Distributional Semantics Approach to Thai Word Sense Disambiguation

Sunee Pongpinigpinyo, and Wanchai Rivepiboon

***Abstract*—**Word sense disambiguation is one of the most important open problems in natural language processing applications such as information retrieval and machine translation. Many approach strategies can be employed to resolve word ambiguity with a reasonable degree of accuracy. These strategies are: knowledge-based, corpus-based, and hybrid-based. This paper pays attention to the corpus-based strategy that employs an unsupervised learning method for disambiguation. We report our investigation of Latent Semantic Indexing (LSI), an information retrieval technique and unsupervised learning, to the task of Thai noun and verbal word sense disambiguation. The Latent Semantic Indexing has been shown to be efficient and effective for Information Retrieval. For the purposes of this research, we report experiments on two Thai polysemous words, namely **หัว** /hua4/ and **เก็บ** /kep1/ that are used as a representative of Thai nouns and verbs respectively. The results of these experiments demonstrate the effectiveness and indicate the potential of applying vector-based distributional information measures to semantic disambiguation.

***Keywords*—**Distributional semantics, Latent Semantic Indexing, natural language processing, Polysemous words, unsupervised learning, Word Sense Disambiguation.

I. Introduction

WORD Sense Disambiguation (WSD) is the process of resolving natural language ambiguities, which is one of the most important problems in the computational linguistics. It refers to the process of selecting the most appropriate meaning or sense to a given ambiguous word within a given context. Resolving the word ambiguity is considered as the major bottleneck for large scale language understanding applications and their associate tasks such as machine translation (MT), information retrieval (IR), natural language understanding (NLU) and others. These various range applications of natural language processing need knowledge of word meaning to select the correct word sense in a context. For example, the Thai word หัว /hua4/ has many different senses, one of which can be translated into English as head, and another as chief.

As with other languages, working with Thai single words and Thai compound words also deals with the ambiguity of word meanings. Polysemy refers to a word that has more than one related meaning or sense, which is derived from the same word form and listed in the same lexical entry in a dictionary [15]. Homonymy is a type of word that has a completely different meaning or sense, which accidentally has the same word form and is listed in the same lexical entry in a dictionary [15].

Many approaches have been proposed for eliminating the ambiguous. Most of the word sense researching is to assess in English sentence and to assist in English language translation. Most Thai words, especially Thai compound words, have several meanings or senses and depend on context words which ambiguous between these senses. The Thai compound words make ambiguity of word meanings more complex problem. This type of word is composed of a combination of different words where each part of the combined word has a complete meaning by itself. The Thai new compound word may retain a partial meaning of the original combined words or completely different meaning from each word. For example, the compound word *หางเสือ* (*หาง* (tail), *เสือ* (tiger)) has two senses which mean tiger tail and rudder. The word *หาง* (tail) maintains its meaning that is an end body part of animal. When it is combined with *เสือ* (tiger), it still retains its relation meaning in the equivalent level as the end body part of a tiger. But the compound word sometimes has a completely different meaning. For example, *หางเสือ* (*หาง* (tail), *เสือ* (tiger)) which means rudder has the meaning which is completely different from each word.

Three main approaches have been applied in the WSD field.

1. Machine Readable Dictionaries (MRD) rely on information provided by Machine Readable Dictionaries (MRD) [1, 11, 12].

2. Supervised learning approaches use information gathered from training on a corpus that has sense-tagged for semantic disambiguation [13]. A major obstacle of this approach is the difficulty of manual sense-tagged in a training corpus that impedes the applicability of many approaches to domains.

3. Unsupervised leaning approaches determine the class membership of each object to be classified in a sample without using sense-tagged training examples [14, 19]. These approaches are considered to have an advantage over supervised learning approach as they do not require costly hand-tagged training data.

This paper describes an investigation of the use of Latent Semantic Indexing (LSI) to solve the semantic ambiguity of

Manuscript received March 8, 2005.

Sunee Pongpinigpinyo is with the Department of Computer Engineering, Chulalingkorn University, Bangkok, 10330, Thailand (phone: 662-218-6992, e-mail: sunee.po@student.chula.ac.th).

