าง ใน แล้ว ยอม แลก ศักดิ์ศรี กับ จะ ไม่ กล้า พูด กล้า แสดง ออก มาก **เก็บ** งำ ไว้ ใน ใจ จะ เป็น คน ที่ รัก และ นาย ธงชัย ได้ ลาออก ไป แล้ว แต่ เขา ก็ ยัง **เก็บ** ตัว เงียบ ไม่ เปิดเผย หรือ ติดต่อ ควบคุมดูแล มี ปัญหา เลย เกิด ความ **เก็บ** กด เมื่อ มี การ ใช้งาน อย่าง สุด มี อะไร โลดโผน หวือหวา หรือ **เก็บ** กด คน ใกล้ชิด แหม่ม ย่อม รู้ดี ดังนั้น

In this cluster, *เก็บ* /kep1/ usually co-occurs with verb such as *กด* /kot1/ "to suppress", *งำ* /ŋam/ "to hide", and *ตัว* /tuua/ "person, individual". When substituting the sense in these contexts with the sense "to hide", we got the suitable sense of *เก็บ* /kep1/ for these contexts.

Cluster 2:

เขา บอก ว่า ถ้า ผม ถูก **เก็บ** ไป จริงๆ เงิน นี้ ก็ จะ เป็น กำไร ของ ประเทศ

ตัว เต็ง คว้า ที่นั่ง อบต. ห้วยไผ่ ถูก **เก็บ** อีก 1 ขณะที่ บางน้ำเปรี้ยว สุด พิลึก มี
แล้ว ทั้ง งาน ใหญ่ และ งาน เล็ก งาน **เก็บ** นักการเมือง ระดับประเทศ และ นัก
เกิดขึ้น เพราะ ความ ต้องการ ที่ จะ **เก็บ** คู่แข่ง ทาง การเมือง จริงๆ เพราะ บางครั้ง รุน
แรง ส่ง มือปืน เข้าไป **เก็บ** เสีย ซึ่ง ก็ เกิดขึ้น มาแล้ว หลาย จังหวัด
เมย์ ภัทรวรินทร์ เพื่อน ร่วม ทีม ก็ ถูก สั่ง **เก็บ** ฉาก นี้ ทั้ง คู่ ต้อง วิ่งหนี การ ไล่ล่า
การ ซื้อ เสียง ทำ ได้ ยาก ขึ้น แต่ การ **เก็บ** คู่แข่ง ทำ ได้ ง่าย กว่า ไม่เพียงแต่ ศึก

In this cluster, *เก็บ* /kep1/ usually co-occurs with *ถูก* /thuuk1/ "-ed (passive voice)", *สั่ง* /saŋ1/ "to order", *คู่แข่ง* /khuu2khɛɛŋ1/ "competitors", and *นักการเมือง* /nak3kaanmɯɯaŋ/ "politician". When substituting the sense in these contexts with "to kill", we got the suitable sense for these contexts.

Cluster 3:

ปิด ที่ ระดับ 19.2 บาท และ มี การ ไล่ **เก็บ** หุ้น ตัว นี้ จน ดัน ราคา หุ้น ขึ้น ไป
กัน มาก ที่สุด โดย มี การ ไล่ **เก็บ** มา ตั้งแต่ วัน ที่ 13 ธันวาคม 25 จาก การ
มากกว่า 1 พันล้าน บาท และ มา ไล่ **เก็บ** อย่าง หนัก อีกครั้ง ในช่วง ราคา 12 -
ราย ใหญ่ ต่าง ชิง กัน เข้ามา **เก็บ** หุ้น ใน กลุ่ม นี้ กัน อย่าง หนาแน่น เกือบ ตลอด ช่วง
แต่ ในขณะนี้ บริษัท ก็ ยัง ได้ **เก็บ** หุ้น ที่ จะ ขาย ให้กับ กลุ่ม ทุน บรูไน อยู่ แต่
โดยเฉพาะ กองทุน ต่าง ทยอย **เก็บ** หุ้น ใน ระดับ ราคา ดังกล่าว รวมถึง
นักลงทุน อเมริกัน เดิน หน้า ไล่ **เก็บ** หุ้น ไฮเทค เก็งกำไร ขานรับ ข่าว การ เคลื่อน

In this cluster, *เก็บ* /kep1/ usually co-occurs with *ไล่* /lai2/ and *หุ้น* /hun2/ "stock". The sense "to buy" is the suitable sense of *เก็บ* /kep1/ in these contexts.

Finally, we got ten additional senses of *หัว* /hua4/ and three additional senses of *เก็บ* /kep1/ as shown in Table 7 and table 8 respectively.

SENSE	DEFINITION	EXAMPLE
Entity	Metonyms use of "head" to refer to an individual; extended metaphorically to refer to an organization; and used also as a classifier.	…มีการเสนอเงินซื้อเสียงในราคา**หัว**ละ 300-500 บาท… …นอกจากจะพบว่ามีการแจกจ่ายเงิน**หัว**ละ 50 บาทแล้ว ยังจัดงานเลี้ยง… …เกิดในกรณีนี้ คือการที่รัฐ 50 รัฐ รวม**หัว**กันฟ้อง บริษัทบุหรี่ยักษ์ใหญ่ทั้งหมด…
Hair	Hair on the head of a human or an animal; hairstyle.	…เธอจะต้องโกน**หัว** ปฏิบัติธรรมอยู่ในวัดนี้… …เห็นคุยกับหนุ่ม**หัว**ตั้งสีทองแว็บๆ หันมาอีกทีกลืนไปในฝูงชน… …แต่กลับมอบให้ ดอน คิง โปรโมเตอร์**หัว**ฟูชื่อก้องโลกขึ้นป้าย อัลวาเรซ ชกชิงแชมป์… …เดินเซ็นเตอร์พอยท์มีคน**หัว**แดง**หัว**เขียว**หัว**เหลือง**หัว**บานเย็นเหมือนตุ๊กตุ่นตุ๊กตา…
Brain	Brain. It also refers to the seat of consciousness, thought, memory and emotion. Also, refers to a very intelligent or intellectual person as the chief planner of an organization or enterprise.	…เขาเครียดช็อคและเกร็งปวด**หัว**ข้างเดียว… …คณะกรรมการบริษัทคือส่วน**หัว** หรือมันสมองของบริษัท… …เป็นสื่อที่แปรความคิดจาก**หัว**สู่ร่างก็ขอให้ช่วยกันถกเถียงหา… …ไม่ต้องพูดพร่ำทำเพลงขอเพียงมี**หัว**เป็นไอเดีย มีปากเป็นภาษาอังกฤษ…
Emotion	Emotional and psychological state	…มันเริ่มทำให้ผมอารมณ์พลุ่งพล่าน**หัว**เสียมากขึ้น ผมอยากจะตะโกนด่า พวกมันว่า… …ซึ่งเป็นวันขึ้นปีใหม่ของจีน เจ้าคุณปัจจนึก**หัว**เสีย เพราะไม่มีขนมปังรับประทาน… …กุนซือปีศาจแดงดำ กล่าวอย่าง**หัว**เสียว่าเราไม่มี โอกาสที่จะผ่านเข้ารอบต่อไป… …โชคร้ายที่ไม่เป็นเช่นนั้น อาการ**หัว**เสียจึงเกิดขึ้น กับเสธ.หนั่น อย่างหลีกเลี่ยงไม่ได้…

Table 7: Definitions and derived senses of *หัว* /hua4/, which are not listed in the dictionary, and examples of these senses found in training corpus.

SENSE	DEFINITION	EXAMPLE
Machine part	Vital part of machine. For example, part which pulls the rest of an engine or a machine, part which cuts or emanates sound or energy.	…ปอนด์ที่มีประกอบด้วย สายคล้องไหล่ **หัว**ฟัง และรีโมท ได้รับเสียง… …ขยายธุรกิจเพื่อให้การบริการขนส่งทางเรือธุรกิจ**หัว**ลาก และแอร์คาโก้ รวมถึงการร่วมทุนกับบริษัท…
Chief	Top position of leadership, importance and honor, an individual holding these positions.	…นอกจากกลุ่มยังเติร์ก แล้ว ในระดับ**หัว**มีเพียงพลเอกจิ๋วคนเดียว… …ของเจ้าหน้าที่ตำรวจว่าเขาคือ**หัว**โจกในการถล่ม คุกเมืองหมอแคน… …ถ้าจะให้ดับทางการเมืองต้องเด็ดที่ **หัว** คือ ทักษิณ และเมื่อผู้นำดับพรรคก็จะ…
Heading	Information shown at the top of a page; title, heading; letterhead.	…เพราะมีชื่อผู้แปลทั้งสามคนปรากฏหราอยู่ที่**หัว**บทความนั้นผู้ทำงานไม่ได้คาดคิดกัน… ..และวันที่ 1 พ.ค. เป็น**หัว**กระดาษจากวัด ป่าบ้านตาด มีเลขที่หนังสือเรียบร้อยและลงนามโดยหลวง…
Headline	Headlines in newspapers	…นสพ.แทบทุกฉบับพากันพาด**หัว**ตัวโตกันเกรียวกราวเพราะไม่… …หนังสือพิมพ์รายวันของอิตาลีได้พาด**หัว**ตัวใหญ่ว่าเกิดข้อกังขาแม้ในวงการพระ…
Topics	Information represented in headlines, titles, and headings.	…ตัวอย่าง เช่น โฆษณาในหนังสือพิมพ์จั่ว**หัว**ว่า ข้อเท็จจริงของการล่าปลาวาฬ… …เปิดประเด็นระดมความคิดเรื่องม็อบ จั่ว**หัว**เรื่องเป็นภาษาไทยแต่บรรยายเป็น…
Titles or names	Titles or names of newspapers, book and magazine. Also used as a classifier.	…ต้องออกหนังสือมาเยอะๆ หลาย**หัว** หลายเล่มแต่พิมพ์จำนวนที่จำกัด… …นิตยสารลิซา แม้จะเป็น**หัว**นอกแต่เมื่อเข้ามาในประเทศไทยแล้วดูเหมือน… …ได้ข่าวแว่วๆ มาว่า ขณะนี้เว็บไซต์หนังสือพิมพ์**หัว**เขียวอย่าง www.thairath.co.th… …จากนามปากกาก็กลายมาเป็น**หัว**หนังสือ

		คือ ตุ๊กตา หนังสือการ์ตูนสำหรับ…

Table 7: Definitions and derived senses of *หัว* /hua4/, which are not listed in the dictionary, and examples of these senses found in training corpus.

SENSE	DEFINITION	EXAMPLE
To hide	To keep out of sight; to keep hidden from others; to hide.	…ไม่กล้าทำความเข้าใจเรื่องเพศศึกษา **เก็บ**กด ความคิดความรู้สึก… …กว่าผู้ชายธรรมดาอีกด้วย บางครั้งคนพวกนี้ต้อง**เก็บ**ตัวตนไว้ลึกๆข้างใน… …ไม่กล้าพูด กล้าแสดงออกมาก**เก็บ**งำไว้ในใจจะ เป็นคนที่รักและหวังดีตลอด… …ทำประชดโดยไปคบเม่นไม้**เก็บ**ความรู้สึกจนกลายเป็นความกดดัน…
To kill	To get rid of; to eliminate; to kill	…ทั้งงานใหญ่และงานเล็ก งาน**เก็บ**นักการเมืองระดับประเทศ… …เลือกตั้งแล้วยังจะต้องรับใบสั่ง**เก็บ**คู่แข่งด้วยอาชีพมือปืนผิดกฎหมาย… …อ้อม เมย์ ภัทรวรินทร์ เพื่อนร่วมทีมก็ถูกสั่ง**เก็บ** ฉากนี้ทั้งคู่ต้องวิ่งหนีการ… …ตัวเต็งคว้าที่นั่งอบต. ห้วยไผ่ ถูก**เก็บ**อีก 1 ขณะที่บางน้ำเปรี้ยวสุดพิลึก…
To purchase	To acquire; to buy especially in stock markets.	…ราคาปิดที่ระดับ 19.2 บาท และมีการไล่**เก็บ**หุ้นตัวนี้จนดันราคาหุ้นขึ้นไปสูงสุดที่ระดับ 39.75 บาท… …เก็งกำไรกันมากที่สุด โดยมีการไล่**เก็บ**มาตั้งแต่วันที่ 13 ธันวาคม 2542 จากการตรวจสอบ… …มีอินไซด์เดอร์ เทรดดิ้งหรือไม่ เนื่องจากการเข้า**เก็บ**หุ้น ของผู้บริหารเมื่อปลายเดือน… …นักลงทุนสถาบัน โดยเฉพาะกองทุน ต่างทยอย**เก็บ**หุ้น ในระดับราคาดังกล่าว…

Table 8: Definitions and derived senses of *เก็บ* /kep1/, which are not listed in the dictionary, and examples of these senses found in training corpus.

Figure 3: Semantic network representing all senses of หัว /hua4/ and other related concepts

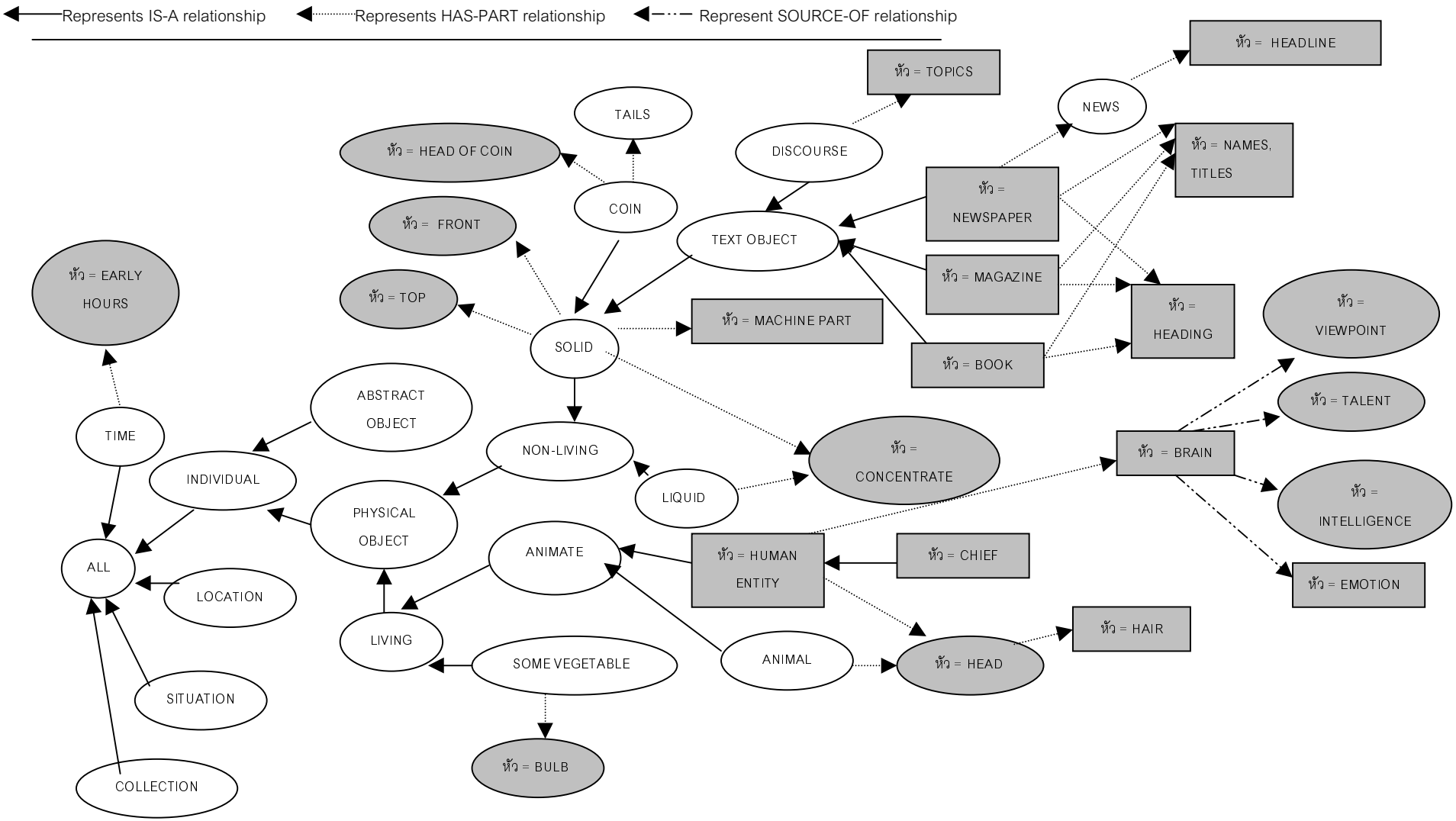


Figure 3 is a semantic network representing all senses of หัว /hua4/ and other related concepts. The network shows the polysemic development of senses of หัว /hua4/ which starts from "head", and extends to "bulb", to "top" and "front", to "early hours" and further to other senses.

In the network, a node represents a concept, the darker nodes with oval shape represent senses of หัว /hua4/ from the Thai dictionary of "The Royal Institute". The darker nodes with the rectangular shape represent senses of หัว /hua4/ that are not described in the dictionary. These nodes are connected by different kinds of arrow that represent the relationships among these concepts. There are three semantic relations in this network. The first is, IS-A relationship, which indicates that a concept is a kind or type of another concept. The second is, PART-OF relationship, which indicates that a concept is a part of another concept. The third is SOURCE-OF relationship, which indicates that a concept is a source of another concept. The interpretation of this network is, for example, PERSON is a kind of ANIMATE, which has HEAD as its part. BRAIN is a part of HEAD, which is a source of INTELLIGENCE, TALENT, VEIWPOINT and EMOTION.

3.3 Processes in WSD Using Decision List Algorithm

This study applied the decision list algorithm proposed by Yarowsky (1994) with some adaptations to suit our tasks. In this study, WSD using decision list algorithm consists of four processes as follows. (1) Data preparation process, which consists of data collection, word segmentation, word sense analysis, and word sense tagging. (2) The training process, in which decision lists for different spans of collocation are created. (3) The testing process, in which each span of collocation is tested by comparing whether word forms in the test data match with word forms in the decision list. The algorithm will choose the sense that co-occurs with the matched

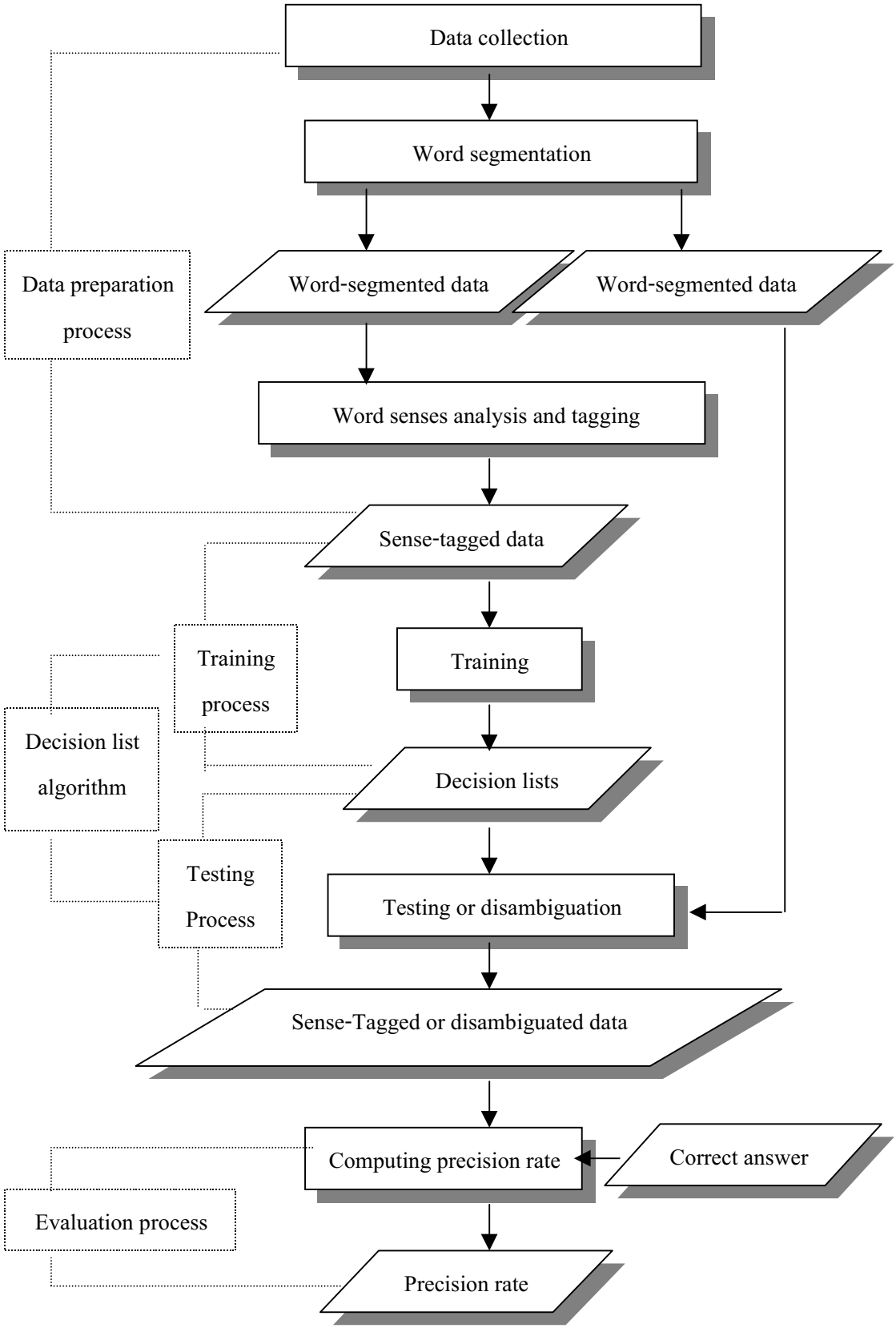


Figure 4: Processes in WSD using decision list algorithm.

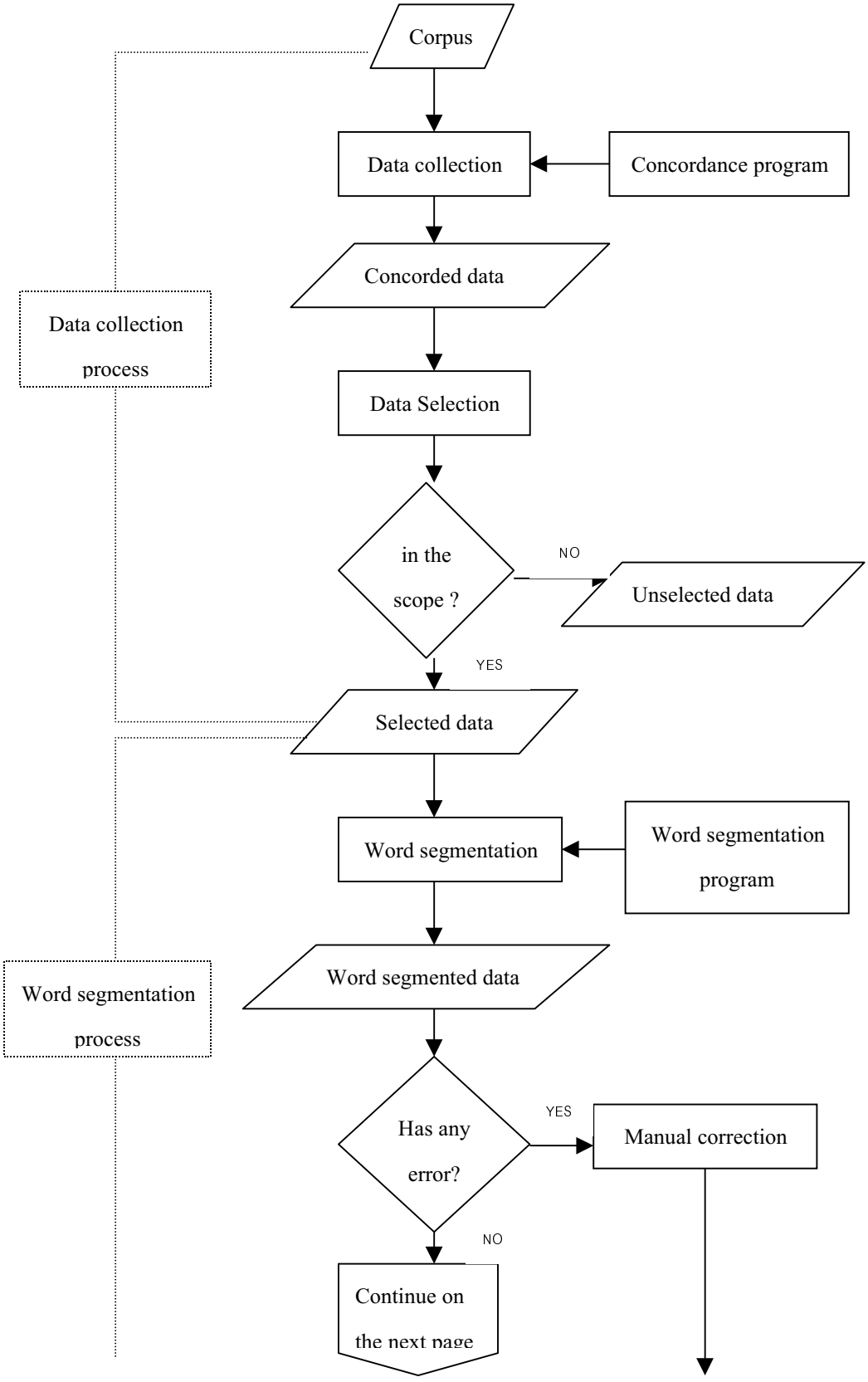


Figure 5: Data preparation processes

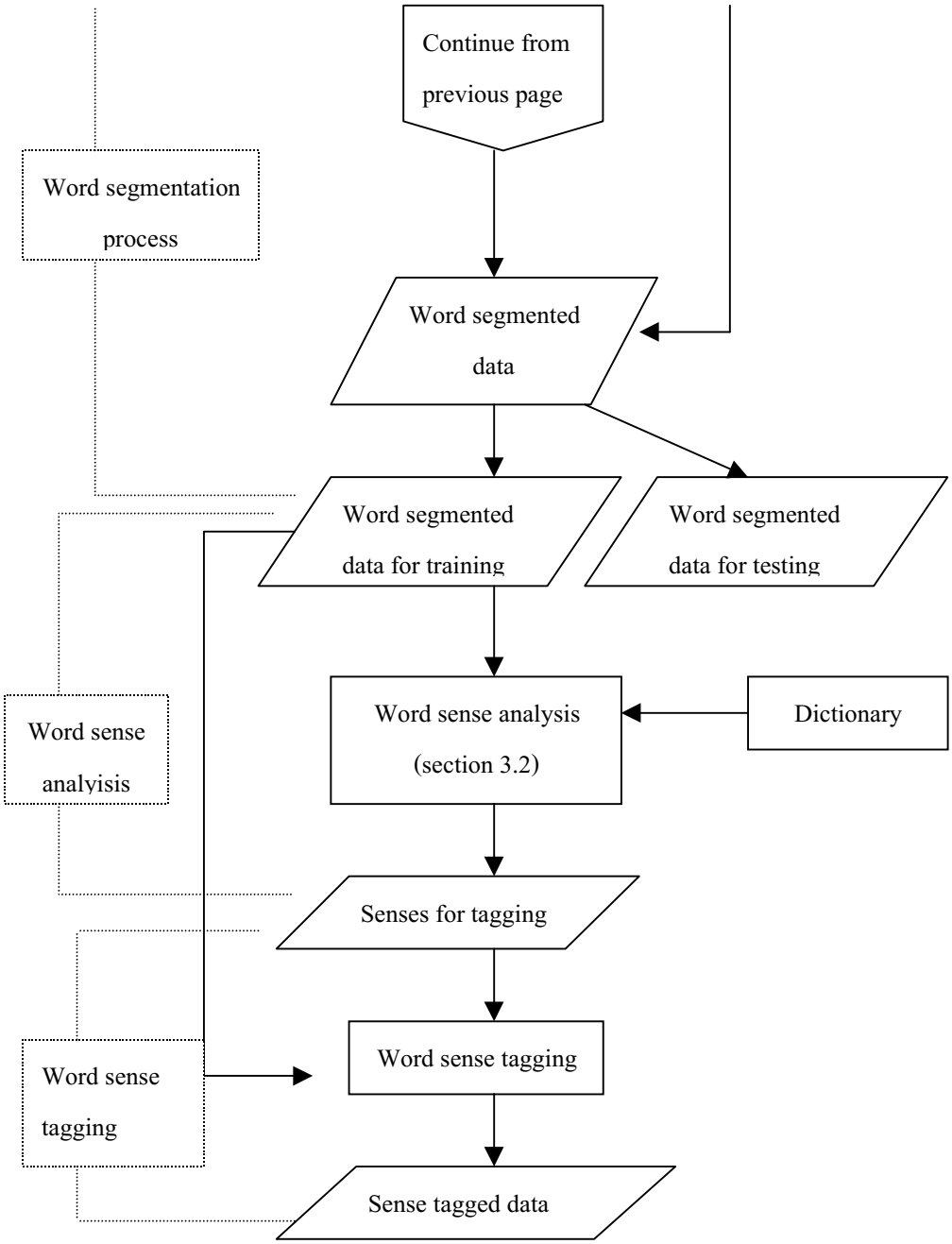


Figure 5: Data preparation processes.

word form that has the maximum collocational weight. (4) The evaluation process, in which precision rates of the performance in all tests are evaluated against the lower bound and upper bound performances. These four processes in this study are presented by the flow chart in figure 4 and are explained in details as follows.

3.3.1 Data preparation process

In a corpus based WSD, a large number of sense-tagged data is required for training the algorithm so that the algorithm can have sufficient knowledge for the disambiguation. In this study, the preparation of the data consists of four tasks as shown in figure 5 and explained in details as follows.

3.3.1.1 Data collection

In section 3.1, we have already explained the details about the source (section 3.1.1), the scope (section 3.1.2) and the size of the data (section 3.1.3). In this section we explains the method of collecting the data.

Data containing the ambiguous words *หัว* /hua4/ and *เก็บ* /kep1/ are collected using a concordance program[3], which randomly retrieves concordance lines containing the ambiguous words *หัว* /hua4/ and *เก็บ* /kep1/ at the center of the lines. For example,

> อักษรหรือตัวเลข เพื่อนำข้อมูลที่จัด**เก็บ**ไปประมวลผลก่อนออกเป็นรายงานข้อมูล
> แก่ผู้มีรายได้น้อย โดยจะ**เก็บ**ค่าใช้จ่ายในการไปตรวจรักษาครั้งละ 30 บาทเท่านั้น
> มักจะเป็นข้อมูลที่มีการ**เก็บ**รวบรวมโดยหน่วยงานราชการที่เกี่ยวข้องกับการ
> ต่อวันเท่านั้น และจะไม่**เก็บ**สินค้าค้างคืนมาขายอีก เพราะทำให้เสียรสชาติ"
> ออกจากตำแหน่ง หากว่าไม่สามารถ**เก็บ** 3 แต้ม โดยมีการเปลี่ยน แปลงทีม

[3] A program that searches for a specified word and usually shows the results in the form of key-word-in-context (KWIC).

Since not all the concorded data are fit with our scope of the data, these concorded data must be manually chosen according to our scope (which are discussed in section 3.1.2). Only the data that are in accordance with our scope will be chosen and used for the training and the testing processes.

3.3.1.2 Word segmentation

Since there is no word boundary in Thai written text, the concorded data must be word-segmented in order to be used in statistical processing of the algorithm such as counting, computing the collocational weight. In this study, the segmentation involves 2 steps as follows.

Step 1: Perform word segmentation automatically by a Thai word segmentation program.

Step 2: At the lexical level, manually correct any mistakes based on the context.

For example,

สอ ง	is changed to	*สอง*
หล่อ น	is changed to	*หล่อน*
อา การ	is changed to	*อาการ*
ไอ เดีย	is changed to	*ไอเดีย*

From the examples above, which are the results from running the word segmentation program, *สอ ง* is changed to *สอง*, even though *สอ* /รวว4/ could be a word, but from the context, *สอง* /รววฦ4/ is the correct word for segmentation. The

results, *หล่อ น*, *อา การ*, *ไอ เดีย*, are changed to *หล่อน*, *อาการ*, *ไอเดีย*, respectively for the same reason.

. .1. or sense anal sis

In section 3.2, we have already explained the details about word sense analysis, which involves the analysis of word senses based on the definitions in the Thai dictionary of "The Royal Institute" (section 3.2.1) and the information from the training corpus (section 3.2.2). The results of the analysis are twenty senses of *หัว*/hua4/ and nine senses of *เก็บ* /kep1/, which are used for sense tagging.

3.3.1.4 Word sense tagging

In order to be convenient for manually tagging and statistical processing, the senses of *หัว* /hua4/ and *เก็บ* /kep1/ are tagged in the form of "<number>". The sense of each number is shown in table 9 for *หัว* /hua4/ and table 10 for *เก็บ* /kep1/. The sense-tagged training corpus of *หัว* /hua4/ and *เก็บ* /kep1/ are in appendix B. The sense-tagged testing corpus of *หัว* /hua4/ and *เก็บ* /kep1/ are in appendix C.

We would like to note here that, these numbers do not indicate any hierarchical relationship between senses. However, they have some relationship to each other as some senses are extended from the others due to the polysemous development process (See section 3.2.2 in details). These senses also have some relationship to the other concepts too. These relationships are shown as a semantic network in figure 3, section 3.2.2.

Tag sets	Senses
<1>	Head
<2>	Entity
<3>	Chief
<4>	Hair
<5>	Brain
<6>	Intelligence
<7>	Talent
<8>	Viewpoint
<9>	Emotion
<10>	Top
<11>	Heading
<12>	Headlines
<13>	Front
<14>	Machine part
<15>	Early hours
<16>	Bulb
<17>	Head of coin
<18>	Concentrate
<19>	Topics
<20>	Titles or names

Table 9: Tag sets representing senses of หัว /hua4/.

Tag sets	Senses
<1>	To take
<2>	To pick up
<3>	To arrange
<4>	To keep
<5>	To hide
<6>	To gather
<7>	To charge
<8>	To kill
<9>	To buy

Table 10: Tag sets representing senses of*เก็บ* /kep1/.

The next three sections involved the decision list algorithm, which consists of the training (section 3.3.2), the testing (section 3.3.3), and the evaluation processes (section 3.3.4). Figure 6 illustrates an overview of these processes. The details of each process are as follows.

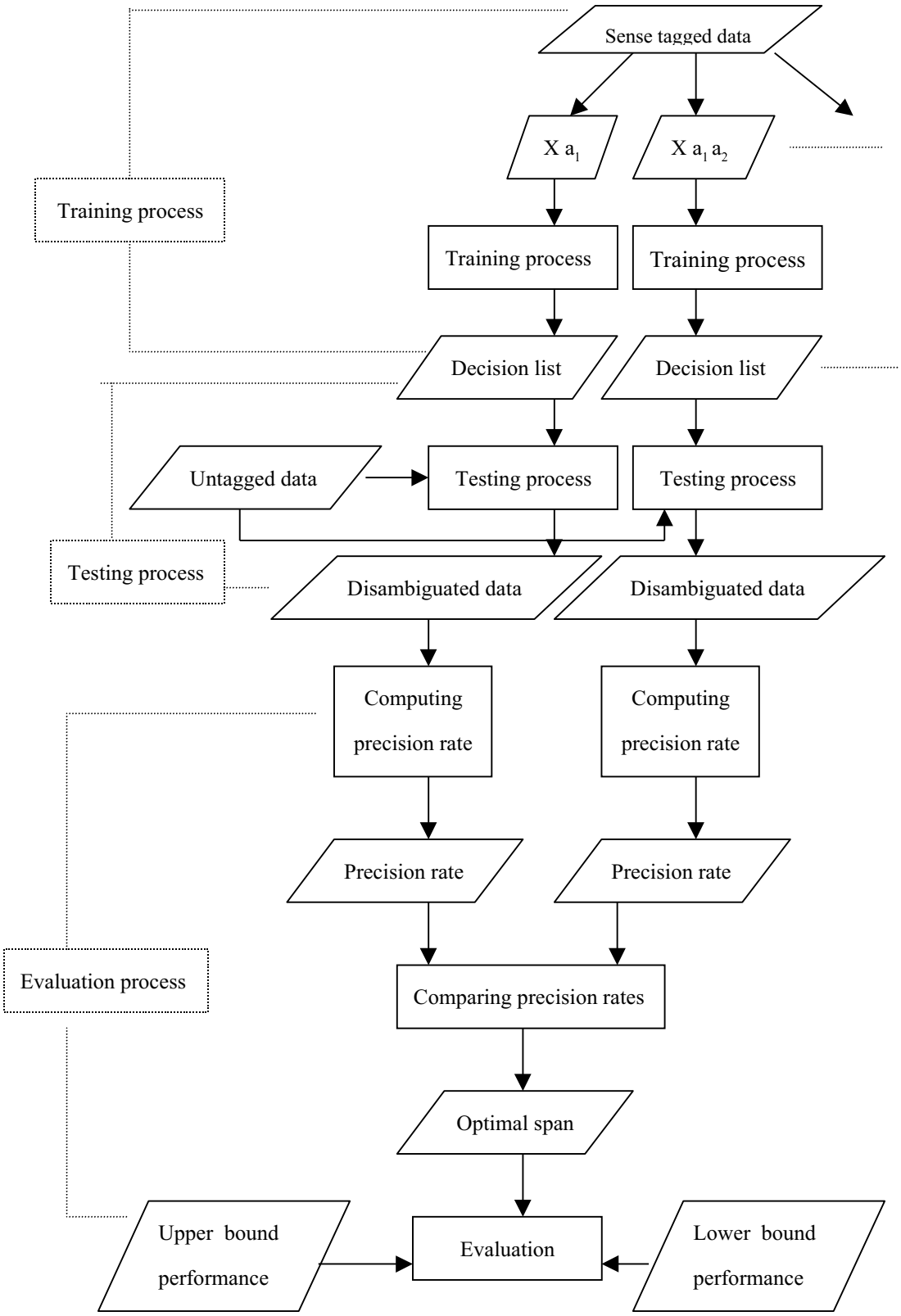


Figure 6: Training, testing and evaluation processes.

3.3.2 Training process

Comment: Training process (section 3.3.2)

1 **for** all spans of word **do**

2 **for** all line containing an ambiguous word **do**

3 **for** all word forms (W_k) in a span **do** #select word forms co-occur with ambiguous word in a running span.

4 $C(S_i,W_k)$ #count the frequency of occurrences of senses with word forms.

5 $C(W_k)$ #count the frequency of occurrences of word forms.

6 **end**

7 **end**

8 **end**

Comment: Computing collocational weight

9 **if** $C(W_k) >= 3$ **do** # if word forms occurs more than or equal 3 times do…

10 $P(S_i \mid W_k) = C(S_i,W_k) / C(W_k)$ #compute the probability of co-occurrences, and after that…

11 **if** $P < 1$ # if the probability is less than 1 do…

12 $Weight(S_i,W_k) = Log \left(\frac{P(S_i \mid W_k)}{\sum_{j \neq i} P(S_j \mid W_k)} \right)$ #compute the weight of co-occurrences.

13 **else** Weight = 9 # if P equal 1, assign value 9 to the weight

14 **end** # continue …

Table 11: Decision list algorithm.

Comment: Testing Process: Disambiguation (section 3.3.3)

#continue…

15 **for** all spans **do**

16 **for** all lines containing an ambiguous word **do**

17 **for** all words (in a testing corpus) (w_k) in a span **do**

18 **if** w_k = W_k # check whether word forms in the test data matches with word form in a decision list, if matches, do…

19 **if** there is only one maxweight

20 **if** weight(S_i | W_k) = maxweight

21 **choose** S_i that has maxweight

22 **return** ambiguous word with S_i

23 **end**

24 **if** there are more than one maxweight

25 **if** word forms are in the different positions

26 **choose** S_i that pointed by the nearest word form

27 **return** ambiguous word with S_i

28 **end**

29 **if** word forms are in the same position

30 **choose** S_i that has the highest frequency among these senses

31 **return** ambiguous word with S_i

32 **end**

33 **end**

34 **if** w_k ne W_k

35 **choose** sense that have the highest frequency in the corpus

36 **return** ambiguous word with S_i

37 **end**

38 **end**

Table 11: Decision list algorithm.

This process creates decision lists for disambiguation. The "decision list algorithm" (Yarowsky, 1994) is adapted to suit this study. The algorithm is shown in table 11, and explained in details as follows.

Since this study wants to find the optimal span for locating sense indicators of WSD, the algorithm will be trained at different spans to create various decision lists for the disambiguation. Since sense indicators are hypothesized to be in the span of five words and they can be found either on the left or on the right side of the ambiguous word, the total settings for testing in this study will be twenty, as illustrated in Figure 7, where X is an ambiguous word, a_1 to a_5 are context words on the right side, and b_1 to b_5 are context words on the left side.

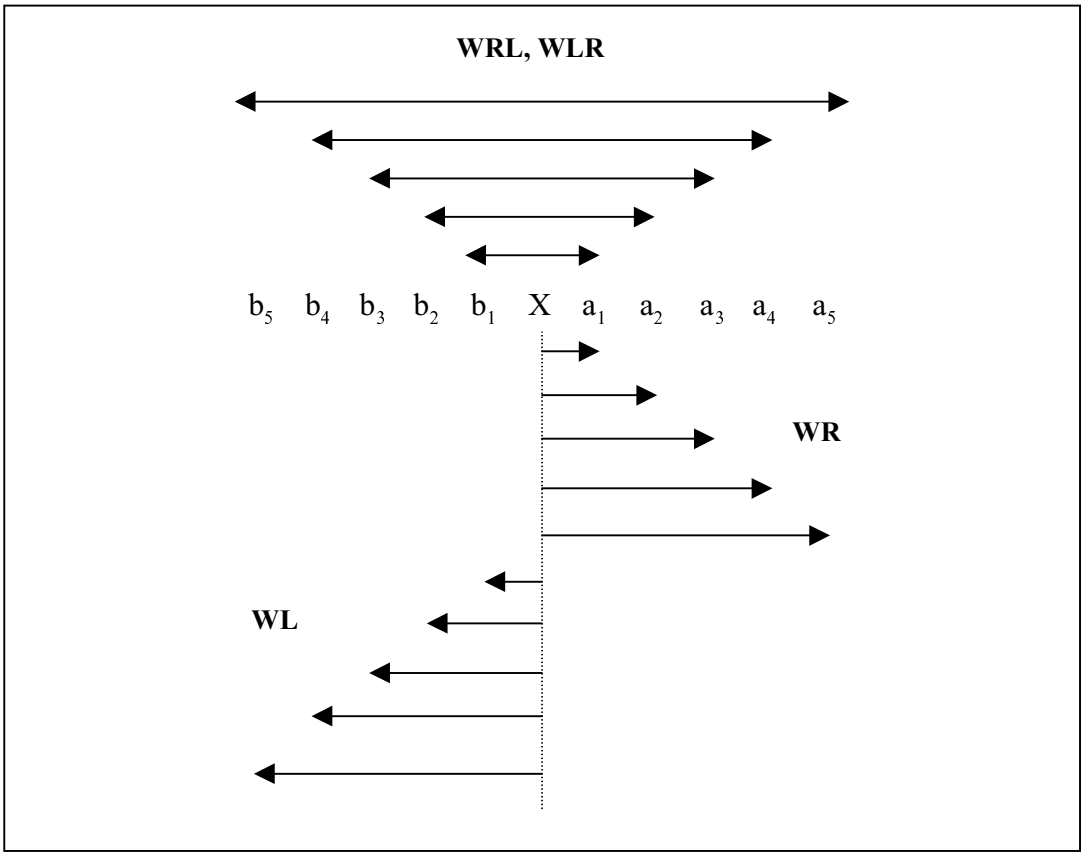


Figure 7: Twenty spans for training and testing.

Word-to-right (WR) consists of 5 spans as follows.

One-word-to-right (1WR) ⟶ X a_1

Two-words-to-right (2WR) ⟶ X a_1 a_2

Three-words-to-right (3WR) ⟶ X a_1 a_2 a_3

Four-words-to-right (4WR) ⟶ X a_1 a_2 a_3 a_4

Five-words-to-right (5WR) ⟶ X a_1 a_2 a_3 a_4 a_5

Word-to-left (WL) consists of 5 spans as follows.