Wanchai Rivepiboon is with the Department of Computer Engineering, Chulalingkorn University, Bangkok, 10330, Thailand (phone: 662-218-6992, e-mail: wanchai.r@chula.ac.th).

two Thai polysemous words, namely *หัว* /hua4/ and *เก็บ* /kep1/, which are used as a representative of nouns and verbs respectively. In this research, we focus on the data that are created by combining an individual word *หัว* /hua4/ and *เก็บ* /kep1/ with other words to be compound words, reduplicative and repetitive words with transparent meanings. หัว /hua4/ and เก็บ /kep1/ which have parts of speech other than noun and verb respectively are excluded in our study.

Latent Semantic Indexing (LSI) [3] is a corpus-based statistical method for inducing and representing aspects of the meanings of words and passages (of natural language) reflective in their usage. The method generates a real valued vector description for documents of text. Basically, the central concepts of LSI is that the information about the contexts in which a particular word appears or does not appear provides a set of mutual constraints to determine the similarity of meaning of sets of words to each other. The advantage of LSI is that it is a fully automatic corpus based statistical procedure that does not require syntactic analysis.

The remainder of this paper is organized as follows. Section 2 describes the related work using corpus statistics to disambiguate word sense meaning. Section 3 briefly describes the Latent Semantic Indexing. Section 4 explains the experimental methodology to Thai word sense disambiguation. Section 5 reports and discusses the results. In section 6 some conclusions are dawn and some suggestion for future work on the area of WSD offered.

II. Related Works

A wide rang of approaches have been investigated and a large amount of effort devoted to tackle WSD. Currently one of the most successful line of research is the corpus-based approach using statistical or Machine Learning (ML) algorithms. Both supervised learning and unsupervised learning are applied to learn statistical models or classifiers from corpora in order to perform WSD. For the research reported in this paper, the focus will be on the use of unsupervised methods for WSD.

The method adopted by Schutze and Zernik [16, 20] avoids tagging each occurrence in the training corpus and associates each sense of a polysemous word with a set of its co-occurring words. If a word has several senses, then the word is associated with several different sets of co-occurring words, each of which corresponds to one of the senses of the word.

Yarowsky [19] used an unsupervised learning procedure with noun WSD. The proposed algorithm starts with a set of labeled data (seeds) and builds a classifier which is then applied on the set of unlabeled data. Only those instances that can be classified with a precision exceeding a certain minimum threshold are added to the labeled set. The result of Yarowsky [19]'s method shows that the average percentage attained was 96.1% for 12 nouns when the training data was a 460 million-word corpus, although Yarowsky used only nouns and did not discuss more than two senses of a word.

Pedersen and Bruce [14] presented three unsupervised learning algorithms to distinguish the sense of an ambiguous word in untagged text. These were McQuitty's similarity analysis, Ward's Minimum-Variance method and the EM algorithm. These algorithms assign each instance of an ambiguous word to a known sense definition based solely on the values of automatically identifiable features in text. Pedersen and Bruce reported that the disambiguating of nouns is more successful than adjectives or verbs. The best result for verbs was provided through the use of McQuitty's method (71.8%), although they tested only 13 ambiguous words of which only 4 were verbs.

III. Latent Semantic Indexing

Latent Semantic Indexing (LSI) is a psychological model and computational simulation intended to help explain the way that humans learn and represent the meaning of words, text, and other knowledge.

LSI is a vector based model of semantic based on word co-occurrences [4]. Words that occur together in the same or similar contexts are considered to be semantically similar. Likewise, words that occur together co-occurrence and have a syntagmatic and/or thematic connection are considered to be semantically similar. The LSI algorithm is trained on a corpus of documents. Documents here are any semantically cohesive set of words, such as sentences, paragraphs, any articles, etc. To build the LSI model for the experiments in this paper, a large co-occurrence matrix of documents is then created, with rows corresponding to words in the vocabulary, and columns to documents. Each entry in the matrix is a weighted frequency of the corresponding term in the corresponding document.