One-word-to-left (1WR) ⟶ X b_1

Two-words-to-left (2WR) ⟶ X b_1 b_2

Three-words-to-left (3WR) ⟶ X b_1 b_2 b_3

Four-words-to-left (4WR) ⟶ X b_1 b_2 b_3 b_4

Five-words-to-left (5WR) ⟶ X b_1 b_2 b_3 b_4 b_5

Word-to-right-and-left giving priority to word-to-right (WRL) consist of 5 spans as follows

One-word-to-right-and-left (1WRL) ⟶ X a_1 b_1

Two-words-to-right-and-left (2WRL) ⟶ X a_1 b_1 a_2 b_2

Three-words-to-right-and-left (3WRL) ⟶ X a_1 b_1 a_2 b_2 a_3 b_3

Four-words-to-right-and-left (4WRL) ⟶ X a_1 b_1 a_2 b_2 a_3 b_3 a_4 b_4

Five-words-to-right-and-left (5WRL) ⟶ X a_1 b_1 a_2 b_2 a_3 b_3 a_4 b_4 a_5 b_5

Word-to-left-and-right giving priority to word-to-left (WLR) consist of 5 spans as follows

One-word-to-left-and-right (1WLR) ⟶ X b_1 a_1

Two-words-to-left-and-right (2WLR) ⟶ X b_1 a_1 b_2 a_2

Three-words-to-left-and-right (3WLR) ⟶ X b_1 a_1 b_2 a_2 b_3 a_3

Four-words-to-left-and-right (4WLR) ⟶ X b_1 a_1 b_2 a_2 b_3 a_3 b_4 a_4

Five-words-to-left-and-right (5WLR) ⟶ X b_1 a_1 b_2 a_2 b_3 a_3 b_4 a_4 b_5 a_5

For each span (line 1, table 11)**, then, for each line containing an ambiguous word** (line 2, table 11)**,**

Step 1: count the frequencies of the co-occurrences of the senses of *หัว* /hua4/ and *เก็บ* /kep1/ and the word forms ($C(S_i, W_k)$ within the span, and the frequencies of occurrences of the word forms ($C(W_k)$). The following is the explanation of this step.

Suppose that, we have the following data in our training corpus.

(i) ลักษณะ เป็น โครง หลัก โปร่ง สูง ระดับ *หัว*<1> คน มี ช่อง ให้ เสียบ หนังสือ ซ้อน

(ii) เพิ่มขึ้น เนื่องจาก หอม จะ ลง *หัว*<16> ช้า เพราะ ต้อง หา อาหาร ไป เลี้ยง ดอก หอม

(iii) ที่ ออกแบบ เป็น สาย รัด และ *หัว*<13> เข็มขัด เข้า ชุด กับ กระเป๋า ถือ

(iv) หนุ่ม ๆ ที่มา ยืน รอ หน้า งาน ตั้งแต่ *หัว*<15> ค่ำ ก็ ตบ เท้า เข้า งาน กัน อย่าง

When training the program at the span of 2WR, in line (i) (which assumed to be the first line of the data), the program will consider only *คน* /khon/ and *มี* /mii/, then count the frequency of *หัว<1>* co-occurred with *มี* /mii/, the frequency of *หัว<1>* co-occurred with *คน* /khon/, and count the frequency of *คน* /khon/ and the frequency of *มี* /mii/. Below are the information after the training algorithm processes the data.

S_i,W_k	No. of occurrence ($C(S_i,W_k)$)	W_k	No. of occurrence ($C(W_k)$)
หัว<1> คน	1	*คน*	1
หัว<1> มี	1	*มี*	1
หัว<16> ช้า	1	*ช้า*	1
หัว<16> เพราะ	1	*เพราะ*	1
หัว<13> เข็มขัด	1	*เข็มขัด*	1
หัว<13> เข้า	1	*เข้า*	1
หัว<15> ค่ำ	1	*ค่ำ*	1
หัว<15> ก็	1	*ก็*	1

If the data size is increased so that the program further finds *คน* /khon/ co-occurred with *หัว*<1> within the span, it will add one to the frequency of the co-occurrence of "*หัว<1> คน"*, henceforth C(*หัว<1>, คน*), so the frequency of C(*หัว<1>, คน)* will equal to two. The program will also add one to the frequency of occurrence of *คน*, so the frequency will equal to two.

Step 2: compute the probability of $P(sense_i \mid word\ form_k)$, which is the probability of the ambiguous word being marked with $sense_i$ when the word $form_k$ is found in the span, and the probability of $P(sense_j \mid word\ form_k)$, which is the probability of the ambiguous word being senses other than $sense_i$, when the word $form_k$ is found in the span. The ratios of these two proabilities are used for computing the weight or strength of co-occurrence between $sense_i$ and word $form_k$. The formula for computing collocational weight is shown below:

$$P(\text{sense}_i \mid \text{word form}_k) = \frac{C(\text{sense}_i, \text{word form}_k)}{C(\text{word form}_k)}$$

$$= \frac{\text{Total occurrence of word form}_k \text{ with sense}_i}{\text{Total occurrence of word form}_k}$$

(line 10, table 11)

$$\text{Weight}(\text{sense}_i, \text{word form}_k) = \text{Log}\left(\frac{P(\text{sense}_i \mid \text{word form}_k)}{\sum_{j \neq i} P(\text{sense}_j \mid \text{word form}_k)}\right)$$

$$= \text{Log}\left(\frac{P(S_i \mid W_k)}{1 - P(S_i \mid W_k)}\right)$$

(line 12, table 11)

The formula is provided by Agirre, and Martinez (2000), which is adapted from that of Yarowsky (1994) [4] to suit WSD task, in which a word can have more than two ambiguities (senses).

[4] See the formula proposed by Yarowsky in section 2.5.2.4, which is used for lexical ambiguity resolution such as homograph disambiguation, in which there are only two ambiguities of a target ambiguous word.

Since the collocational weight is the logarithm of the ratio between the probability of the co-occurrence of word $form_k$ and $sense_i$, and the probability of the co-occurrence of word $form_k$ and other senses (excluding $sense_i$), if the weight is higher than 0, word $form_k$ will has higher probability to co-occur with sense $_i$ than other senses. Thus, word $form_k$ is a better indicator of sense $_i$ than other senses.

In this study, the collocational weights are computed under the following conditions.

(1) The program will compute the probability only if $C(W_k)$ is greater than or equal to 3 (line 9, table 11). This is to lessen the effect caused by the small size of training data because using small frequency might result in an unjustified decision.

The following example shows that an incorrect decision is the result of the computation when $C(W_k)$ is less than 3. In (i), when training at span two-word-to-right-and-left (2WRL), *ที่ดิน* /thii2din/ co-occurs with *เก็บ*<4> 2 times and *ที่ดิน* /thii2din/ occurs only 2 times in a training corpus, if the program computes its collocational weight, P(*เก็บ*<4> | *ที่ดิน*) will be equal to 1 and P(*เก็บ*-other senses | *ที่ดิน*) will be equal to 0. This makes the co-occurrence of *เก็บ*<4> and *ที่ดิน* /thii2din/ highly significant. Thus, when disambiguates the following sentence,

(i) ของ เกษตรกร การ จัด รูป ที่ดิน เพื่อ **เก็บ** เงิน คืน ทุน สิ่ง สำคัญ ก่อน ที่ รัฐ จะ เข้าไป ลงทุน

the program will choose sense <4>. But this is not the correct sense. The program is expected to choose the sense <7>. This sense (<7>) co-occurs with *เงิน* /ŋən/ with the highest weight of 0.4214 (excluding the collocational weight of *เก็บ*<4> and *ที่ดิน*). Since in the corpus, *เงิน* /ŋən/ occurs 132 times, and occurs with *เก็บ*<7> *95* times, the co-occurrence of *เก็บ*<7> and *เงิน* /ŋən/ should be more significant than the co-occurrence of *เก็บ*<4> and *ที่ดิน* /thii2din/.

(2) The program will compute the collocational weight only if the probability (P) is less than one (line 11, table 11). This is because, when P is equal to 1, the weight will equal log (1/1-1) which is infinity, and the program will stop working. So, if P is equal to 1, the program will assign the highest value to the weight. In this study, the assigned value is 9 because the highest weight computed from this study is less than 3 (line 13, table 11).

Step 3: Continue training until the last line of the training data, and then, continue training at other spans till all spans are trained.

After the training process, there are twenty decision lists for twenty different spans of word. Each list consists of collocational patterns (senses co-occur with word forms) and their weights. At the time of disambiguation, for each line containing the ambiguous word *หัว* /hau4/ or *เก็บ* /kep1/, the program will compare word forms within the span with the word forms in the decision list, and choose the sense co-occurred with the word form that has the highest weight.

3.3.3 Testing process

The testing process applies the decision lists created from the training process to the testing corpus. Since there are twenty decision lists for twenty different strategies (spans) for disambiguation, the testing process is run twenty times for every decision list. The process has two steps as follows:

For each span, and then for each line containing an ambiguous word,

Step 1: Check whether the collocational word forms (w_k) of the ambiguous words in the testing corpus are the same as in the decision lists (W_k) (line 18, table 11).

(1) If they are the same, then makes the decision according to the following cases.

Case 1: In case that there is only one sense that has maximum weight, the program chooses this sense (lines 19, 20 and 21, table 11). This is the ordinary case performed by the program based on the statistical view that, collocational patterns that receive higher weight are more statistically significant than those with lower weight . Therefore, they should be better sense indicators. The example below shows the use of decision list for disambiguating.

Example 1: Disambiguating at 2WR

เขื่อน ภูมิพล และ เขื่อน สิริกิติ์ จะ มี การ เก็บ กัก น้ำ เอาไว้ ใช้ มาก กว่า ปี ที่ ผ่าน มา

Weight	Collocational pattern	Sense
9	เก็บ กล้วยไม้	**<1>**
9	เก็บ กัก (matched)	**<4>**
9	เก็บ ผัก	**<1>**
…	...	**…**
2.1106	เก็บ ภาษี	**<7>**
1.4914	เก็บ แต้ม	**<6>**
…	…	**…**
0.9700	เก็บ น้ำ (matched)	**<4>**
…	…	**…**
-1.0263	เก็บ น้ำ (matched)	**<7>**
…	…	**…**
-1.4914	เก็บ น้ำ (matched)	**<1>**
…	**…**	**…**

Table 12: Decision list[5] for *เก็บ* /kep1/ at 2WR.

[5] This decision list and other lists shown in this section are the abbriviated decision lists and sorted by weight from the highest to the lowest weight.

In example 1, the program is tested at the span 2WR, it will select word forms within this span, which are *กัก* /kak1/ "to detain" and *น้ำ* /naam3/ "water". Then, the program checks these word forms whether they match word forms in the decision list of this span, if match, the program will extract these word forms and their collocational senses (which can be more than one sense) and their weights. In this example, the program extracts 4 collocational patterns as follows.

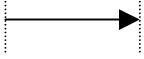
Collocational pattern	Weight
กัก เก็บ<4>	9
น้ำ เก็บ<4>	0.9700
น้ำ เก็บ<7>	-1.0263
น้ำ เก็บ<1>	-1.9638

Then, the program compares among these patterns to see whether which one has the highest weight. In this example, sense <4> indicated by *กัก* /kak1/ has the highest weight, so the program chooses sense <4> for *เก็บ* /kep1/ in this context.

Case 2: In case that there are many maximum weights, and the positions of word forms are different, the program will choose sense indicated by the nearest word form to the ambiguous word (line 24, 25 and 26, table 11). If both left and right words are considered, the nearest word is determined from the priority setting in the span. For example, in the span 2WRL, b_2 b_1 X a_1 a_2, the order of words sorted by the nearness is a_1 b_1 a_2 b_2. This decision can be explained as follows.

Example 2: Disambiguating at 2WR

ทาง เดียว ที่ จะ ไม่ อันตราย จาก การ เก็บ ค่า ใช้ ซอฟแวร์ จะ เกิดขึ้น เมื่อ ซอฟต์แวร์



Weight	Collocational pattern	Sense
9	เก็บ กล้วยไม้	<1>
9	เก็บ แมลง	<1>
9	เก็บ กัก	<4>
9	เก็บ ผัก	<1>
…	…	…
9	เก็บ ใช้ (matched)	<7>
…	…	…
9	เก็บ ค่า (matched)	<7>
…	...	…

Table 13: Decision list for *เก็บ* /kep1/ at 2WR.

In example 2, the program is tested at the span 2WR, it will select word forms that are within this span, which are *ค่า* /khaa2/ "cost" and *ใช้* /chai3/ "use". Then, the program checks these word forms whether they match word forms in the decision list of this span, if match, the program will extract these word forms and their collocational senses and their weights. In this example, the program extracts 2 collocational patterns as follows.

Collocational pattern	weight
ค่า เก็บ<7>	9
ใช้ เก็บ<4>	9

Then, the program compares among these patterns to see whether which one has the highest weight. In this example, both patterns have the same weight. According to this case, the program will choose sense <7>, indicated by the nearest word form (*ค่า* /khaa2/) to the ambiguous word *เก็บ* /kep1/.

Example 3: Disambiguating at 1WLR

เดินทาง ไป ดู พื้นที่ เกิดเหตุ เพลิง ไหม้ ถัง **เก็บ** น้ำมัน ของ บริษัท ไทยออยล์ จำกัด ว่า ได้

Weight	Collocational pattern	Sense
…	…	…
9	ซื้อ เก็บ	**<4>**
9	อ่าง เก็บ	**<4>**
9	เก็บ น้ำมัน (matched)	**<4>**
9	เก็บ น้ำมันหล่อลื่น	**<4>**
9	เก็บ ความลับ	**<4>**
9	สถานที่ เก็บ	**<4>**
…	…	**…**
9	เก็บ ข้าวเปลือก	**<4>**
9	ถัง เก็บ (matched)	**<4>**
…	…	**…**

Table 14: Decision list for *เก็บ* /kep1/ at 1WRL.

In example 3, the program is tested at the span 1WLR, it w ill select word forms within this span, which are *ถัง* /thaŋ/ "tank" and *น้ำมัน* /naam3man/ "oil". Then, the program checks these word forms whether they match word forms in the decision list of this span, if match, the program will extract these word forms and their collocational senses and their weights. In this example, the program extracts 2 collocational patterns as follows.

Collocational pattern	weight
น้ำมัน เก็บ<4>	9
ถัง เก็บ<4>	9

Then, the program compares among these patterns to see which one has the highest weight. In this example, both patterns have the same weight. According to the span 1WLR, the program will choose sense <4> indicated by *ถัง* /thaŋ4/ "tank", which is in the nearest position to the ambiguous word *เก็บ* /kep1/.

Thus, the difference between right priority, and left priority is that, if it is the right priority, program will choose word form at the right first, if it is left priority, the program will choose word form at the left first.

Case 3: In case that the weights are equal and the word form indicate more than one senses, the program will choose the sense that has the highest frequency of occurrence. (line 24, 29 and 30, table 11) This decision can be explained as follows.

Example 4: Disambiguating at 1WR

กล่าว ให้ เอกชน เข้ามา ดำเนินการ และ **เก็บ** ผลประโยชน์ แน่นอน ว่า ประชาชน โดยทั่ว

Weight	Collocational pattern	Sense
…	…	…
0.1250	เก็บ เบี้ย	**<7>**
0.0792	เก็บ หนังสือ	**<4>**
0	เก็บ ผลประโยชน์ (matched)	**<7>**
0	เก็บ ยา	**<4>**
0	เก็บ ผลประโยชน์ (matched)	**<1>**
…	…	**…**

Table 15: Decision list for *เก็บ* /kep1/ at 1WR.

In example 4, the program is tested at the span 1WR, it will select word forms within this span, which in this example, is *ผลประโยชน์* /phon4pra1yoot1/ "benefits".

Then, the program checks this word form and extracts 2 collocational patterns as follows.

collocational pattern	weight
ผลประโยชน์ เก็บ<7>	0
ผลประโยชน์ เก็บ<1>	0

Then, the program compares among these patterns. In this example , both patterns have the same weight. According to this case, the program will choose <7> because this sense is found more than sense <1> in the training corpus.

(2) If collocational words of the ambiguous words are not found in the decision lists, the program will simply choose the sense that has the highest frequency in the training corpus, which is "top part of human or front part of other animals" for *หัว* /hua4/, and "to maintain, store and keep" for *เก็บ* /kep1/ (line 34 and 35, table 11), in this study.

As stated earlier that case 1 is the ordinary case that the program expects to find. The reason that there are cases 2, 3 (there are more than one sense with the maximum weights) and (2) (collocational word is not found in the decision list) is because of the size of the training data, which may not be large enough to enable a word form to be a clear distinctive sense indicator. So, there are many word forms that indicate many senses, and there are also many senses indicated by the same word form. However, this problem can be solved by increasing the size of the training data [6].

---

[6] Chapter 5, section 5.1.1 discusses about the effect of the size of the training data, by testing the disambiguation at different training data size, and found that the higher the size of the data, the better the algorithm's performance.

Thus, in this study, case 2, case 3, and (2), are the problems with their reasonable resolutions. In case 2, the decision of choosing the sense indicated by the nearest word form is in accordance to our hypothesis that the word form that is close to the ambiguous word is the better sense indicator than words that are in a further distance. In case 3 and (2), the reason that the program chooses the most frequent sense is in accordance to the human retrieval of sense, which states that human always thinks or retrieves the sense that is the most dominant or frequent first if given a neutral context or no context at all (Simpson, 1981).

Step 2: Return the chosen sense to the ambiguous word. Thus, in example 1 the program returns sense <4> to *เก็บ* /kep1/ as follow.

เขื่อน ภูมิพล และ เขื่อน สิริกิติ์ จะ มี การ เก็บ<4> กัก น้ำ เอาไว้ ใช้ มาก กว่า ปี ที่ ผ่าน มา

Step 3: Continue to disambiguate other lines till the last line of the testing data, and then, continue to disambiguate at other spans till all twenty spans are tested.

3.3.4 Evaluation process

In this study, since it is intended to find the optimal span for locating sense indicators, the performance of each span ranged from 1 to 5 will be computed for the precision rate. The precision rate of each span will be compared to get the optimal span, which will be used to evaluate against the lower bound and upper bound performances (the evaluation will be discussed in chapter 4). The precision rate is computed as follows (Agirre and Rigau, 1996).

$$\text{Precision rate} = \left(\frac{\text{Number of times the sense is correctly disambiguated}}{\text{Total number of answered senses}} \right) * 100$$

In this research, the precision rate will be computed automatically as follows.

The program checks whether each sense selected is the correct sense, if it is correct, the program will increase the number of the sense correctly disambiguated. The total number of answered senses will be the same as the number of testing data which is 400 for *หัว* /hua4/ and 400 for *เก็บ* /kep1/.

There are 3 precision rates for the evaluation.

(1) Precision rate of the algorithm tested at twenty spans, which is computed as explained above.

(2) Precision rate for lower bound performance, which is the performance of the simple algorithm using highest frequent sense as a cue for disambiguation. This means that the algorithm always returns the sense that has the highest frequency in a testing corpus for disambiguation.

- The precision rate for lower bound performance disambiguating *หัว* /hua4/ is 24%, which is computed as follows.

$$\text{Precision rate} = \left(\frac{\text{Frequency of } S_x \text{ in the testing corpus}}{\text{Total number of answered senses}} \right) * 100$$

Where, S_x is the sense that has the highest frequency in the training corpus.

= (96 / 400) * 100 = 24%

- The precision rate for lower bound performance disambiguating เก็บ /kep1/ is 40%, which is computed as follows.

Precision rate = (160/400)*100 = 40%

(3) Precision rate for upper bound performance. In this study, the upper bound performance is the performance of the researcher's tagging which is assumed to be 100% correctly.

CHAPTER IV

RESULTS

This chapter consists of three sections. Section 4.1 and section 4.2 report and discuss the results of the algorithm's performance on disambiguating หัว /hua4/ and เก็บ /kep1/ respectively. The results consist of the precision rates of twenty tests of หัว /hua4/ and เก็บ /kep1/ and twenty tests of each sense of หัว /hua4/ and เก็บ /kep1/. The precision rates of all results are compared to get the optimal spans for disambiguating both words, and the optimal span for disambiguating each sense of them. The results of the optimal spans of หัว /hua4/ and เก็บ /kep1/ will be discussed based on three perspectives, namely computational, syntactic and semantic perspectives. Section 4.3 summarizes the results and further discusses why some senses have high precision rate and some senses have low precision rate. The details of each section are as follows.

4.1 The Results of the Disambiguation of หัว /hua4/

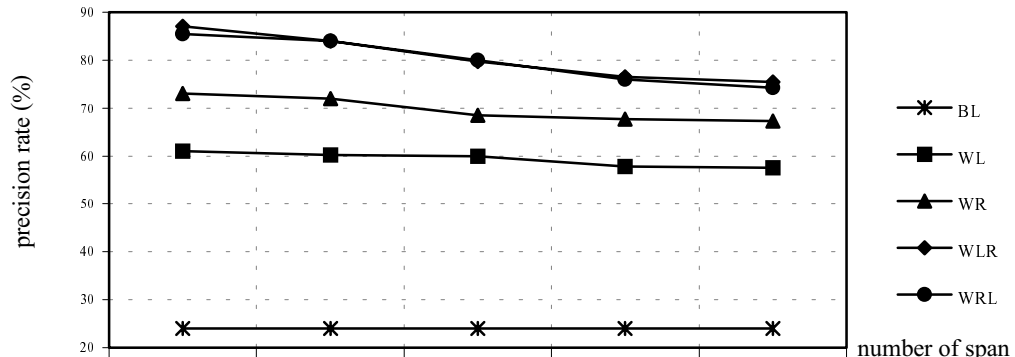
Figure 8 shows that the optimal span for the disambiguation of หัว /hua4/ is 1WLR as indicated by the precision rate of 87%, 3.625 times[1] higher than the lower bound performance and 0.87 time[2] lower than the upper bound performance. Since the precision rates of WLR and WRL are not significantly different, the presentation and the explanation of the results, from now on, will use WRL only, whether, in fact, it refers to WRL or WLR. The poorest span for the disambiguation of หัว /hua4/ is 5WL

---

[1] 3.625 times is computed as 87 / 24 (which is the precision rate of the lower bound performance).

[2] 0.87 times is computed as 87 / 100 (which is the precision rate of the upper bound performance).

as indicated by the precision rate of 57.50%. From the evaluation, we can see that the algorithm's performance is very good. The strengths of the algorithm that enable it to achieve high performance as well as the limitations that unable it to perform as good as human will be discussed in section 5.1.1.



	1	2	3	4	5
BL	24.00	24.00	24.00	24.00	24.00
WL	61.00	60.25	60.00	57.75	57.50
WR	73.00	72.00	68.50	67.75	67.25
WLR	87.00	84.00	79.75	76.50	75.50
WRL	85.50	84.00	80.00	76.00	74.25

Figure 8: Precision rate of disambiguation of หัว /hua4/. BL is base line, WR is word-to-right of an ambiguous word, WL is word-to-left of an ambiguous word, WRL is word-to-right-and-left of an ambiguous word giving the priority to word-to-right, WLR is word-to-left-and-right of an ambiguous word giving the priority to word-to-left.

In terms of the number of context words for disambiguation of หัว /hua4/, the optimal is one, which is in accordance with our hypothesis. The precision rate decreases respectively as the number of word increases from 1 to 2, 3, 4 and 5 words. However, the 2W is also good for the disambiguation, the precision rate at 2WRL span is only 3% lower than that of 1WRL. These results are supported by the results of disambiguating on each sense of หัว /hua4/, in which, there are 14 senses that have the optimal span as 1W, while 2 senses have the optimal span as 2W and only one sense needs 3W span. Thus, 3W span is sufficient for the disambiguation of หัว /hua4/.

In terms of side, the optimal side for disambiguating is both right and left, which contradicts to our hypothesis. However, this is not because both right and left sides play equal role in the disambiguation. But, because WR plays more role and with some influence from WL, considering WRL yields a better result than considering WR or WL alone. These results are supported by the results on disambiguating each sense of หัว /hua4/, in which, there are 12 senses that their sense indicators are on the right side, 2 senses are on the left side, and 3 senses are on both right and left sides.

Since different senses require different spans for the disambiguation, the next section (section 4.1.1) presents and discusses the results of the algorithm's performance when disambiguating each sense of หัว /hua4/. The results will show the optimal span for disambiguating each sense of หัว /hua4/. The results will be discussed based on three perspectives, namely computational, syntactic and semantic perspectives. However, the results of disambiguating three senses namely, "talent", "heading" and "head of coin" will not be presented because their occurrences in the training data are too low. Thus, they could not be used as representatives of these senses. The results of disambiguating those remaining 17 senses will be presented and discussed in according to the optimal span for the disambiguation as follows.

4.1.1 The results with the optimal span as one

There are fourteen senses presented here that need only one word or immediately adjacent word for the disambiguation. This is because, from the decision lists at 1W span, the immediately adjacent words of หัว /hua4/ are mostly content words including noun, adjective, and verb, which have some semantic relationship with the word หัว /hua4/. This is in accordance with our hypothesis that, in Thai, the content word is usually immediately adjacent to another content word without any

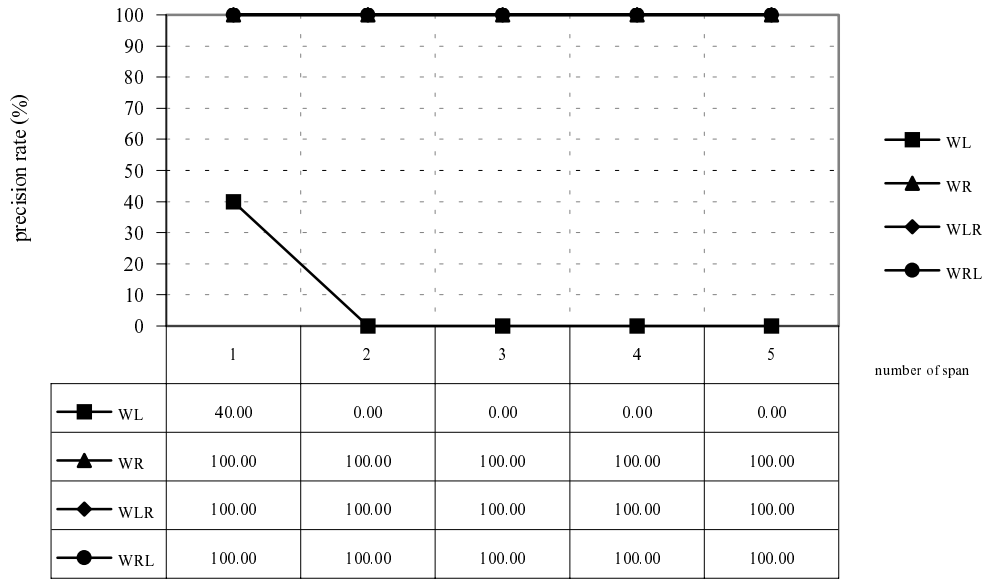
function word in between. These content words add some meanings to another content word it co-occurred with.

Within these fourteen senses, there are nine senses that the sense indicators are on the right side, two senses that are on the left side, and three senses that are on both right and left sides. The detail presentations and explanations are as follows.

4.1.1.1 The sense indicator are on the right side

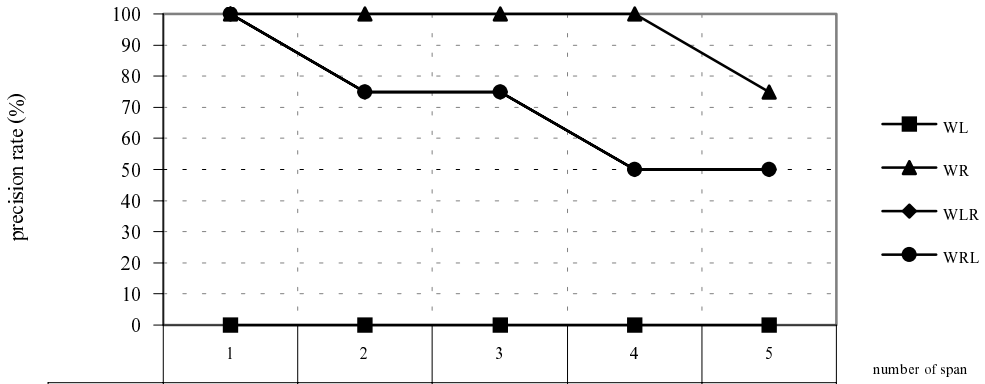
There are nine senses that the sense indicators are on the right side, namely "chief", "emotion", "machine part", "early hours", "hair", "intelligence", "top", "bulb" and "topics". In fact, WL of หัว /hua4/ play no role at all in indicating the sense “emotion” and plays very little role on disambiguating the other eight senses.

Figure 9 to figure 17 present the results on disambiguating these nine senses of หัว /hua4/ that the sense indicators are on the right side.



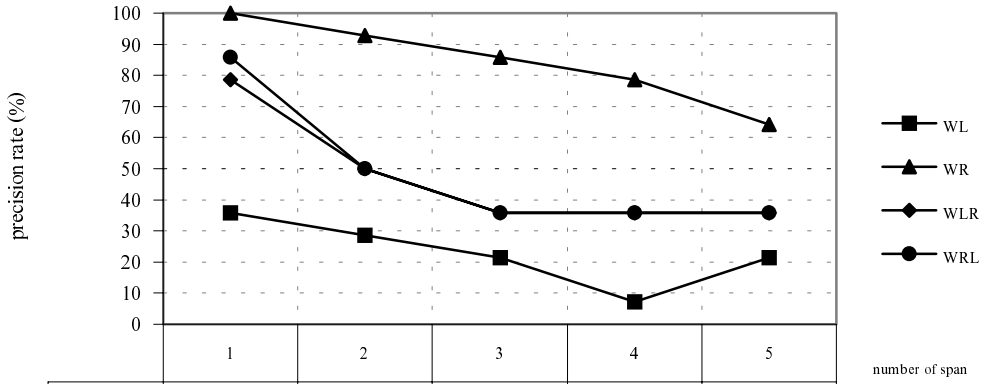
	1	2	3	4	5
WL	40.00	0.00	0.00	0.00	0.00
WR	100.00	100.00	100.00	100.00	100.00
WLR	100.00	100.00	100.00	100.00	100.00
WRL	100.00	100.00	100.00	100.00	100.00

Figure 9: Results on disambiguating หัว /hua4/ "chief".



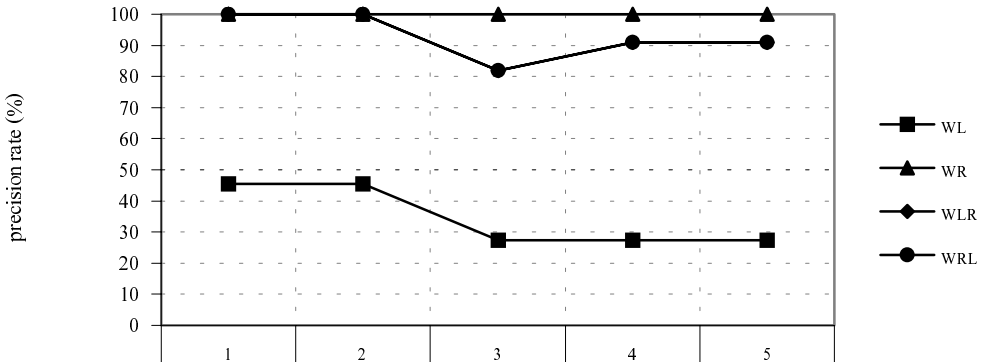
	1	2	3	4	5
WL	0.00	0.00	0.00	0.00	0.00
WR	100.00	100.00	100.00	100.00	75.00
WLR	100.00	75.00	75.00	50.00	50.00
WRL	100.00	75.00	75.00	50.00	50.00

Figure 10: Results on disambiguating หัว /hua4/ "emotion".



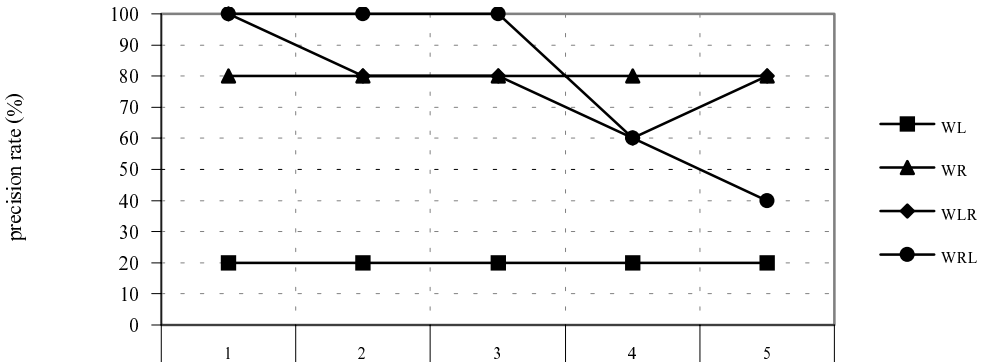
	1	2	3	4	5
WL	35.71	28.57	21.43	7.14	21.43
WR	100.00	92.86	85.71	78.57	64.29
WLR	78.57	50.00	35.71	35.71	35.71
WRL	85.71	50.00	35.71	35.71	35.71

Figure 11: Results on disambiguating หัว /hua4/ "machine part".



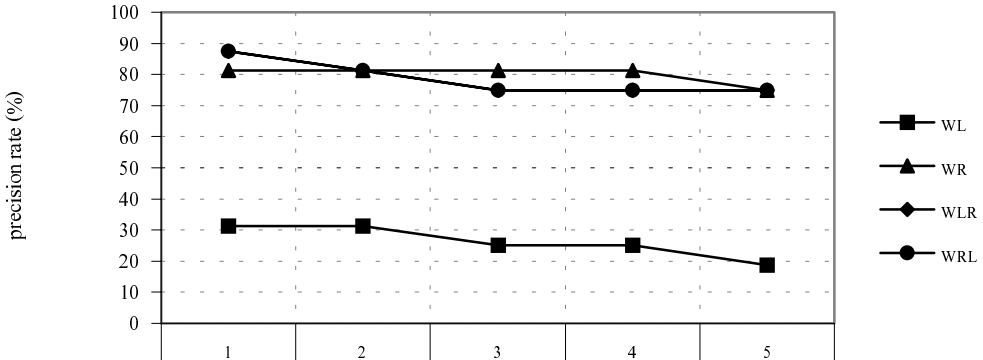
	1	2	3	4	5
WL	45.45	45.45	27.27	27.27	27.27
WR	100.00	100.00	100.00	100.00	100.00
WLR	100.00	100.00	81.82	90.91	90.91
WRL	100.00	100.00	81.82	90.91	90.91

Figure 12: Results on disambiguating หัว /hua4/ "early hours".



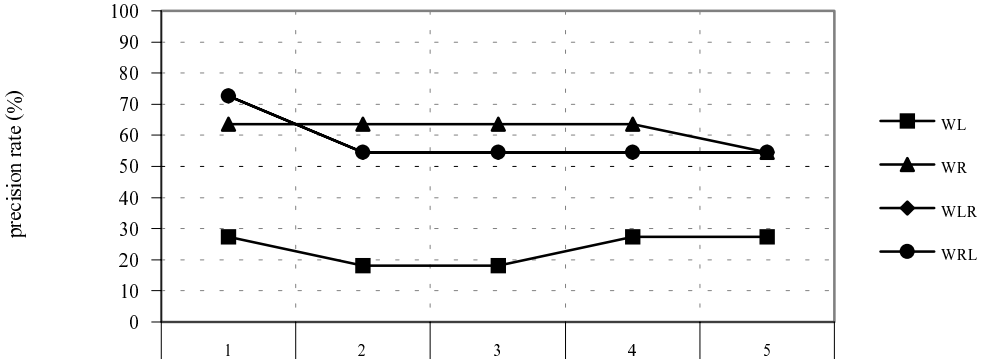
	1	2	3	4	5
WL	20.00	20.00	20.00	20.00	20.00
WR	80.00	80.00	80.00	80.00	80.00
WLR	100.00	80.00	80.00	60.00	80.00
WRL	100.00	100.00	100.00	60.00	40.00

Figure 13: Results on disambiguating หัว /hua4/ "hair".



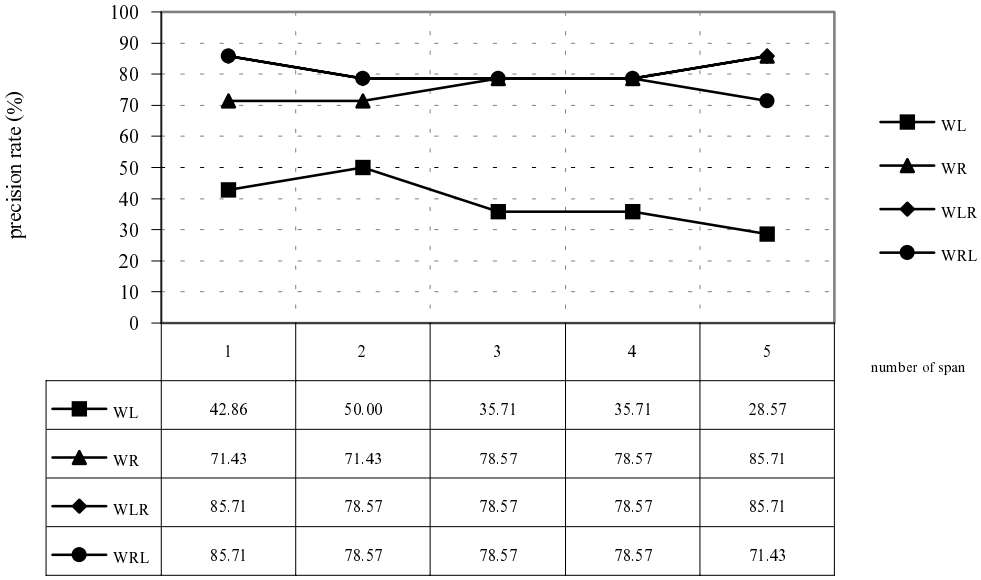
	1	2	3	4	5
WL	31.25	31.25	25.00	25.00	18.75
WR	81.25	81.25	81.25	81.25	75.00
WLR	87.50	81.25	75.00	75.00	75.00
WRL	87.50	81.25	75.00	75.00	75.00

Figure 14: Results on disambiguating หัว /hua4/ "intelligence".



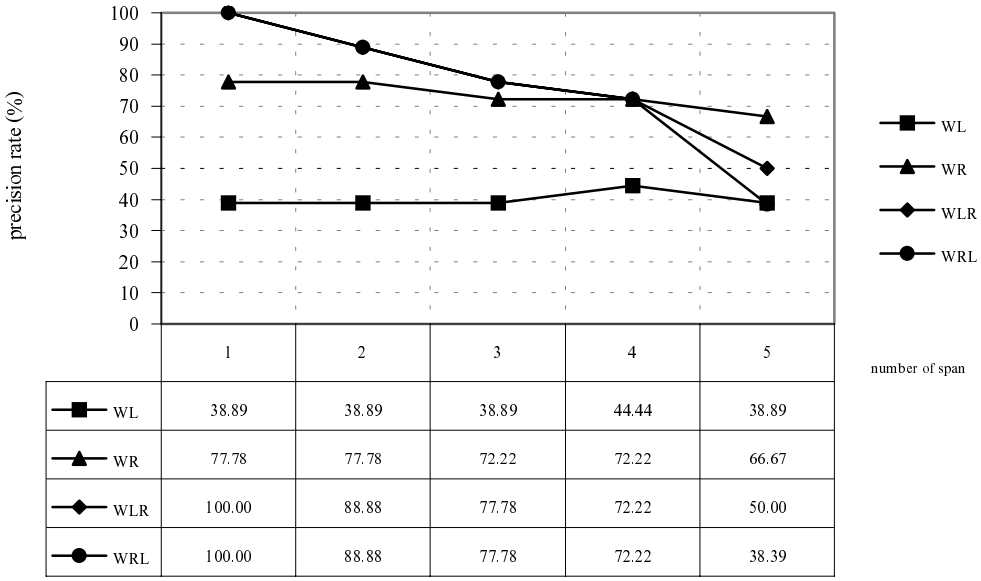
	1	2	3	4	5
WL	27.27	18.18	18.18	27.27	27.27
WR	63.64	63.64	63.64	63.64	54.55
WLR	72.73	54.55	54.55	54.55	54.55
WRL	72.73	54.55	54.55	54.55	54.55

Figure 15: Results on disambiguating หัว /hua4/ "top".



	1	2	3	4	5
WL	42.86	50.00	35.71	35.71	28.57
WR	71.43	71.43	78.57	78.57	85.71
WLR	85.71	78.57	78.57	78.57	85.71
WRL	85.71	78.57	78.57	78.57	71.43

Figure 16: Results on disambiguating หัว /hua4/ "bulb".



	1	2	3	4	5
WL	38.89	38.89	38.89	44.44	38.89
WR	77.78	77.78	72.22	72.22	66.67
WLR	100.00	88.88	77.78	72.22	50.00
WRL	100.00	88.88	77.78	72.22	38.39

Figure 17: Results on disambiguating หัว /hua4/ "topics".

Finding the sense indicators of หัว /hua4/ at the right side is in accordance with our hypothesis, as we expect that the modifiers of noun หัว /hua4/ are mostly found at the right side. This can be explained in details as follows.

- Computational explanation

In term of computational, sense indicators of these nine senses are on the right side because, from the decision lists of words at the right side, there are many collocational word forms co-occur with these senses. Their co-occurrences are significant as they have high collocational weights. This statistical evidence indicates that the optimal side for the disambiguation of หัว /hua4/ is the right side.

- Syntactic explanation

The statistical evidence is also in accordance with the explanation of the syntactic structure of Thai language, which the head and modifier relationship plays dominant role in the disambiguation. This relationship can be explained as follows.

From the decision lists of words at the right side of หัว /hua4/, collocations can be grouped according to their parts of speech. Their syntactic patterns are shown as follows.

Pattern (1): Head and modifier relationship.

NP

head	modifier
N หัว	N; ADJ; V

In pattern (1), *หัว* /hua4/ is the head of NP, which can have another noun, a verb or an adjective as its modifier. The modifier adds some meaning to the head.

Example 1 is the examples of collocational patterns of *หัว* /hua4/ meaning "chief", "emotion", "machine part", "early hours", "hair", "intelligence", "top", "bulb" and "topics" respectively, which are in accordance with the relationship in the pattern (1).

Example 1:

(i) NP(Head: N(*หัว*)),(Mod: N(*โจก*))

(ii) NP(Head: N(*หัว*)),(Mod: ADJ(*เสีย*))

(iii) NP(Head: N(*หัว*)),(Mod: V(*รบ*))

(iv) NP(Head: N(*หัว*)),(Mod: N(*ค่ำ*))

(v) NP(Head: N(*หัว*)),(Mod: ADJ(*ดำ*))

(vi) NP(Head: N(*หัว*)),(Mod: ADJ(*ใส*))

(vii) NP(Head: N(*หัว*)),(Mod: N(*ไม้ขีดไฟ*))

(viii) NP(Head: N(*หัว*)),(Mod: N(*หอม*))

(ix) NP(Head: N(*หัว*)),(Mod: N(*เรื่อง*))

- Semantic explanation

Beside syntactic relationships, there are also semantic relationship between *หัว* /hua4/ and word forms on the right side. From the decision lists of words at the right side of *หัว* /hua4/, words co-occur with each sense of *หัว* /hua4/ can be grouped according to their semantic fields, with different fields indicate different senses. There seems to be a coherence between the semantic fields of the word *หัว* /hua4/ and its sense indicators, as shown below.

Senses	Semantic fields of co-occurred word forms	Examples of co-occurred word forms
Chief	"leader" (inconclusive)	*โจก*
Emotion	"bad" (inconclusive)	*เสีย*
Machine part	OPERATION VERB	*จ่าย รบ อ่าน ฟัง ลาก*
Early hours	HOUR	*ค่ำ*
Hair	COLOR	*ดำ แดง*
Intelligence	ATTRIBUTE	*ใส ไว ดี*
Top	THING	*ไม้ กระสุน นม ไม้ขีดไฟ*
Bulb	PLANT	*หอม ผัก มัน เผือก*
Topics	DISCOURSE	*เรื่อง*

Table 16: Semantic relationship between *หัว* /hua4/ and words at 1WR.

The relationship of these nine senses of *หัว* /hua4/ and semantic fields of their co-occurred word forms on the right side can be explained according to the semantic network (in figure 3, section 3.2) which presents the relationship of *หัว* /hua4/ and other concepts as follows.

หัว /hua4/ meaning "machine part" is a part of a machine that perform some chief operation and it often co-occurs with some OPERATION VERB. For example, an OPERATION VERB, *ลาก* /laak2/ "to trail, pull" co-occurs with *หัว* /hua4/ as in *หัวลาก* indicates that *หัว* /hua4/ means "machine part" and *หัวลาก* means "part which pulls the rest of a machine".