LSI relies on a Singular Value Decomposition (SVD) [17] of a matrix (word × context) derived from a corpus of natural context that pertains to knowledge in the particular domain of interest. SVD is a form of factor analysis and acts as a method for reducing the dimensionality of a feature space without serious loss of specificity. Typically, the word by context matrix is very large and quite often sparse. SVD reduces the number of dimensions without great loss of descriptiveness. SVD is the underlying operation in a number of applications including statistical principal component analysis [8], text retrieval [2, 6] pattern recognition and dimensionality reduction [5] and natural language standing [10].

The next step is to reduce the very large sparse matrix into a compressed matrix which based on Singular Value Decomposition (SVD). The result of the SVD is a k-dimensional vector space containing a vector for each term and each document.

$M = T \times S \times D'$

The original $t \times d$ matrix M is decomposed into a reduced rank $t \times k$ term matrix T, a diagonal matrix of singular values, S and a $d \times k$ document matrix D.

Decreasing k, the number of dimensions retained, reduces the accuracy with which M can be recreated from its component matrixes, but importantly, it reduces the noise from the original matrix. As empirical results in Information Retrieval (IR), we chose 300, 400 and 500 respectively as a value of k in our experiments. We find that best value of k is 500 in this our Thai word senses disambiguation experiment although 300 is generally chosen to be a good value of k in IR experiments [7]. In this task, we are only interested in the term matrix, T. Each row of T is a vector representation of the semantics of a particular word, in a k dimensional space. We can now compare the semantic distance between any two words by looking at the cosine of the angle of the two corresponding rows (vectors) in the matrix T. In this research, the cosine of the angles between the context vectors is used to calculate the correlation between words. The cosine measure for two vectors $\vec{x}$ and $\vec{y}$ can be calculated as follows:

$$Similarity = \cos(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}| \, |\vec{y}|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_1^2} \sqrt{\sum_{i=1}^{n} y_1^2}} \quad (1)$$

The advantages of using LSI with word meaning are that it is a fully automatic; corpus based statistical procedure that does not require syntactic analysis. LSI uses a fully automatic mathematical/statistical technique to extract and infer semantic relations between the meanings of words from their contextual usage in large collections of natural discourse. The analysis yields a mathematically well-defined representation of a word that can be thought of as a kind of average meaning.

IV. Experimental Methodology

We evaluate our method using sources of sense-tagged corpus. In supervised learning sense-tagged corpus is used to induce a classifier that is then applied to classify test data. Our approach, however, is purely unsupervised and the sense-tagged corpus is used to carry out an evaluation of the discovered sense groups.

Since there is no available Thai corpus-based, which contains Thai polysemous words in public, so in this research, we use a Thai corpus-based, which contains Thai polysemous words หัว /hua4/ and เก็บ /kep1/, were created by [9]. According to [9] the polysemous words and their contexts were randomly extracted from the corpus of "Bangkok Business" newspaper from November 1st, 1999 to October 31st, 2000 with the total size of 132 MB. The corpus contains sentences, which have sense of *หัว* /hua4/ and *เก็บ* /kep1/. The data contain 2,200 samples of *หัว* /hua4/ and *เก็บ* /kep1/ of each word. Each instance of *หัว* /hua4/ and *เก็บ* /kep1/ was hand-tagged with its sense defined in the Thai Royal Institute Dictionary BE (Buddhist Era) 2525. The characteristic of Thai text language is that there is no word boundary in Thai written text. Therefore, the collected data which contained the polysemous words *หัว* /hua4/ and *เก็บ* /kep1/ must be word-segmented. The segmentation was processed automatically by SWATH [18] which is a Thai word segmentation program from the NECTEC. The error correction was verified manually based on the context. The distributions of senses of *หัว* /hua4/ and เก็บ /kep1/ are presented in Table 1 and Table 2 respectively.