As for *หัว* /hua4/ meaning "early hours", there seems to be a coherence of semantic fields, TIME PERIOD, between this word and its sense indicator. As we can see that *หัว* /hua4/ meaning "early hours" is a part of HOUR or TIME PERIOD, and *ค่ำ* /kham2/ "dark, night", for example, is also a TIME PERIOD. Thus, *ค่ำ* /kham2/ "dark, night", when co-occurs with *หัว* /hua4/, as in *หัวค่ำ*, indicates the sense of *หัว* /hua4/ as "early hours".

The rest can be explained in the similar way. For examples, *แดง* /dɛɛŋ/ "red" indicates the sense of *หัว* /hua4/ "hair" because *แดง* /dɛɛŋ/ "red" is a COLOR and *หัว* /hua4/ "hair" is a part of "head" which can have COLOR or style or can be colored. *ไว* /wai/ "quick" indicates the sense of *หัว* /hua4/ as "intelligence" because *หัว* /hua4/ meaning "intelligence" often co-occurs with some ATTRIBUTE, which are words used to describe a characteristic of "intelligence" and *ไว* /wai/ "quick" is an ATTRIBUTE. *ไม้ขีดไฟ* /maai3khiit1fai/ "match" indicates the sense of *หัว* /hua4/ as “top” because "top" is a part of OBJECT or THING, and *ไม้ขีดไฟ* /maai3khiit1fai/ "match", is also an OBJECT. *เผือก* /phɯɯak1/ "taro" indicates the sense of *หัว* /hua4/ as “bulb” because "bulb" is a part of some PLANT, and เผือก /phɯɯak1/ "taro" is a PLANT.

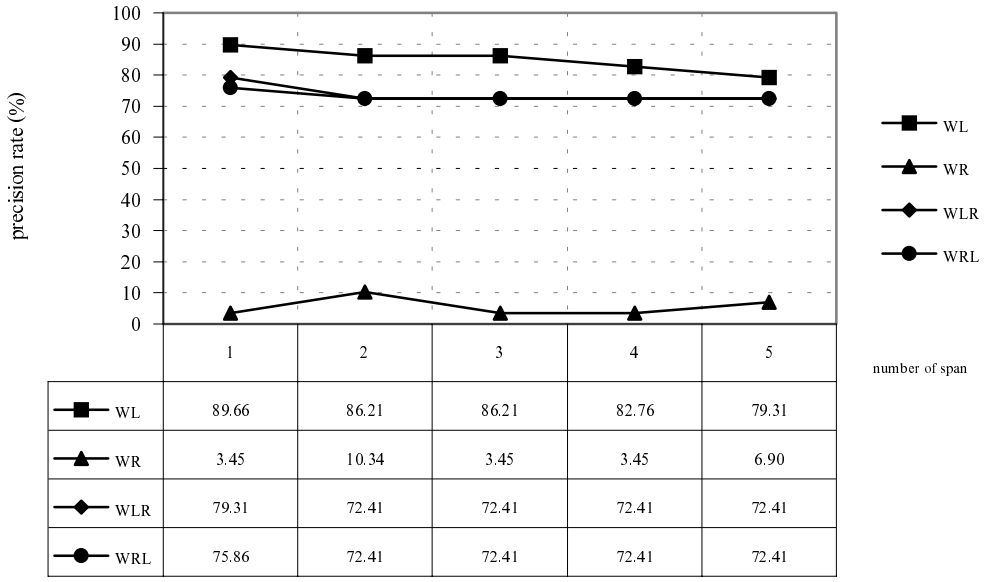
As for *หัว* /hua4/ "emotion" and *หัว* /hua4/ "chief", since there is only one word co-occurred with them, namely *เสีย* /siia4/ "bad" and *โจก* /cook1/ "leader" respectively, the semantic relationship between them is inconclusive.

The collocational patterns presented above are in the form of the parts of speech and semantic fields. However, the word forms that are sense indicators for each sense of *หัว* /hua4/ can be different. Words co-occurred with these nine senses of *หัว* /hua4/ at the right and the left side are shown in the appendix E, from table 1 to table 17, for sense "chief", "emotion", "machine part", "early hours", "hair", "intelligence", "top", "bulb" and "topics" respectively.

4.1.1.2 The sense indicators are on the left side

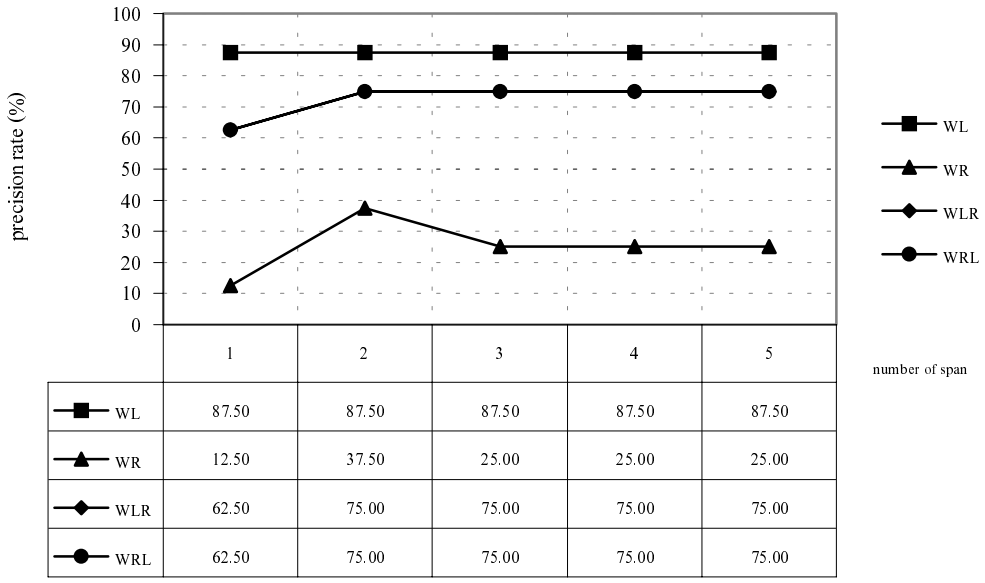
There are two senses that sense indicators are located at the left side, namely "brain" and "headline". Words on the right side play very little role for disambiguating these two senses.

Figures 18 and 19 present the results on disambiguating these two senses of หัว /hua4/.



	1	2	3	4	5
WL	89.66	86.21	86.21	82.76	79.31
WR	3.45	10.34	3.45	3.45	6.90
WLR	79.31	72.41	72.41	72.41	72.41
WRL	75.86	72.41	72.41	72.41	72.41

Figure 18: Results on disambiguating หัว /hua4/ "brain".



	1	2	3	4	5
WL	87.50	87.50	87.50	87.50	87.50
WR	12.50	37.50	25.00	25.00	25.00
WLR	62.50	75.00	75.00	75.00	75.00
WRL	62.50	75.00	75.00	75.00	75.00

Figure 19: Results on disambiguating หัว /hua4/ "headline".

Finding sense indicators on the left side of หัว /hua4/ is opposed to our hypothesis, as we do not expect that any modifier of noun หัว /hua4/ will be at the left side. The reasons can be explained as follows.

- Computational explanation

The left is the optimal side for the disambiguation of these two senses because, from the decision list at the left side, there are many collocational word forms co-occurred with these senses. Their co-occurrences are significant as they have high collocational weights.

- Syntactic explanation

From the decision lists at the left side, we found the following collocational pattern, which has syntactic pattern as follow.

Pattern (2): Verb and object relationship

S

verb	object
V	N หัว

In pattern (2) หัว /hua4/ is the object of the preceding verb. In this case, selectional restriction between the verb and its object is the reason why the preceding verb can act as a sense indicator of หัว /hua4/. Example 2 is the examples of

collcoational patterns of *หัว* /hua4/ meaning "brain" and "headline", respectively, which are in accordance with the relationship in the pattern (2).

Example 2:

(i) S(Verb: V(*ปวด*)),(Object: N(*หัว*))

(ii) S(Verb :V(*พาด*)), (Object: N(*หัว*))

- Semantic explanation

From the decision lists of words at the left side of *หัว* /hua4/, words co-occur with each sense of *หัว* /hua4/ can be grouped according to their semantics fields, with different fields indicates different senses. These semantic fields of word forms co-occur with different senses of *หัว* /hua4/ are shown as follows.

Senses	Semantic fields of co-occurred word forms	Examples of co-occurred word forms
Brain	PHYSICAL STATE	*ปวด มึน เวียน*
Headline	"to headline" (inconclusive)	*พาด*

Table 17: Semantic relationship between *หัว* /hua4/ and words at 1WL.

The semantic relationship between *หัว* /hua4/ meaning "brain" and its sense indicators can be explained in terms of semantic coherence above. As *หัว* /hua4/ “brain” is a physical part in "head" which can have some physical state as *ปวด* /puuat1/ "pain, ache", thus, *ปวด* /puuat1/, which is a PHYSICAL STATE indicates that *หัว* /hua4/ means "brain".

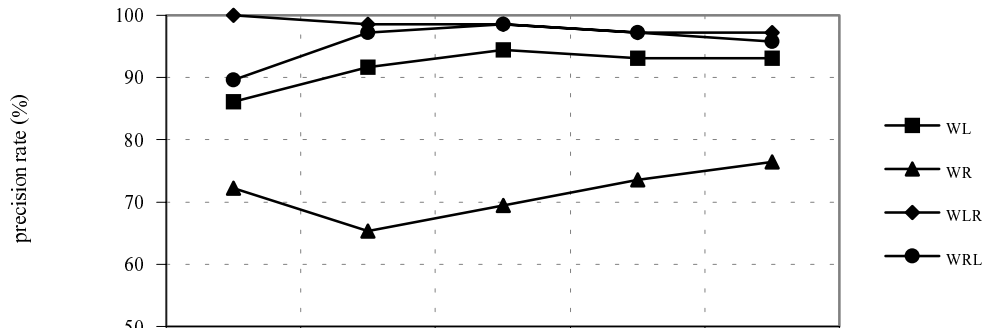
As for *หัว* /hua4/ meaning "headline", since there is only one word form, which is *พาด* /phaat2/ "to headline", that co-occurs with it, the semantic relationship is inconclusive.

The examples of word forms that co-occur with these two senses are in the appendix E, from table 18 to table 20, for senses "brain" and "headline" respectively.

4.1.1.3 The sense indicators are on both right and left side

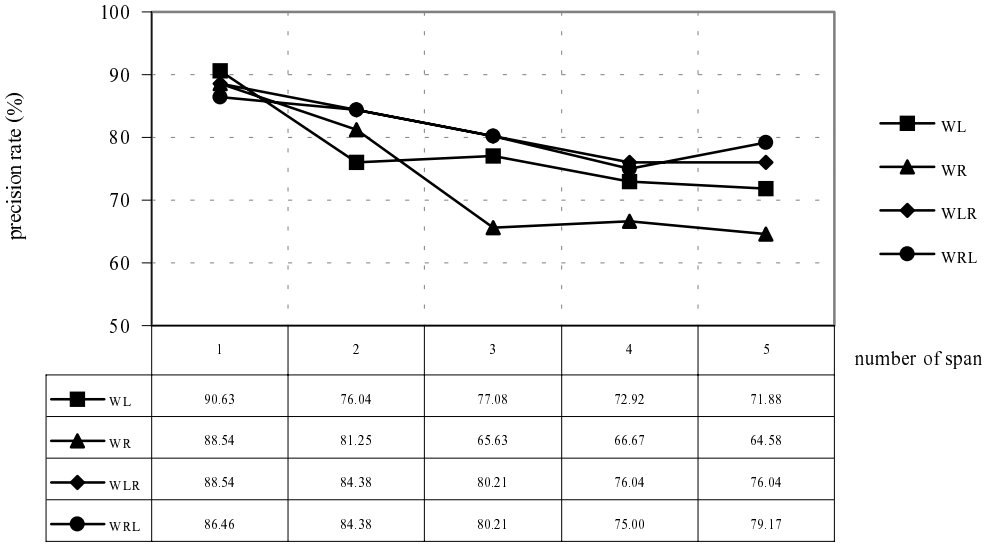
There are three senses that the sense indicators are on both right and left sides namely, "head", "entity", and "titles or names".

Figures 20 to 22 present the results of disambiguation these three senses of หัว /hua4/.



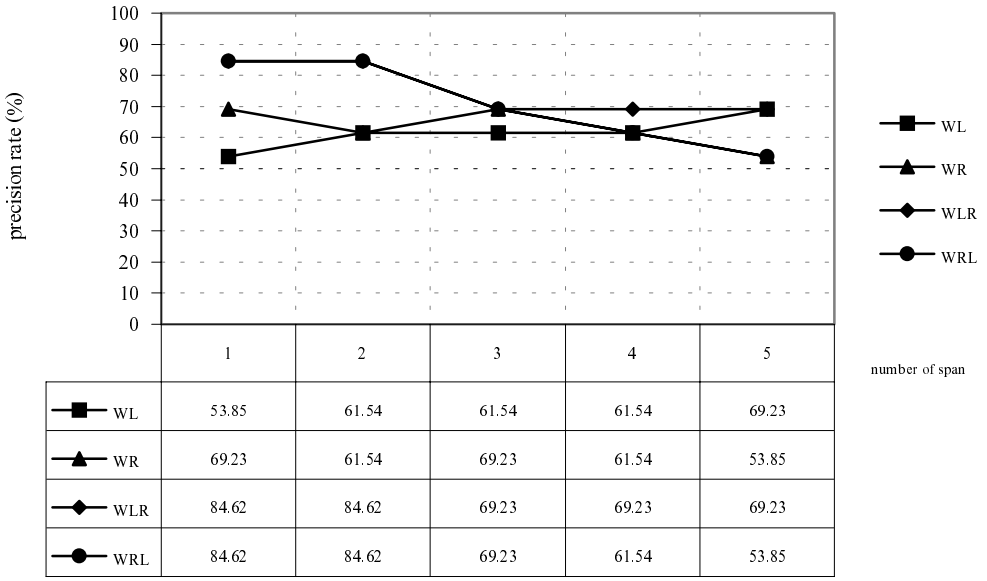
	1	2	3	4	5
WL	86.11	91.67	94.44	93.06	93.06
WR	72.22	65.28	69.44	73.61	76.39
WLR	100.00	98.61	98.61	97.22	97.22
WRL	89.61	97.22	98.61	97.22	95.83

Figure 20: Results on disambiguating หัว /hua4/ "head".



	1	2	3	4	5
WL	90.63	76.04	77.08	72.92	71.88
WR	88.54	81.25	65.63	66.67	64.58
WLR	88.54	84.38	80.21	76.04	76.04
WRL	86.46	84.38	80.21	75.00	79.17

Figure 21: Results on disambiguating หัว /hua4/ "entity".



	1	2	3	4	5
WL	53.85	61.54	61.54	61.54	69.23
WR	69.23	61.54	69.23	61.54	53.85
WLR	84.62	84.62	69.23	69.23	69.23
WRL	84.62	84.62	69.23	61.54	53.85

Figure 22: Results on disambiguating หัว /hua4/ "titles or names".

Finding sense indicators on both right and left sides is contradict to our hypothesis as we expect that only WR is sufficient information for the disambiguation.

- Computational explanation

Right-and-left is the optimal side for the disambiguation of these three senses because, the number of collocational word forms at both right and left sides is almost equal. Their co-occurrences are significant as they have high collocational weights.

- Syntactic explanation

Sense indicators are on both right and left side because both pattern (1), in which sense indicators are on the right side, and pattern (2), in which sense indicators are on the left side, play almost equal role in the disambiguation.

Example 3 is the examples of collcoational patterns of *หัว* /hua4/ meaning "head", "entity" and "titles or names", respectively, which are in accordance with the relationship in the pattern (1).

Example 3:

(i) NP(Head: N(*หัว*)),(Mod: N(*สัตว์*))

(ii) NP(Head: N(*หัว*)),(Mod: N(*ประชาชนชน*))

(iii) NP(Head: N(*หัว*)),(Mod: N(*หนังสือ*))

Example 4, (i) is the example of collocational pattern of *หัว* /hua4/ "entity" and (ii), (iii) are the examples of collocational patterns of *หัว* /hua4/ "head" which are in accordance with the relationship in the pattern (2).

Example 4:

(i) S(Verb: V(*รวม*)),(Object: N(*หัว*))

(ii) S(Verb: V(*ตี*)),(Object: N(*หัว*))

(iii) S(Verb: V(*ก้ม*)),(Object: N(*หัว*))

In addition to pattern (1), from the decision lists of *หัว* /hua4/ meaning "head", we also found another pattern that sense indicators are on the right side. This pattern is shown as follows.

Pattern (3): Subject and verb relationship.

S

subject	verb
N *หัว*	V

In pattern (3), *หัว* /hua4/ acts as the subject of the verb. The verb at the right side can act as the sense indicator of *หัว* /hua4/ because there are some selectional restrictions between the subject and the verb.

Example 5 is the examples of collocatioal patterns of this sense of *หัว* /hua4/ which are in accordance with the relationship in the pattern (3).

Example 5:

(i) S(Subject: N(*หัว*)),(Verb: V(*แตก*))

(ii) S(Subject: N(*หัว*)), Verb: V(*กระแทก*))

(iii) S(Subject: N(*หัว*)), Verb: V(*ชน*))

This pattern is not what we have in mind when setting the hypothesis that the sense indicator of *หัว* /hua4/ should be on the right side, though it is not opposed to the hypothesis.

In addition to pattern (2), from the decision lists of these three senses of *หัว* /hua4/ we also found another pattern that sense indicators are on the left side. This pattern is shown as follows.

Patterns (4): Head and modifier relationship

NP

head	modifier
N	N *หัว*

In pattern (4), *หัว* /hua4/ itself acts as the modifier of its head verb or noun. Since there are selectional restriction constraints between the head and the modifier *หัว* /hua4/, the head can act as a sense indicator of *หัว* /hua4/ in these cases.

Example 6 is the examples of collocational patterns of *หัว* /hua4/ meaning "titles or names", "head" and "entity" respectively, which are in accordance with the relationship in the pattern (4).

Example 6:

(i) NP(Head: N(*นิตยสาร*)),(Mod: N(*หัว*))

(ii) NP(Head: N(*กุนซือ*)),(Mod: N(*หัว*))

(iii) NP(Head: N(*ค่า*)),(Mod: N(*หัว*))

There are also many cases that both WL and WR help disambiguating in the same context. For example, in example 7, both *กุนซือ* /kunsɯɯ/ ɪ"coach", and *โล้น* /loan3/ "bald" occur in the same context and are in the same construction, which both help disambiguating *หัว* /hua4/ meaning "head".

Example 7:

NP

head		modifier	
N		NP	
	head		modifier
	N		ADJ
กุนซือ	หัว		โล้น

- Semantic explanation

From the decision lists of words at the right side of *หัว* /hua4/, words co-occur with each sense of *หัว* /hua4/ can be grouped according to their semantics fields, with different fields indicates different senses. These semantic fields are shown as follows.

Senses	Semantic fields of co-occurred word forms	Examples of co-occurred word forms
Head	HUMAN or ANIMAL	*คน สัตว์ หมู งู นก เขา ไก่*
Entity	HUMAN or ORGANIZATION	*ประชาชน แรงงาน นักศึกษา*
Titles or names	PRINTED MATERIAL	*หนังสือ*

Table 18: Semantic relationship between *หัว* /hua4/ and right words at 1WRL.

The explanation of the semantic relationship can be explained again by the semantic coherence between the word *หัว* /hua4/ and its sense indicators. For example, *คน* /khon/ "human" indicates the sense of *หัว* /hua4/ as "head" because *คน* /khon/ is a HUMAN, and “head” is a part of HUMAN or ANIMAL. As for *หัว* /hua4/ “entity”, since HUMAN or ORGANIZATION is an "entity", *ประชาชน* /pra1chaachon/ "people or group of people" indicates the sense “entity” because *ประชาชน* /pra1chaachon/ is a HUMAN. *หนังสือ* /naŋ4sɯɯ4/ indicates the sense of *หัว* /hua4/ as “titles or names” because *หนังสือ* /naŋ4sɯɯ4/ is a PRINTED MATERIAL, and PRINTED MATERIAL has title or name.

In addition, from the decision lists of words at the left side of *หัว* /hua4/, words co-occur with each sense of *หัว* /hua4/ can be grouped according to their semantic fields, with different fields indicates different senses. These semantic fields are shown as follows.

Senses	Semantic fields of co-occurred word forms	Examples of co-occurred word forms
Head	ACTION VERB	*ตี ทุบ เด็ด สั่น ตบ ยิง ส่าย*
Entity	Inconclusive	*ค่า รวม หมาย เก็บ หาย ต่อ*
Titles or names	PRINTED MATERIAL	*หนังสือ นิตยสาร หนังสือพิมพ์*

Table 19: Semantic relationship between *หัว* /hua4/ and left words at 1WRL.

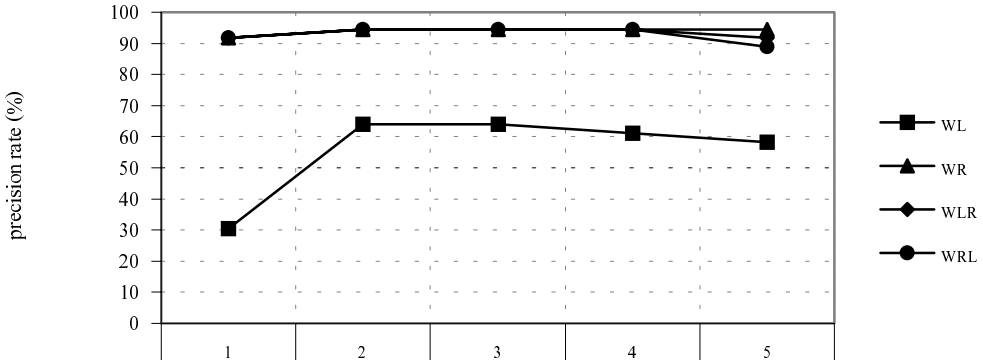
Again, the semantic relationship between *หัว* /hua4/ and its sense indicators on the left can be explained in terms of semantic coherence. For example, *ตี* /tii/ "hit" indicates the sense "head" because "head" is a part of human and "hit" is an action being done on someone. However, for *หัว* /hua4/ meaning "entity", since there are many word forms, which are in many different semantic fields, that co-occur with it, the semantic relationship between them is inconclusive.

The lists of word forms used as sense indicators for these three senses, namely "head", "entity" and "titles or names", are shown in the appendix E, from table 21 to table 26 respectively.

4.1.2 The results with the optimal span higher than one

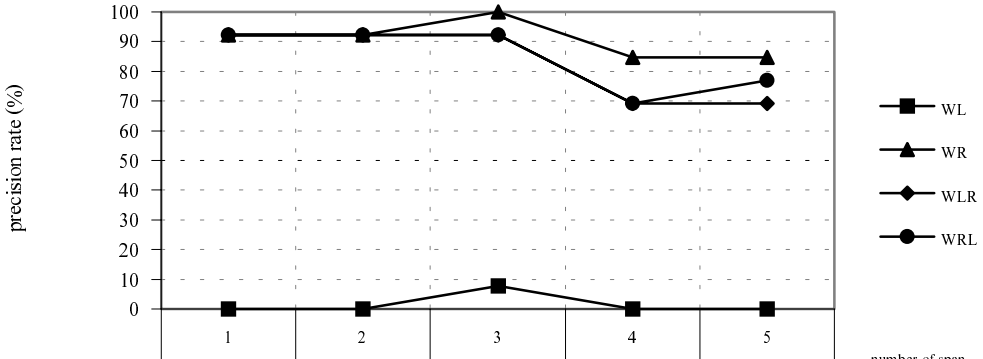
There are three senses that need more than 1W span for the disambiguation. *หัว* /hua4/ in the senses of "viewpoint" and "front" need two words for the disambiguation, *หัว* /hua4/ meaning "concentrate" needs three words for the disambiguation.

Figure 23 to 25 present the results on disambiguating three senses of *หัว* /hua4/ that the optimal span is higher than one and the sense indicators are on the right side.



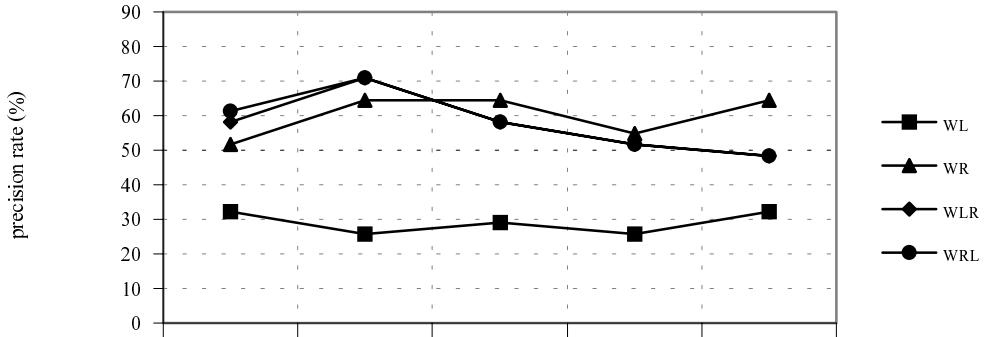
	1	2	3	4	5
WL	30.55	63.89	63.89	61.11	58.33
WR	91.67	94.44	94.44	94.44	94.44
WLR	91.67	94.44	94.44	94.44	91.67
WRL	91.67	94.44	94.44	94.44	88.89

Figure 23: Results on disambiguating หัว /hua4/ "viewpoint".



	1	2	3	4	5
WL	0.00	0.00	7.69	0.00	0.00
WR	92.31	92.31	100.00	84.62	84.62
WLR	92.31	92.31	92.31	69.23	69.23
WRL	92.31	92.31	92.31	69.23	76.92

Figure 24: Results on disambiguating หัว /hua4/ "concentrate".



	1	2	3	4	5
WL	32.26	25.81	29.03	25.81	32.26
WR	51.61	64.52	64.52	54.84	64.52
WLR	58.06	70.97	58.06	51.61	48.39
WRL	61.29	70.97	58.06	51.61	48.39

Figure 25: Results on disambiguating หัว /hua4/ "front".

In terms of side, the sense indicators of these three senses are on the right sides, which is in accordance with our hypothesis. WL play no role in disambiguating หัว /hua4/ in the sense of "concentrate" and plays little role in disambiguating หัว /hua4/ in the senses of "viewpoint" and "front". The sense indicators of these three senses are on the right side can be explained in the same ways as those already discussed in section 4.1.1.1, which are as follows.

- Computational explanation

There are many collocational word forms co-occur with these senses on the right side and their co-occurrences are significant as they have high collocational weights. This statistical evidence indicates that the optimal side for the disambiguation of these three senses of หัว /hua4/ is the right side.

- Syntactic explanation

Example 8 is the examples of collocational patterns of *หัว* /hua4/ meaning "viewpoint", "front" and "concentrate" respectively, which are in accordance with the relationship in the pattern (1). Although they are not immediately adjacent to *หัว* /hua4/, they function as a modifier of the head noun *หัว* /hua4/.

Example 8:

(x) NP(Head: N(*หัว*)),(Mod: ADJ(*แข็ง*))

(xi) NP(Head: N(*หัว*)),(Mod: ADJ(*กะทิ*))

(xii) NP(Head: N(*หัว*)),(Mod: V(*เตียง*))

- Semantic explanation

From the decision lists of words at the right side of *หัว* /hua4/, collocations can be grouped according to their semantics fields, with different fields indicates different senses. These semantic fields are shown in the table below.

Senses	Semantic fields of co-occurred word forms	Examples of co-occurred word forms
Viewpoint	ATTRIBUTE	*แข็ง อ่อน รุนแรง รั้น เก่า สมัยใหม่*
Front	THING	*รถ เรือ จักร เพลง*
Concentrate	THING or LIQUID	*เชื้อ กะทิ น้ำหอม น้ำอัดลม พันธุ์*

Table 20: Semantic relationship between *หัว* /hua4/ and words at 2WR and 3WR.

The semantic relationship between *หัว* /hua4/ and its sense indicator can be explained in the same way in terms of semantic coherence between them. For

examples, *เก่า* /kao1/ "old-fashioned" indicates the sense "viewpoint" because “viewpoint” often co-occurs with some ATTRIBUTE used to describe a characteristic of "viewpoint", and "old-fashioned" is an ATTRIBUTE. *รถ* /rot3/ "car" indicates the sense "front" because “front” is a part of OBJECT or THING, and "car" is an OBJECT. *กะทิ* /ka1thi3/ indicates the sense "concentrate" because "concentrate" is a part of THING or LIQUID, and *กะทิ* /ka1thi3/ "coconut cream" is a LIQUID.

The lists of word forms that co-occur with these three senses of *หัว* /hua4/ namely, "viewpoint", "concentrate" and "front" are shown in the appendix E, from table 27 to table 31, respectively.

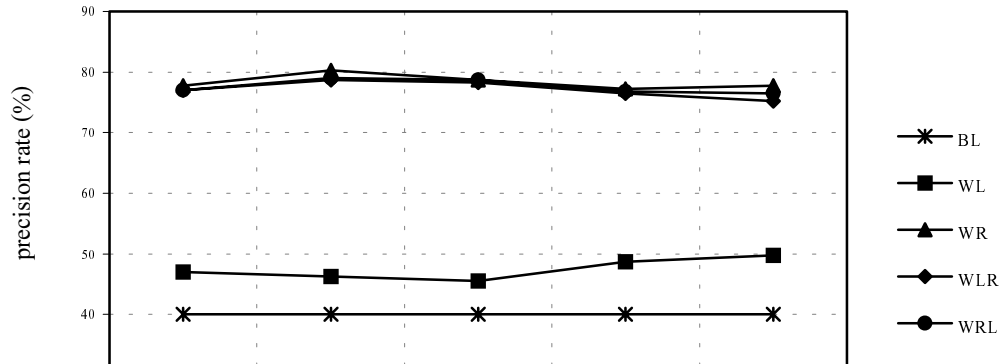
In terms of the span, the optimal span as more than 1W is opposed to our hypothesis. There are two explanations for this.

(1) The number of the training data is not large enough. For example, in "concentrate", in *หัวน้ำเชื่อมของโค๊ก*, *น้ำเชื่อม* /naam3chwwam2/ "syrup" could not be a sense indicator because it does not occur in the training data. But, instead, *โค๊ก* /cok3/ "coke" is found in the training corpus at 3W span. Thus it is used as a sense indicator during the test.

(2) Even though the size of the training data is large enough, some word forms at 1W span can co-occur with more than one sense. For example, *ปีก* /peek1/ "wing" and *ลำตัว* /lamtua/ "body" usually co-occur with *หัว* /hua4/ meaning "head". However, *ปีก* /peek1/ and *ลำตัว* /lamtua/, sometimes, also co-occur with *หัว* /hua4/ sense "front" too, as in "…ซากเรือดำน้ำเหล็ก ส่วน**หัว ลำตัว**และส่วนท้ายเรือ…" and "…สายการบินนี้ทำงานอย่างตรงไปตรงมา มีศีลธรรม ส่วน**หัว ปีก** ดัดแปลงมาจากเรือสุพรรณหงส์…". Thus, they can not be used to distinguish between these two senses.

However, from the results of the disambiguation of these senses, we can see that 1W is as good as 2W and 3W because the precision rate at 1W span is only a little bit lower than the precision rates at 2W and 3W spans. This is the important evidence indicating that 1W span is still the optimal span with 2W and 3W position give little additional information for the disambiguation. For example, when considering the raw scores of correct answering, in sense "concentrate", at 1W span, the program can correctly disambiguate 12 out of 13 tokens, and at 3W span, all 13 tokens are correctly disambiguated. In other words, the program can disambiguate only one more token at 3W span. Besides, in the decision list of these senses, there is a dominant collocational pattern of *หัว* + N, which indicates that only 1W span is sufficient.

4.2 The Results of the Disambiguation of *เก็บ* /kep1/



	1	2	3	4	5
BL	40.00	40.00	40.00	40.00	40.00
WL	47.00	46.25	45.50	48.75	49.75
WR	77.75	80.25	78.75	77.25	77.75
WLR	77.00	78.75	78.25	76.50	75.25
WRL	77.00	79.00	78.75	76.75	76.50

Figure 26: Precision rate of disambiguation of*เก็บ* /kep1/.

Figure 26 shows that the optimal span for the disambiguation of *เก็บ* /kep1/ is 2WR as indicated by the precision rate of 80.25%, 2 times higher than the lower bound

performance and only 0.8 times lower than the upper bound performance. The poorest span for the disambiguation is 3WL as indicated by the precision rate of 45.50%.

In terms of the optimal number of context words for the disambiguation, 2W span is the optimal. This is opposed to our hypothesis. However, the reason that 2W span is the optimal is not because of the influence of the second word. From figure 21, we can see that the result from 1W span is not different from that of 2W span. The second word adds a little improvement on the disambiguation. As the number of span increases to 3, 4, and 5 the precision rate decreases, but very little. In this study, we found that 3W span is sufficient. This is supported by the results on disambiguating each sense of *เก็บ* /kep1/, in which, there are 4 senses of *เก็บ* /kep1/ that have the optimal spans as one, 2 senses, namely "to hide" and "to charge" that have the optimal spans as two and only one sense, namely "to keep" that the optimal span is three.

In terms of side, the sense indicators of *เก็บ* /kep1/ for all senses is on the right side. WL play very little role in the disambiguation. This is opposed to our hypothesis as we expect that subject at the left side of the verb *เก็บ* /kep1/ plays equal role as the object at the right side. WRL[3] are also good sense indicators, however, because of the influence of WR alone. This is why the line representing WR and WRL are almost the same line.

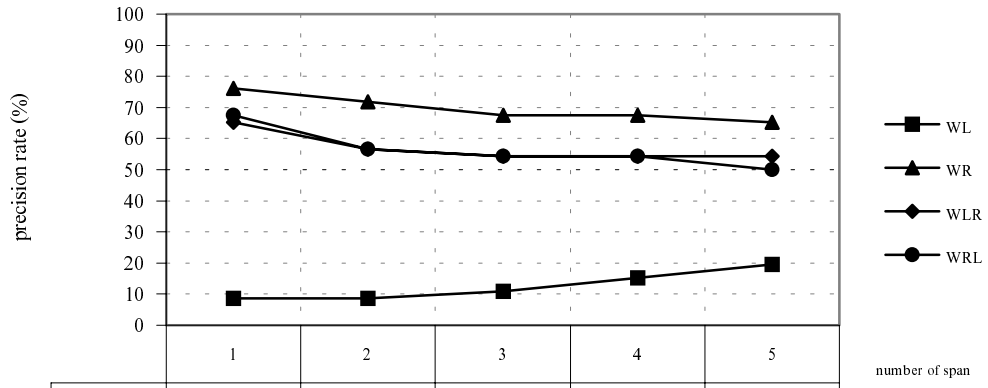
Since there is no significant difference between the precision rate of 1W, 2W and 3W spans, and only WR play role in the disambiguation, we will not present and discuss the results of disambiguating at different spans. We will present and discuss

---

[3] Since the precision rates of WLR and WRL are not significantly different, the presentation and the explanation of the results, from now on, willuse WRL only, whether, in fact, it refers to WRL or WLR.

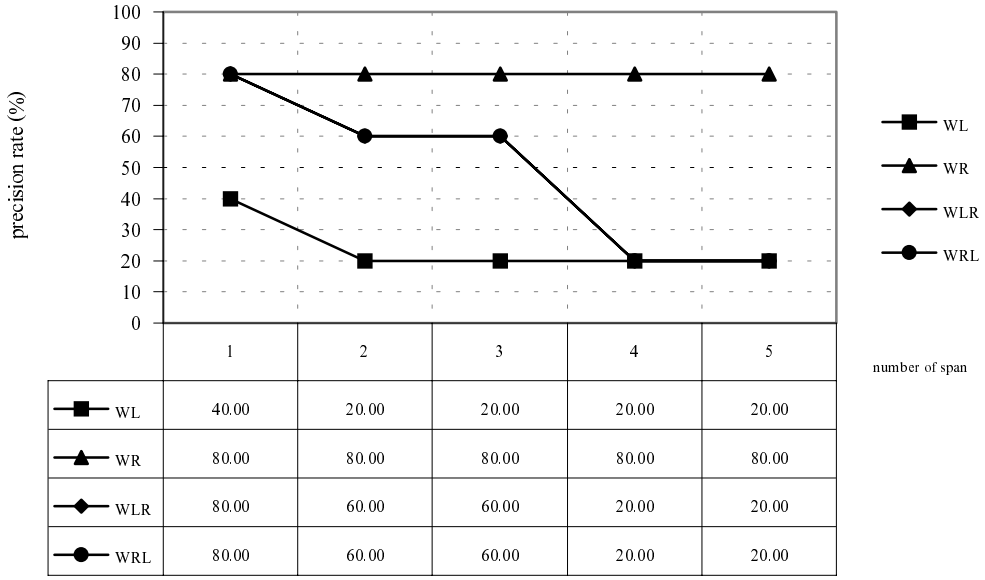
the results of each sense altogether. However, the results of the sense "to pick up" will not be presented because the training data on this sense is too small. Thus, the program cannot disambiguate this sense at all.

Figures 27 to 34 present the results on disambiguating eight senses of เก็บ /kep1/ that the sense indicators are on the right side.



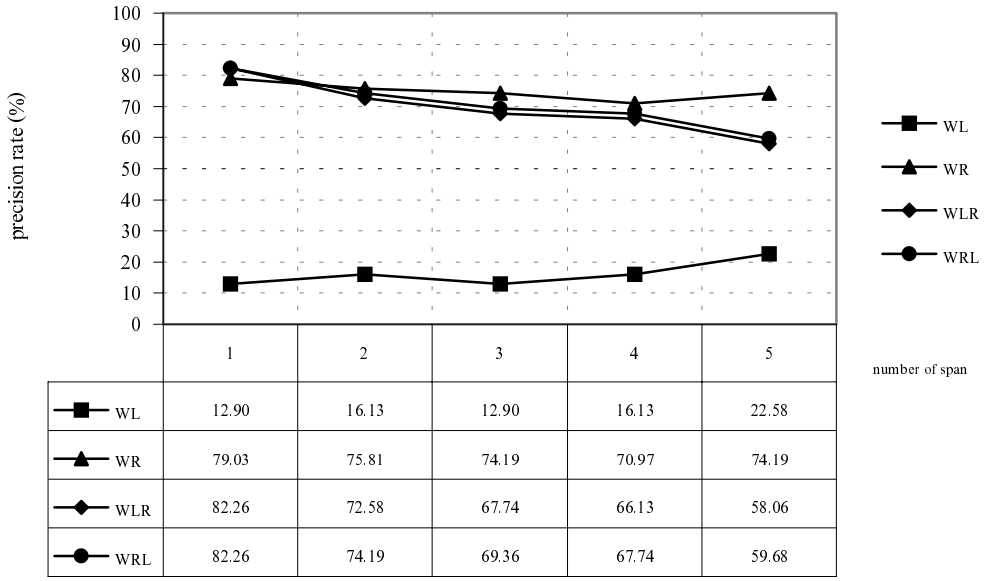
	1	2	3	4	5
WL	8.70	8.70	10.87	15.22	19.57
WR	76.09	71.74	67.39	67.39	65.22
WLR	65.22	56.52	54.35	54.35	54.35
WRL	67.40	56.52	54.35	54.35	50.00

Figure 27: Results on disambiguating เก็บ /kep1/ "to take".



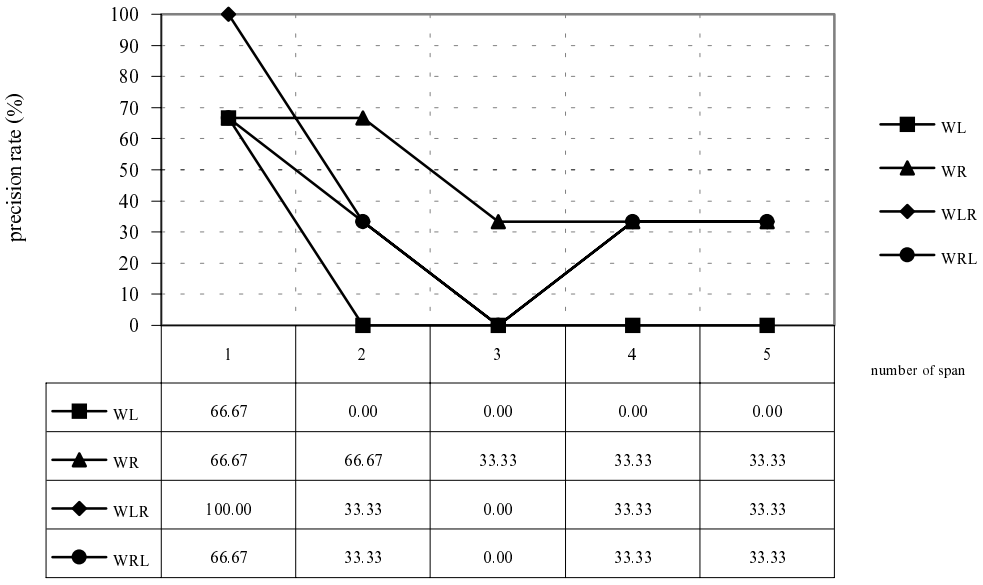
	1	2	3	4	5
WL	40.00	20.00	20.00	20.00	20.00
WR	80.00	80.00	80.00	80.00	80.00
WLR	80.00	60.00	60.00	20.00	20.00
WRL	80.00	60.00	60.00	20.00	20.00

Figure 28: Results on disambiguating เก็บ /kep1/ "to buy".



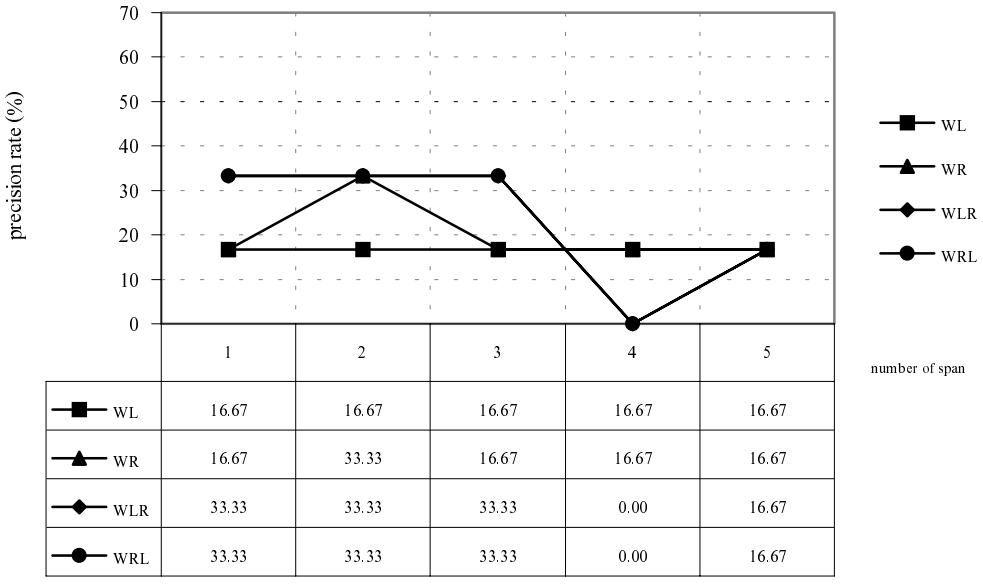
	1	2	3	4	5
WL	12.90	16.13	12.90	16.13	22.58
WR	79.03	75.81	74.19	70.97	74.19
WLR	82.26	72.58	67.74	66.13	58.06
WRL	82.26	74.19	69.36	67.74	59.68

Figure 29: Results on disambiguating เก็บ /kep1/ "to gather".



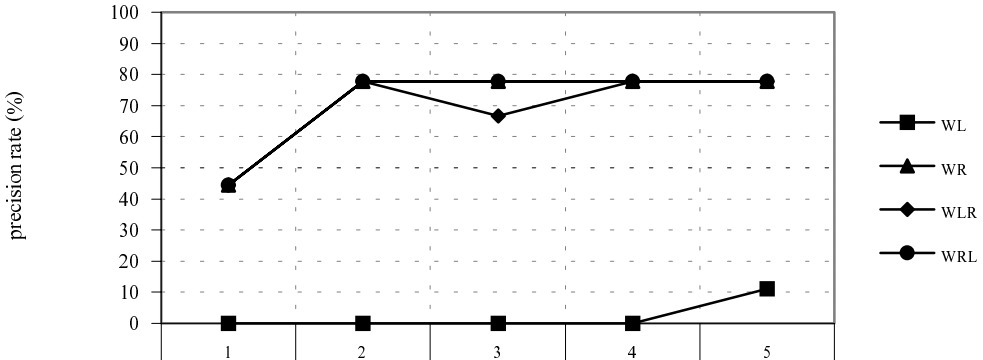
	1	2	3	4	5
WL	66.67	0.00	0.00	0.00	0.00
WR	66.67	66.67	33.33	33.33	33.33
WLR	100.00	33.33	0.00	33.33	33.33
WRL	66.67	33.33	0.00	33.33	33.33

Figure 30: Results on disambiguating เก็บ /kep1/ "to kill".