TABLE I
Definitions of Sense หัว /hua4/

Senses	Definitions
Head	Body part, which contains the brain.
Head of coin	Side of coin where a person's profile is represented, opposite of tails.
Intelligence	Ability of a person's brain.
View point	The way of a person views or thinks about as issue.
Talent	Talent or special ability to do something.
Top	Top part or pointed ends of and object.
Front	Front or pointing part of an object.
Early hours	The early hours or part of a time
Bulb	Globular base of stem of some plants sending roots downward and leaves upwards.
Concentrate	Concentrated substance.
Entity	Metonyms use of head to refer to an individual; extended metaphorically to refer to an organization.
Hair	Hair on the head of a human or an animal; hairstyle.
Brain	Brian. It also refers to the seat of consciousness, thought, memory and emotion.
Emotion	Emotional and psychological state.
Machine part	Vital part of machine. For example, part which pulls the rest of an engine, part that cuts or emanates sound or energy.
Chief	Top position of leadership, importance and honor, an individual holding these positions.
Heading	Information shown at the top of a page; title, heading; letterhead.
Headline	Headlines in newspaper.
Topics	Information represented in headlines, titles, and headings.
Titles or Names	Titles or names of newspapers, book and magazine.

TABLE II
Definitions of Sense เก็บ /kep1/

Senses	Definitions
To pick up	To pick something up from the ground or the floor.
To arrange	To put away; to arrange objects in a cabinet.
To take	To collect; to harvest; to take under one's care.
To keep	To keep or store, to prevent loss or damage.
To gather	To gather; to save.
To charge	To collect or to charge a fee.
To hide	To keep out of sight; to keep hidden from others; to hide.
To kill	To get rid of; to eliminate.
To purchase	To acquire; to buy in stock markets.

V. Results

The input of the testing component is the testing corpus, which is already segmented. The output is the most likely senses of words given by the WSD systems. The performance of the method was computed as precision rates by applying the following formula:

$$Precision\,Rates = \frac{Total\,number\,of\ correct\,answer}{Total\,number\,of\ answered\,senes} * 100 \quad (2)$$

A. Baseline system

As a baseline system, the most frequent sense (MFS) of a word is chosen as the correct sense. The frequency of word senses is calculated from the occurrences of the word senses in the corpus, with ties broken randomly.

B. Experimental Results

Table 3 and Table 4 show experimental results compare with the baseline system for the disambiguation of *หัว* /hua4/ and *เก็บ* /kep1/ respectively.

The first column of Table 3 and Table 4 are sense definitions. The number of sentences of occurrences of each sense is shown in the second column of Table 3 and Table 4. The final column shows the total number of correct answers that could be estimated correctly. Table 3 shows that the polysemous word หัว /hua4/ has 20 senses the percentage attained at 71.27 %. Table 4 shows that the polysemous word เก็บ /kep1/ has 9 senses the percentage attained at 75.58 %.

TABLE III
PRECISION RATE OF DISAMBIGUATION OF หัว /HUA4/

Sense Definitions	No. of Sentences	Precision (%)
Brain	138	77.5
Bulb	159	75.9
Chief	30	76.5
Concentrate	55	76.5
Early hours	41	70.4
Emotion	13	72.8
Entity	460	61.2
Front	133	69.8
Hair	37	78.7
Head	506	62.4
Head of coin	6	60.1
Heading	7	70.2
Headline	41	74.5
Intelligence	88	73.2
Machine part	50	77.2
Talent	5	63.7
Titles or names	56	70.9
Top	77	71.3
Topics	60	77.2
View point	238	65.4
	Average	71.27

Baseline = 23.00 %

TABLE IV
PRECISION RATE OF DISAMBIGUATION OF เก็บ /KEP1/

Sense Definitions	No. of Sentences	Precision (%)
To arrange	41	80.7
To charge	627	72.1
To gather	295	79.8
To hide	61	76.5
To keep	832	70.2
To kill	10	75.8
To pick up	7	79.8
To purchase	20	73.5
To take	322	71.8
	Average	75.58

Baseline = 37.81 %

As a result, it can be pointed out that less polysemous words are the better performances of the method. The reasons why less polysemous words have clearer sense indicators as is that their senses are not closely related. Different senses occur with a totally different context.

VI. Conclusion

In this paper, we have presented a novel method of Thai word sense disambiguation by using Latent Semantic Indexing (LSI). LSI has one property that is very attractive for processing in solving ambiguity word semantic. No knowledge sources are required for the analysis. Thus it eliminates the need for dictionaries. It needs neither any training or uses word sense tags from corpus. The use of unlabelled data is especially important in corpus-based natural language processing because raw corpora are ubiquitous while sense tags data are expensive to obtain. This is shown that it is a suitable method of further developing a Thai word sense disambiguation program.