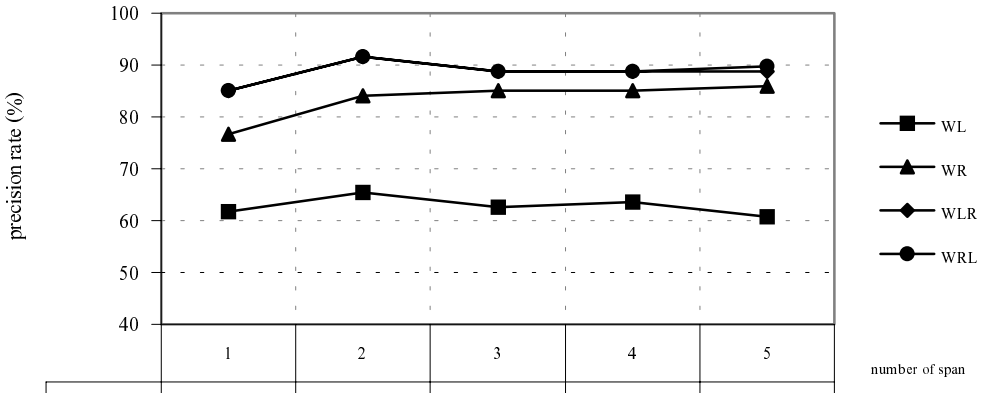
	1	2	3	4	5
WL	16.67	16.67	16.67	16.67	16.67
WR	16.67	33.33	16.67	16.67	16.67
WLR	33.33	33.33	33.33	0.00	16.67
WRL	33.33	33.33	33.33	0.00	16.67

Figure 31: Results on disambiguating เก็บ /kep1/ "to arrange".



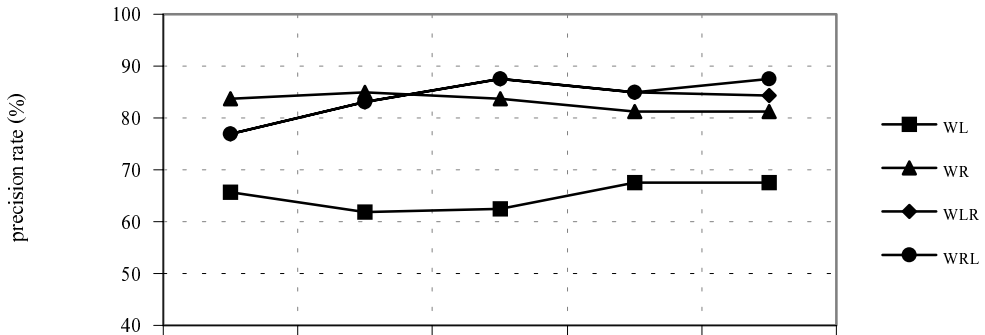
	1	2	3	4	5
WL	0.00	0.00	0.00	0.00	11.11
WR	44.44	77.78	77.78	77.78	77.78
WLR	44.44	77.78	66.67	77.78	77.78
WRL	44.44	77.78	77.78	77.78	77.78

Figure 32: Results on disambiguating เก็บ /kep1/ "to hide".



	1	2	3	4	5
WL	61.68	65.42	62.62	63.55	60.75
WR	76.64	84.11	85.05	85.05	85.98
WLR	85.05	91.59	88.79	88.79	88.79
WRL	85.05	91.59	88.79	88.79	89.72

Figure 33: Results on disambiguating เก็บ /kep1/ "to charge".



	1	2	3	4	5
WL	65.63	61.88	62.50	67.50	67.50
WR	83.75	85.00	83.75	81.25	81.25
WLR	76.88	83.13	87.50	85.00	84.38
WRL	76.88	83.13	87.50	85.00	87.50

Figure 34: Results on disambiguating เก็บ /kep1/ "to keep".

From the results of all senses of เก็บ /kep1/ presented above, we can see that only WR play role in the disambiguation and 1W is sufficient. However, as stated above, there are three senses that 2W and 3W spans are also helpful. These results can be explained in details as follows.

In terms of side, finding that only WR play role in the disambiguation is opposed to our hypothesis. The explanations are as follows.

- Computational explanation

In term of computational, a lot of words on the right side of เก็บ /kep1/ have higher collocational weight than words on the left. Their co-occurrences are significant as they have high collocational weights. This statistical evidence indicates that the optimal side for the disambiguation of เก็บ /kept1/ is the right side.

- Syntactic explanation

The statistical evidence is in accordance with the explanation of the structure of Thai language, which can be explained from the following syntactic patterns as follows.

Pattern (1): Verb and object relationship

S

verb	object
V *เก็บ*	N

In this pattern, noun is the object of the verb *เก็บ* /kep1/. Thus, it can be used as a sense indicator. This is the pattern that we expect when setting the hypothesis of this study.

Example 1 is the examples of collocational patterns of *เก็บ* /kep1/ meaning "to take", "to buy", "to gather", "to kill", "to hide", "to charge" and "to keep" respectively, which are in accordance with the relationship in the pattern (1).

Example 1:

(i) S(Verb: V(*เก็บ*)),(Object: N(*กล้วยไม้*))

(ii) S(Verb: V(*เก็บ*)),(Object: N(*หุ้น*))

(iii) S(Verb: V(*เก็บ*)),(Object: N(*คะแนน*))

(iv) S(Verb: V(*เก็บ*)),(Object: N(*คู่แข่ง*))

(v) S(Verb: V(*เก็บ*)),(Object: N(*อารมณ์*))

(vi) S(Verb: V(*เก็บ*)),(Object: N(*ค่าบริการ*))

(vii) S(Verb: V(*เก็บ*)),(Object: N(*สต๊อก*))

Pattern (2): Serial verb construction

SERIAVL VERB

verb verb

V *เก็บ* V

Thepkanjana (1986) defined serial verb as "verb(s) dependent on the first verb." From this definition, we can see that verbs in a serial verb construction are semantically related, as one verb is the element, which is modified, and another verb is the modifier. Here *เก็บ* /kep1/, which is the first verb in the construction is modified by another verb, which is its modifier. This pattern is not what we expect when setting the hypothesis.

Example 2 is the examples of collocational patterns of *เก็บ* /kep1/ meaning "to take", "to arrange", "to gather", "to hide" and "to keep" respectively, which are in accordance with the relationship in the pattern (2).

Example 2:

(i) Serial verb(Verb: V(*เก็บ*)),(Verb: V(*หา*))

(ii) Serial verb(Verb: V(*เก็บ*)),(Verb: V(*ใส่*))

(iii) Serial verb(Verb: V(*เก็บ*)),(Verb: V(*ออม*))

(iv) Serial verb(Verb: V(*เก็บ*)),(Verb: V(*กด*))

(v) Serial verb(Verb: V(*เก็บ*)),(Verb: V(*สำรอง*))

Although words on the left side of *เก็บ* /kep1/ play little role in the sense disambiguation, there are some collocational patterns of words from the left, namely, V + *เก็บ*, N + *เก็บ*. However, their collocational weights are lower than those at the right side. This statistical evidence is opposed to our hypothesis as we expect that noun as a subject of verb should have a semantic relationship with the verb, which can be used as a sense indicator of the verb *เก็บ* /kep1/.

The following patterns are the collocational pattern found at the left side, which have the syntactic relationship as follows.

Pattern (3): Subject and verb relationship

S

subject	verb
N	V *เก็บ*

In this pattern, subject can be used to disambiguate the senses of *เก็บ* /kep1/. Example 3 is the examples of collocational patterns of *เก็บ* /kep1/ meaning "to charge", which are in accordance with the relationship in the pattern (3).

Example 3:

(i) S(Subject: N(*รัฐบาล*)),(Verb: V(*เก็บ*))

(ii) S(Subject: N(*พนักงาน*)),(Verb: V(*เก็บ*))

Pattern (4): Head and modifier relationship

NP

head	modifier
N	V *เก็บ*

In this pattern, *เก็บ* /kep1/ is a modifier of the head noun. Thus there is a semantic relationship between the head noun and the modifier *เก็บ* /kep1/. This pattern plays some role in disambiguating the sense "to keep" as shown in figure 34.

Example 4 is the examples of collocational patterns of *เก็บ* /kep1/ meaning "to keep", which are in accordance with the relationship in the pattern (4).

Example 4:

(i) NP(Head: N(*สถานที่*)),(Mod: V(*เก็บ*))

(ii) NP(Head: N(*โกดัง*)),(Mod: V(*เก็บ*))

Pattern (5): Serial verb construction

SERIAL VERB

Verb	verb
V	V *เก็บ*

Here, verb *เก็บ* /kep1/ is a modifier at the left side in a serial verb construction which modifies its preceding verb. Both verb at the left side or at the right side of verb

เก็บ /kep1/ play role in the disambiguation, however, verb at the right sides play more role. We do not expect to see this pattern too.

Example 5 is the examples of collocational patterns of *เก็บ* /kep1/ meaning "to buy", "to arrange", "to kill", and "to keep", which are in accordance with the relationship in the pattern (5).

Example 5:

(i) Serial verb(Verb :V(*เข้า*)),(Verb : V(*เก็บ*))

(ii) Serial verb(Verb :V(*พับ*)),(Verb : V(*เก็บ*))

(iii) Serial verb(Verb :V(*สั่ง*)),(Verb : V(*เก็บ*))

(iv) Serial verb(Verb :V(*กัก*)),(Verb : V(*เก็บ*))

However, it should be noted again that patterns (3), (4) and (5) do not have much influences on the disambiguation because there are only few cases of these patterns. The fact that, the left side plays very little roles for the disambiguation of all senses of *เก็บ* /kep1/ is discussed in more details in section 5.1.2.

- Semantic explanation

From the syntactic relationship, we can see that verb and object relationship plays dominant role in sense disambiguation of *เก็บ* /kep1/. These objects can be grouped into their semantic fields with different fields indicate different senses. These semantic fields are shown as follows.

Senses	Semantic fields of co-occurred word forms	Examples of co-occurred word forms
To take	PLANT, THING, ANIMATE	*ผัก กล้วยไม้ ไม้ ไข่ หอย แมลง ขยะ ป้าย*
To buy	"Stock"	*หุ้น*
To gather	THING	*คะแนน หลักฐาน รายละเอียด ข้อมูล แต้ม*
To kill	HUMAN, PERSON	*คู่แข่ง นักการเมือง*
To arrange	Inconclusive	No co-occurred objects
To hide	EMOTION	*ความรู้สึก อารมณ์*
To charge	FEES	*ค่าใช้จ่าย ค่าบริการ ค่าเช่า ค่าเล่าเรียน*
To keep	THING	*รถ สารเคมี น้ำมันหล่อลื่น ข้าวเปลือก*

Table 21: Semantic relationship between senses of *เก็บ* /kep1/ and their co-occurred words-to-right.

According to the relationship between verb and object, the semantic relationships between *เก็บ* /kep1/ and its sense indicators, which are the object noun, can be explained in terms of semantic coherence. As we can see from the table above, “to take” occurs with an object that could be PLANT, ANIMATE, or THING; “to buy” can co-occur with “a stock”; “to gather” can co-occur with THING; “to kill” can co-occur with HUMAN, or PERSON; and so on.

However, for *เก็บ* /kep1/ meaning "to arrange", there is no co-occurred objects, thus its semantic relationship is inconclusive.

In terms of the number of context words, the reasons that 2W and 3W spans are also helpful in the disambiguation of three senses of *เก็บ* /kep1/, namely, "to hide", "to charge" and "to keep" can be explained in the similar way as discussed in section 4.1.2. It is because the training data is too small and some word forms can immediately co-occur with many senses of *เก็บ* /kep1/. For example *ข้อมูล* /khɔɔ2muun/" usually co-occurs with *เก็บ* /kep1/ meaning "to gather" and it also, sometimes, co-occurs with *เก็บ* /kep1/ meaning "to keep". Thus, the program would need more than one word for disambiguating these senses. For example, in *เก็บข้อมูลไว้,* if consider only 1W span, *ข้อมูล* /khɔɔ2muun/ would indicate the sense of *เก็บ* /kep1/ as "to gather" which is the wrong sense. The more correct sense, which is "to keep" can be indicated by *ไว้* /wai 2/ "used as a suffix of some verbs for strengthening their meanings", which is found at the 2W span.

The example above shows that sense indicators of *เก็บ* /kep1/ may not be a subject or object noun as we hypothesized, but they can be another verb. In fact, serial verb construction is very common in Thai. The examples of serial verbs that have *เก็บ* /kep1/ as their constituents are *เก็บเอามา, เก็บเอาไป, เก็บเอาไว้.*

Besides, since these serial verbs do not need to be immediately adjacent to each other, the unit in between a serial verb has less semantic relation than the constituent of a serial verb, which are in a further distance.

The examples of word forms co-occurred with eight senses of *เก็บ* /kep1/ at the right side as well as the left side are in appendix E, from table 31 to table 46, for sense "to take", "to buy", "to gather", "to kill", "to arrange", "to hide", "to charge", and "to keep" respectively.

4.3 Discussion and Conclusion

In section 4.1 and 4.2, we presented and discussed the results of sense disambiguation of *หัว* /hua4/ and *เก็บ* /kep1/ based on three perspectives, namely computational, syntactic and semantic perspectives by grouping word forms that are in the same syntactic and semantic classes, and explaining the relationship based on these groupings. On computational perspective, we explained the relationship between the number of co-occurred word forms and senses of ambiguous words. On syntactic perspective, we explained the relationship between syntactic classes or parts of speech of co-occurred word forms and senses of ambiguous words. On semantic perspective, we explained the relationship between semantic classes or semantic fields of co-occurred word forms and senses of ambiguous words. However, we would like to note here that, in WSD, it is better to consider word forms than syntactic classes or semantic fields because word forms are more specific than these features, thus, they are better sense indicators (see section 5.1.1 for further discussion about the limitations of syntactic classes and semantic fields).

The results and discussions are summarized in table 22 and 23 in this section for sense disambiguation of *หัว* /hau4/ and table 24 and 25 for sense disambiguation of *เก็บ* /kep1/.

In this summarization, we divided senses of *หัว* /hua4/ and *เก็บ* /kep1/ into two groups, namely senses that have high precision rates (those that have precision rate more than or equal 80% for *หัว* /hua4/ and more than or equal 75% for *เก็บ* /kep1/), and senses that have low precision rates (those that have precision rate lower than 80% for *หัว* /hua4/ and lower than 75% for *เก็บ* /kep1/). The reasons why some senses have high precision rate, and some senses have low precision rate are as follows.

- Senses that have high precision rate are senses that meet the following conditions.

1. Senses that co-occur with only one word form, such as *โจก* /cook1/, *เสีย* /siia4/, *ค่ำ* /kham2/, *ปวด* /puuat1/, *พาด* /phaat2/, for senses "chief", "emotion", "early hours", "brain" and "headline" respectively.

2. Senses that co-occur with classes that have few word forms such as OPERATION VERB and COLOR, for senses "machine part" and "hair" respectively.

3. Senses that have high frequency in a training corpus such as "head", "entity" and "brain".

- Senses that have low precision rate are senses that meet the following conditions.

1. Senses that co-occur with many word forms. For example, there are too many word forms that can co-occur with *หัว* /hua4/ sense "front", such as *เรือ* /rɯɯa/ "boat", *เครื่องบิน* /khrɯɯaŋ2bin/ "airplane", *รถ* /rot3/ "car", *รถยนต์* /rot3yon/ "car", *เตียง* /tiiaŋ/ "bed", *เตียงนอน* /tiiaŋnɔɔn/ "bed", etc. Thus there is no clear indicator for this sense. This is the reason why the algorithm's performance on disambiguating this sense is very low.

2. Senses that co-occur with classes that have many word forms such as THING, for senses "top" and "front".

3. Senses that have low frequency in a training corpus, such as "talent", "heading" and "head of coin".

These are the same reasons that are discussed in section 5.1.1 for the strengths and weaknesses of the decision list algorithm used in this study.

Senses that have high precision rate					
Senses	Precision rate		Explanation		
	WR	WL	Computational	Syntactic	Semantic
Chief	100%	40%	Number of WR is higher than WL and their collocational weights are higher than WL	NP(Head:N(หัว)),(Mod:N)	Inconclusive
Emotion	100%	0%		NP(Head:N(หัว)),(Mod:ADJ)	"bad" (inconclusive)
Machine part	100%	35.71%		NP(Head:N(หัว)),(Mod:V)	OPERATION VERB
Early hours	100%	45.45%		NP(Head:N(หัว)),(Mod:N)	HOUR
Concentrate	100%	7.69%		NP(Head:N(หัว)),(Mod:N)	THING, LIQUID
Viewpoint	94.44%	63.89%		NP(Head:N(หัว)),(Mod:ADJ,V)	ATTRIBUTE
Intelligence	81.25%	31.25%		NP(Head:N(หัว)),(Mod:ADJ,V,N)	ATTRIBUITE
Hair	80%	20%		NP(Head:N(หัว)),(Mod:N,ADJ)	COLOR

Table 22: Summarization of results and discussions of senses ofหัว /hua4/ that have high precision rate.

Senses that have high precision rate					
Senses	Precision rate		Explanation		
	WR	WL	Computational	Syntactic	Semantic
Head	72.22%	86.11%	Number of WR is almost equal to the number of WL	NP(Head:N(หัว)),(Mod:N,ADJ) S(Subject:N(หัว)),(Verb:V)	HUMAN or ANIMAL
				NP(Head:N),(Mod:N(หัว)) S(Verb:V),(Object :N(หัว))	VERB
Entity	88.54	90.63%		NP(Head:N(หัว)),(Mod:N,ADJ)	PERSON
				S(Verb:V),(Object:N(หัว)) NP(Head:N),(Mod:N(หัว))	Inconclusive
Brain	3.45%	89.66%	Number of WL is higher than WR and weight are higher than WR	S(Verb:V),(Object:N(หัว))	PHYSICAL STATE
Headline	12.50%	87.50%		S(Verb:V),(Object:N(หัว))	"to headline" (inconclusive)

Table 22: Summarization of results and discussions of senses ofหัว /hua4/ that have high precision rate.

Senses that have low precision rate					
Senses	Precision rate		Explanation		
	WR	WL	Computational	Syntactic	Semantic
Topics	77.78%	38.89%	Number of WR is higher than WL and their collocational weights are higher than WL	NP(Head:N(หัว)),(Mod:N)	DISCOURSE
Bulb	71.43%	42.86%		NP(Head:N(หัว)),(Mod:N)	PLANT
Top	63.64%	27.27%		NP(Head:N(หัว)),(Mod:N)	THING
Front	64.52%	25.81%		NP(Head:N(หัว)),(Mod:N)	THING
Titles or names	69.23%	53.85%	Number of WR is almost equal to the number of WL	NP(Head:N(หัว)),(Mod:N,ADJ)	PRINTED MATERIAL
				NP(Head:N,ADJ),(Mod:N(หัว))	PRINTED MATERIAL
Talent	0%	0%	Inconclusive	Inconclusive	Inconclusive
Heading	0%	0%			
Head of coin	0%	0%			

Table 23: Summarization of results and discussions of senses of*หัว* /hua4/ that have low precision rate.

Senses that have high precision rate					
Senses	Precision rate		Explanation		
	WR	WL	Computational	Syntactic	Semantic
To charge	85.05%	62.62%	Number of WR is higher than WL and their collocational weights are higher than WL	S(Verb:V(เก็บ)),(Object:N) Serial verb(Verb:V(เก็บ)),(Verb:V)	FEES
To keep	83.75%	65.63%		S(Verb:V(เก็บ)),(Object:N) Serial verb(Verb:V(เก็บ)),(Verb:V)	THING
To buy	80%	40%		S(Verb:V(เก็บ)),(Object:N)	STOCK
To gather	79.03%	12.90%		S(Verb:V(เก็บ)),(Object:N) Serial verb(Verb:V(เก็บ)),(Verb:V)	THING
To hide	77.78%	0%		S(Verb:V(เก็บ)),(Object:N) Serial verb(Verb:V(เก็บ)),(Verb:V)	EMOTION
To take	76.09%	8.70%		S(Verb:V(เก็บ)),(Object:N) Serial verb(Verb:V(เก็บ)),(Verb:V)	PLANT, THING, ANIMATE

Table 24: Summarization of results and discussions of senses of*เก็บ* /kep1/ that have high precision rate.

Senses that have low precision rate					
Senses	Precision rate		Explanation		
	WR	WL	Computational	Syntactic	Semantic
To kill	66.67%	66.67%	Number of WR is higher than WL and their collocational weights are higher than WL	S(Verb:V(เก็บ)),(Object:N)	HUMAN, PERSON
To arrange	16.67%	16.67%		Serial verb (Verb:V(เก็บ)),(Verb:V))	Inconclusive
To pick up	0%	0%	Inconclusive	Inconclusive	Inconclusive

Table 25: Summarization of results and discussions of senses of*เก็บ* /kep1/ that have low precision rate.

CHAPTER V

DISCUSSIONS CONCLUSIONS AND FURTHER SUGGESTIONS

This chapter consists of three main sections. Section 5.1 discusses the important issues found in this study. Section 5.2 summarizes the main points of this study. Section 5.3 suggests the way to improve the algorithm's performance and to further develop WSD program in Thai. The details about each section are as follows.

5.1. Discussions

There are four important issues discussed in this section. Section 5.1.1 discusses the idea of WSD using decision list collocation, whether it is applicable to the task, and its strengths and limitations. As a result of the last chapter, section 5.1.2 further discusses the reasons why words on the left play very little role in the disambiguation of all senses of เก็บ /kep1/, and section 5.1.3 further discusses the reasons why words on the right are sense indicators of หัว /hua4/ and เก็บ /kep1/. Section 5.1.4 discusses whether the optimal spans of disambiguating หัว /hua4/ and เก็บ /kep1/ can be used for disambiguating other nouns and verbs. The details of the discussions are as follows.

5.1.1 WSD of หัว /hua4/ and เก็บ /kep1/ using decision list collocation

The results presented in chapter 4 suggest that the algorithm's performance is very good even though the size of the training data is not large. The performance is very much higher than the lower bound performance and is very little lower than the

upper bound performance. The strengths of the algorithm can be explained by three factors as follows.

(1) The algorithm performs WSD by always choosing a sense co-occurred with a word form that has the highest collocation weight. From a statistical view, a word form with the highest collocational weight will be the most reliable sense indicator of the ambiguous word.

(2) The algorithm performs WSD by using the information in a small window span of only 1W or 2W. Disambiguating in small window span yields a good result because 1W or 2W (which usually are content words) is usually in the same syntactic construction, (such as NP, VP, serial verb, sentence) with the ambiguous word. Thus it should be more semantically related with the ambiguous word than the word in the further distance.

(3) The algorithm performs feature selection by choosing only the best feature, which is a word form in this study. When comparing word forms with other features like parts of speech (such as noun, verb, adjective etc.) and conceptual features (such as ANIMATE, INANIMATE, PLANT, HUMAN, etc.), word forms are better in many aspects as follows.

First, features like parts of speech or conceptual features are too general classification, that is, while there are not many senses that a word form can co-occur with, a part of speech can co-occur with all senses. For example, while *ใหม่* /mai1/ usually co-occurs with only three senses of *หัว* /hua4/ namely, "viewpoint", "titles or names" and "machine part", an adjective can co-occur with all senses of *หัว* /hua4/. Thus, we cannot use an adjective to disambiguate the sense of *หัว* /hua4/, but we can use the word *ใหม่* /mai1/ to narrow down the possible senses of *หัว* /hua4/.

Conceptual features seem to be a good choice. For example, we found that ANIMATE always co-occurs with *หัว* /hua4/ meaning "head", PLANT always co-occurs with *หัว* /hua4/ meaning "bulb". However, conceptual features can be too general like parts of speech. For example, while *หอม* /hɔɔm4/ "onion" and *ขิง* /kiŋ 4/ "ginger", which are PLANT can indicate sense "bulb", *มะระ* /ma3ra3/ "bitter cucumber" and *มะละกอ* /ma3ra3kɔɔ/ "papaya" which also have the feature PLANT, cannot indicate this sense.

Second, using these features causes problems in preparing manually sense-tagged training data. In addition to tagging sense of the ambiguous words, we have to tag parts of speech or conceptual features on context words too, which is time-consuming and costly.

Thus, the advantages of using word form are as follows.

(3.1) Word form is a good sense indicator because it is more specific that is, one word form usually co-occurs with only one sense. For example, *ค่ำ* /kham2/ "night" always co-occurs with *หัว* /hua4/ "early hours", *พาด* /phaat2/ "to headline" always co-occurs with *หัว* /hua4/ "headline", *ออม* /ʔɔɔm/ "to save", always co-occurs with *เก็บ* /kep1/ "to gather", *กัก* /kak1/ "to detain" always co-occurs with *เก็บ* /kep1/ "to keep".

(3.2) We can prepare the training data easier than when using other features because there is no need to manually tag any information or features into word forms.

(3.3) Beside the advantages of word form to this study, word forms can help further disambiguating other ambiguous words by applying the concept of mutually disambiguation[1], in which there will be only one combination of meanings that fits together. For example, if we know that *หัว* /hua4/ means "hair" when co-occurs with

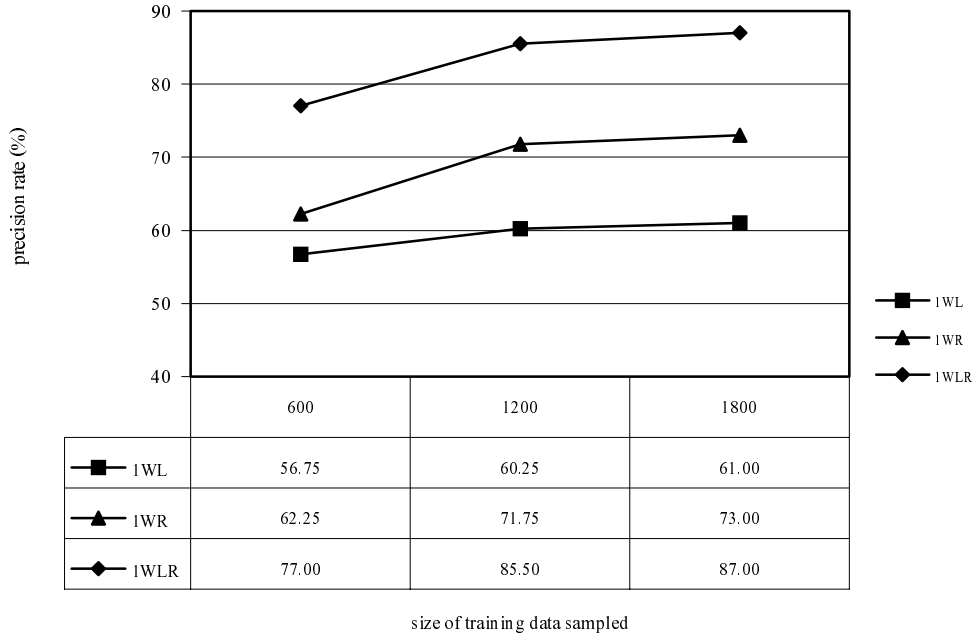
[1] See the explanation of mutually disambiguation in section 2.4.2.1.1.1.

จิก /jik1/, we automatically know that จิก /jik1/ co-occurs with this sense of หัว /hua4/ means "to pull (someone hair)" (instead of "to peck").

However, despite of these strengths, the algorithm cannot perform as good as the human because of the following two problems.

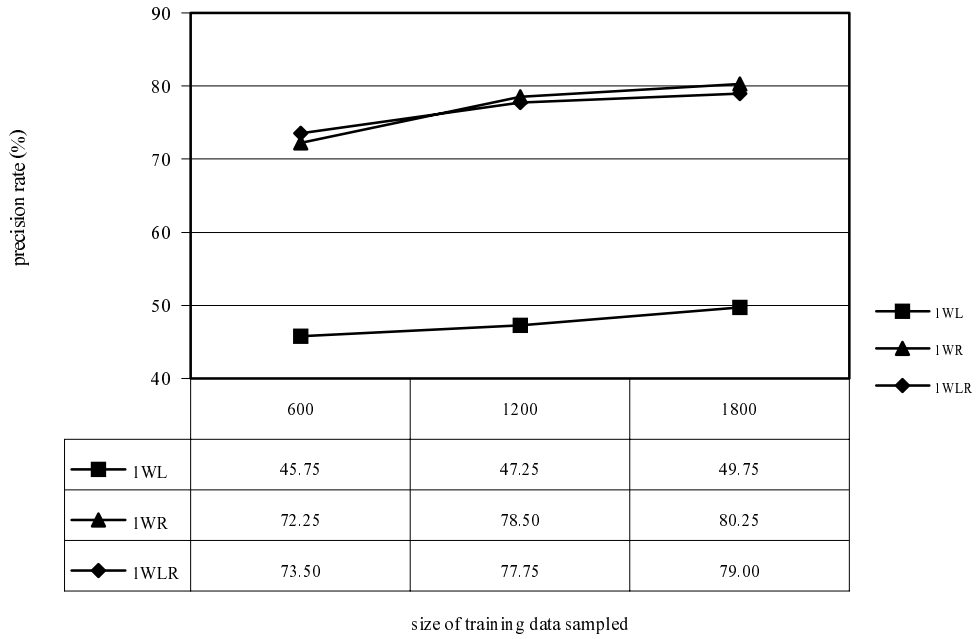
(1) The size of training data is not large enough. This is an important problem faced by a corpus-based WSD as its disambiguation requires a large size of knowledge (sense-tagged data). Thus, the algorithm performance is good if it is trained with sufficient knowledge. However, the limitation is that the preparation of sense-tagged data is time-consuming and costly.

As stated in section 3.1.3, we have tested the algorithm at different sizes of the training data, namely at 600, 1,200, 1,800 samples, to see whether the precision rates would increase as the size of the training data increase. The result is that the precision rates increase as shown in figure 30 and 31. This indicates that the more data trained to the algorithm, the more the precision rate increases.



	600	1200	1800
1WL	56.75	60.25	61.00
1WR	62.25	71.75	73.00
1WLR	77.00	85.50	87.00

Figure 35: Results on disambiguating *หัว* /hua4/ at different sizes of training data.



	600	1200	1800
1WL	45.75	47.25	49.75
1WR	72.25	78.50	80.25
1WLR	73.50	77.75	79.00

Figure 36: Results on disambiguating *เก็บ* /kep1/ at different sizes of training data.

Besides, the quality of the data is also important, that is if adding more training data does not provide more new and useful word forms for the disambiguation, the

algorithm will not perform much better. Thus, the increases in the quality as well as quantity of the training data are both important.

In this study, the small size of the training data has an effect on the algorithm's performance as follows.

First, because the training samples for *หัว* /hua4/ in the sense of "talent" and *เก็บ* /kep1/ in the sense of "to pick up" is not large enough, the program cannot disambiguate these senses at all.

Second, because the training data is not large enough, many word forms then do not have discriminate power, which are cases 2,3, and (2) discussed in chapter 3, section 3.3.3. Thus, the size of the training data must be large enough so that a word form will have a strong discriminated power of only one sense.

(2) The problem of a word form itself. Even though the size of the training data is large enough so that a word form has a strong collocational weight with only one sense, some word forms may not always indicate the correct sense. In other word, choosing the senses that indicated by the word form that has the highest collocational weight is not always correct. For example, the word form *ฟาด* /faad2/ can immediately co-occur with *หัว* /hua4/ in the senses of "entity", and "head". Even though it has the highest collcoational weight when co-occurs with "entity", there are many context as in (i) that the correct sense is "head".

(i) …ผู้สนับสนุนฝ่ายแพ้ งัดกำปั้นและไมค์ลอย**ฟาดหัว**เทศมนตรีคนใหม่เลือดซึม…

Thus, these two problems are the reasons why the decision list algorithm cannot perform as good as human.

However, the fact that algorithm is tested with the highly ambiguous words like *หัว* /hua4/ and *เก็บ* /kep1/, suggests that the idea of the decision list algorithm is effective (in spite of the above two problems). Thus, if the algorithm is tested with less ambiguous words, its performance should be better or even close to human performance. By testing with less ambiguous word, the algorithm will have fewer problems with word form. The reason is because, first, less ambiguous words have clearer sense indicators, as their senses are not closely related, so different senses occur with totally different context. For example, for the ambiguous word *ขัน* /khan4/, *ไก่* /kai1/ "chicken" is a good indicator for sense "to crow, coo", and *ไก่* has very low probability to occur with other senses like "to laugh". *เขา* /khaw4/ "he" is a good indicator for *ขัน* /khan4/ meaning "laugh" and *เขา* /khaw4/ has very low probability to occur with *ขัน* /khan4/ sense "to crow, coo". Second, disambiguating less ambiguous words requires fewer training data. This is because less ambiguous words have fewer numbers of senses, so it is easier to find many samples of all senses in a small size of data.

5.1.2 Why WL are not the sense indicators of *เก็บ* /kep1/?

This question arises as a result of the disambiguation of *เก็บ* /kep1/ presented in section 4.2. The results suggest that WL play very little role in disambiguating all senses of *เก็บ* /kep1/. The explanation of the results in section 4.2 indicates that object noun plays dominant role while subject noun plays very little role in disambiguating all senses of *เก็บ* /kep1/. The reasons are as follows.

First, there are many word forms that can act as objects of verb, and these word forms usually indicate only small numbers of senses. But, there are few word forms that can act as the subject of verb, and these word forms usually indicate many senses of *เก็บ* /kep1/. For example, in *เขาเก็บ*, *เขา* /khaw4/ "he" can co-occur with all senses of

เก็บ /kep1/, while, in *เก็บภาษี*, *ภาษี* /phaasii4/ "tax" co-occurs with sense "to charge" only. This is in accordance with the explanation in section 4.2 that we found more word forms in the decision list of the right side than the left side.

However, we would like to note here that too many word forms could lessen the performance as stated in section 4.3, which discussed about reasons why some senses have low precision rate.

Second, from the information in the corpus, we found that subject nouns do not always occur close to the verb *เก็บ* /kep1/. In other words, it is rare to find simple sentences, in which subjects immediately come before verbs. Thus, the subject nouns are not likely good sense indicators for the verb *เก็บ* /kep1/.

Beside verb and object relationship, serial verb construction also helps disambiguating *เก็บ* /kep1/ as explained in section 4.2. However, *เก็บ* + V play more role than V + *เก็บ*. There can be two explanations.

First, there are many word forms at the right side, and these word forms indicate fewer senses, while there are fewer word forms at the left side, and these word forms indicate more than one sense. This explains why word forms on the right side are better sense indicators than those on the left side. For example, in *เก็บสะสม*, *สะสม* /sa1som4/ "to accumulate, collect" indicates *เก็บ* /kep1/ sense "to gather", while in *จัดเก็บ*, *จัด* /cat1/ "to arrange, organize" can occur with sense "to charge", as in *จัดเก็บภาษี*, or "to keep", as in *จัดเก็บข้อมูล*. This is also in accordance with the explanation in section 4.2 that we found more word forms in the decision list of the right side than the left side.

Second, verbs at the right side of *เก็บ* /kep1/ and *เก็บ* /kep1/ are more semantically related than verbs at the left side. This is in accordance with Filbeck (1975) cited in Thepkanjana (1986:14), that "He argues that all verbs including the initial verb (or verb phrase) in a serial verb construction refer to a single proposition, or in other words, a single event. He stated that the initial verbs carries the true

predicate meaning of the proposition and any subsequent or serial verb (or verb phrase) indicates a functional meaning which is related to the meaning of the initial verb. In other words, a serial verb (or verb phrase) modifies the initial verb (or verb phrase)." Thus, in a serial verb construction, a subsequent verb modifies its preceding verb. From the first reason, *เก็บ* /kep1/ usually acts as the initial verb in a construction rather than as the subsequent verb. Thus, *เก็บ* /kep1/ is modified by its subsequent verb, which is its modifier at the right side. For example, in *เก็บเอาไป*, *เอา* /?aw/ "to take, bring" indicates *เก็บ* /kep1/ sense "to take", which is the correct sense in this context. In *เอาไปเก็บ*, *เอา* /?aw/ "take" or *ไป* /pai/ "to go" at the left side indicate sense "to take", which is a wrong sense in this context. The correct sense is "to keep".

5.1.3 Why WR are sense indicators of *หัว* /hua4/ and *เก็บ* /kep1/?

WR are sense indicators of *หัว* /hua4/ and *เก็บ* /kep1/ can be explained by the theory of language typology in terms of word ordering. Word ordering is the ordering of words in a syntactic structure of a language, such as the ordering of head and modifier in a phrase, the ordering of subject, verb and object in a sentence, and the ordering of a noun and its definite article in a noun phrase.

In Thai, the modifier is at the right side of its head. For example, in a noun phrase, the modifier is at the right side of its head noun. The results in section 4.1.1.1 suggested that head and modifier relationship plays dominant role in sense disambiguation of *หัว* /hua4/, thus sense indicators of *หัว* /hua4/ are WR. However, in other languages, such as English and Chinese, the modifier is at the left side of its head, which suggests that sense indicators of noun in English and Chinese may be at the left side.

For the ordering of subject, verb and object in a sentence, since Thai is SVO language, subject is at the left side of verb and object is at the right side of verb. From the results in section 4.2, we can see that verb and object relationship plays dominant

role in sense disambiguation of เก็บ /kep1/, thus WR are sense indicators of เก็บ /kep1/. The sense indicators of verbs in English and Chinese may be on the right side too as English and Chinese are also SVO language.

The ordering of noun and its definite article are not considered in sense disambiguation of Thai and Chinese words, as Thai and Chinese do not have definite articles. English has definite articles, such as *a the*, which are at the left side of nouns. However, they do not play role in sense disambiguation, as they are function words, which contribute no clue for disambiguation. French has definite articles such as *la le*, which are at the left side of nouns and they have great influence on sense disambiguation of nouns. Since *la* is used with feminine nouns, and *le* is used with masculine nouns, senses of ambiguous nouns can be known if these definite articles are present with ambiguous words. For example, in French, *livre* can have two meanings. If it is feminine, it means "pound", if it is masculine, it means "book". Thus, if *la* is present at the left side of *livre*, we know that *livre* means "pound", and if *le* is present at the left side of *livre*, we know that *livre* means "book".

Thus word order in a syntactic structure has a great influence on setting the hypothesis on locating sense indicators. Since different languages have different word ordering, moreover, some languages are the same in some aspect of word ordering and different in other aspects, different languages require different hypothesis. For example, Thai, English and Chinese are SVO language, the location of sense indicators for verb are the same. However, English and Chinese differ from Thai in that, a modifier is on the left side of its head, thus the location of sense indicators for noun are different from that of Thai. Thai and Chinese may not consider word ordering of noun and definite article. While English definite articles play no role in sense disambiguation, French definite articles play dominant role.

5.1.4 Can หัว /hua4/ be a representative for noun, and เก็บ /kep1/ be a representative for verb?

After knowing that the optimal span for disambiguating หัว /hua4/ is 1WRL and เก็บ /kep1/ is 2WR and the sense indicators of หัว /hua4/ and เก็บ /kep1/ are on the right side, the question that follows is whether these results can be used for disambiguating other nouns, and verbs. It is possible that these optimal spans may not be applicable to other nouns and verbs. For example, the sense indicators of other verbs such as ขัน /khan4/ may be on the left side. This is because of subject and verb relation, in which subject plays a good role in the disambiguation of ขัน /khan4/, as in เขาขัน, เขา /khaw4/ "he", as a subject, is a good indicator of sense "to laugh", while in ไก่ขัน, ไก่ /kai1/ "chicken", as a subject, is a good indicator of sense "to crow, coo".

Thus, there should be further study on sense disambiguation of other ambiguous nouns and verbs to see whether the optimal span and the optimal side are the same as those found in this study.

5.2 Conclusions

The summaries of the main points from this study are as follows.

5.2.1 The Analysis of all Possible Senses of หัว /hua4/ and เก็บ /kep1/

In chapter 3, we prepared manually sense-tagged corpus for the training and testing processes. First, we manually assign the senses listed in the Thai dictionary of "The Royal Institute" and found that these senses are not suitable to some context in the corpus of "Bangkok Business" newspaper. So, we analyzed and established the additional senses based on the data in the corpus. Based on the definitions listed in the

dictionary and from the corpus, we got twenty senses of *หัว* /hua4/ and nine senses of *เก็บ* /kep1/.

5.2.2 WSD Using Decision List Collocation

In chapter 3, to find the optimal span for the disambiguation among the twenty spans, we trained the algorithm twenty times for twenty spans of collocation. We got twenty decision lists for twenty spans trained. Then, we applied each decision list for the disambiguation of *หัว* /hua4/ and *เก็บ* /kep1/ on the unseen text. The algorithm performs the disambiguation by choosing the sense co-occurred with word form that has the highest collocational weight based on the principle that collocational patterns that have higher weight are more statistically significant than those lower weight. Therefore, they should be better sense indicators.

In chapter 4, we reported the results of the performance of the disambiguation of *หัว* /hua4/ and *เก็บ* /kep1/. We found that $\pm$ 5 is sufficient for the disambiguation and the optimal span for the disambiguation of *หัว* /hua4/ is 1WRL with the precision rate of 87%, *เก็บ* /kep1/ is 2WR with the precision rate of 80.25%. The sense indicators of *หัว* /hua4/ and *เก็บ* /kep1/ are on the right side, while WL play some role for disambiguating *หัว* /hua4/, they play very little role in disambiguating *เก็บ* /kep1/.

We explained the results of the disambiguation based on three perspectives namely, computational, syntactic and semantic. However, the main points that we would like to conclude are from the syntactic perspective, which constitutes our hypotheses, as follows.

The reason that the optimal span for the disambiguation of *หัว* /hua4/ is 1WRL because the sense indicators of *หัว* /hua4/ usually immediately co-occur with *หัว* /hua4/. The sense indicators of *หัว* /hua4/ are at the right side because *หัว* /hua4/ usually acts as the head and another noun, verb or adjective acts as the modifier at the right side. The left side also plays role in the disambiguation of some senses of *หัว* /hua4/ because of

the head-modifier relationship, in which *หัว* /hua4/ acts as the modifier of the head, which is another noun, adjective or verb on the left side of *หัว* /hua4/. This indicates that there are some semantic relationship between the head and its modifier. Thus, whether *หัว* /hua4/ is a head or a modifier, its constituent part plays role in the sense disambiguation. However, words on the right side are better sense indicators because most of the samples found in this study have *หัว* /hua4/ as the head of the construction.

For *เก็บ* /kep1/, we found that the optimal span for the disambiguation is 2WR. This is explained by the structure of serial verb in Thai, in which there usually is another word between these serial verbs, and this word is not a better sense indicator as the following verb. WR is the sense indicator of *เก็บ* /kep1/ because of the dominant role of verb and object relation, in which the object of *เก็บ* /kep1/ can be used to disambiguate the sense of *เก็บ* /kep1/. WL is not a sense indicator of *เก็บ* /kep1/ because the subject and verb relationship as well as verb (at the left) and *เก็บ* /kep1/ relationship play very little role in the disambiguation.

5.3 Further Suggestions

We would like to suggest two tasks that can be done further.

(1) To increase the performance of the decision list algorithm disambiguating *หัว* /hua4/ and *เก็บ* /kep1/, the following should be done.

(1.1) Increase the size of the training data. However, increasing the size of the data means increasing both the number of the data and the quality of the data. That is, the data should provide new information for the disambiguation.

(1.2) Resolve the problem of word forms that do not have discriminate power of senses by using the concept of mutually disambiguation. For example, if two senses of *ฟาด* /faad2/ "to buy" and "to hit" have already been disambiguated, when *ฟาด* /faad2/ co-occurs with *หัว* /hua4/, *หัว* /hua4/ can be

disambiguated. This is because there will be only one combination of meanings that fits together. So, ฟาด /faad2/ sense "to buy" fits with หัว /hua4/ sense "entity", and ฟาด /faad2/ sense "to hit" fits with หัว /hua4/ sense "head". By doing this, we have to manually assign the correct senses of ฟาด /faad2/ in the training corpus so that when training and testing, the algorithm will consider both word forms and their senses for the disambiguation.

(2) To further develop this prototype program into the complete WSD program in Thai by

(2.1) Applying the concept of mutually disambiguation for testing other words by using word forms with disambiguated senses to help disambiguate other ambiguous words.

(2.2) Testing with other words to find the representative of the optimal span for WSD of noun and verb.

(2.3) Testing with other parts of speech such as adjective.