The results from the research on word sense disambiguation of Thai polysemous word หัว /hua4/ and เก็บ /kep1/ are promising. The data are free running text and have large number of senses per word (twenty senses for หัว /hua4/ and nine senses for เก็บ /kep1/).

References

[1] E. Agirre and G. Rigau, "A proposal for word sense disambiguation using conceptual distance", *In Proc. the International Conference Recent Advances in Natural Language Processing*, Tzigov Chark, Bulgaria, 1995.

[2] M. W. Berry, S. T. Dumais and G. W. O'Brien, "Using Linear Algebra for Intelligent Information Retrieval", *SIAM: Review*, vol.37 no. 4, 1995, pp. 573-595.

[3] M. W. Berry, "Large Scale Singular Value Computations", *International J. Supercomputer Applications*, vol.6, pp. 13-49, 1992.

[4] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer and R. Harshman, "Indexing by Latent Semantic Analysis", *J. the American Society for Information Science*, vol. 41, 1990, pp. 391-407.

[5] R. O. Duda,. P. E. Hart and D. G. Stork, *Pattern Classification,* 2nd ed., Wiley, 2000.

[6] S. T. Dumais, "Latent Semantic Indexing (LSI) and TREC-2", *In Proc. 2nd Text Retrieval Conf. (TREC-2)*, March, 1994, pp. 105-115.

[7] P. W. Foltz, "Latent Semantic Analysis for text-based research", Behavior Research Methods, Instruments and Computers, vol. 28 no. 2, 1996, pp. 197-202.
[8] I. T. Jolliffe, *Principal Component Analysis,* Springer Verlag, 1986.
[9] W. Kanokrattanukul, "Word Sense Disambiguation in Thai Using Decision List Collocation", Master Thesis, Dept. Linguistics, Chulalongkorn Univ., 2001.
[10] T. K. Landauer and S. T. Dumais, "A Solution to Plato's Problem: The Latent Semantic Analysis Theory of the Acquisition, Induction, and Representation of Knowledge", *Psychological Review*, vol. 104, no. 2, 1997, pp. 211-240.
[11] C. Leacock, M. Chodorow and G. A. Miller, "Using Corpus Statistics and WordNet Relations for Sense Identification", *Computational Linguistics*, vol. 24, no. 1, 1998, pp. 147-165.
[12] G. A. Miller, M. Chodorow, S. Landes, C. Leacock and R. G. Thomas, "Using a semantic concordance for sense identification", *In Proc. the ARPA Human Language Technology Workshop,* 1994.
[13] H. T. Ng and H. B. Lee, "Integrating Multiple Knowledge Sources to Disambiguate Word Sense: An Examplar-Based Approach", *In Proc. 34th Annual Meeting of the Association for Computational Linguistics*, Santa Cruz, 1996.
[14] T. Pedersen and R. Bruce, "Distinguishing word senses in untagged text", *In Proc. 2nd Conf. Empirical Methods in Natural Language Processing*, 1997, pp. 197-207.
[15] J. I. Saeed, *Semantics*, The United Kingdom, Blackweel Publishers, 1997.
[16] H. Schutze, "Dimensions of Meaning", *In Proc. Supercomputing*, 1992, pp. 787-796.
[17] G. Strang, *Algebra and its applications*, 2nd ed., Academic Press, 1980.
[18] "Smart Word Analysis for Thai", 2002, National Electronics and Computer Technology Center (NECTEC), Information Research and Development Division. [Online] Available: http:// www.links.nectec.or.th/.
[19] D. Yarowsky, "Unsupervised Word Sense Disambiguation Rivaling Supervised Methods", *In Proc. 33rd Annual Meeting of the Association of Computational Linguistics*, Cambridge, Massachusetts, 1995.
[20] U. Zernik, "Train1 vs. Train2: Tagging Word Sense in Corpus. Lexical Acquisition: Exploiting on-line Resources to Build a Lexicon", *In Proc. Recherche d'Informations Assistée par Ordinateur*, 1991, pp. 91-112.