CHAPTER 1

INTRODUCTION

Natural language processing (NLP) requires the resolution of various types of *ambiguity*, including, at the word level, the *syntactic* and the *semantic ambiguities*. The resolution of a word's syntactic ambiguity is usually solved by part-of-speech taggers which predict the syntactic category of words in text with high levels of accuracy (Brill, 1995). The problem of resolving semantic ambiguity is generally known as Word Sense Disambiguation and has been proved to be more difficult than syntactic disambiguation (Wilks, 2000).

The intended meaning of an ambiguous word can be determined by considering the context in which it is used. Given an ambiguous word used in a number of different contexts, word sense disambiguation or word sense discrimination is the process of identifying which of those contexts refer to the same meaning of that word. This ambiguous word under consideration is often referred to as the target word. The term context is used to refer to two or three sentences around the target word.

1.1 Motivation

The term *lexical ambiguity* refers to two different concepts: *homonymy* and *polysemy*. Homonymy is the case where two different words have the same lexical form, and polysemy is the case where one word has several related meanings. An example of homonymy is the distinction between bank ("river edge") and bank ("financial institution"). The meaning of "head" which can mean the upper or top part of our body or the top position is an example of polysemy word.

One way to assign the meaning to a word in a particular usage is to examine its context. For example, the English word *bank* (an extensively cited example of lexical ambiguity) can refer to the bank of river or to the financial institution. For this reason, a computer program analyzing the sentence "The boy leapt from the bank into the water" will need to decide which is the correct meaning for the sentence. Although human can resolve this problem very well, this is still a problem for automated NLP systems in practical applications, especially for systems that need to handle discourse in broad domains.

The supervised learning approach has been applied to solve word sense ambiguation successfully for last decade. Typically these approaches train a model by presenting it with some number of manually created sense tagged examples for a particular word. After training, these models are able to assign one of a predefined set of meanings to newly encounter of a target word. However, unsupervised approaches to word sense disambiguation is a different problem. Rather than trying to assign a target word to one of a set of possible meanings, it seeks to group together the words that are used in similar contexts. The motivation behind taking this approach is that a predefined set of meanings (as provided by a dictionary or similar resource) is often too inflexible and limited to account for word usages in actual text. In addition, sense tagged text only exists in small quantities and is expensive to create.

One of methods of unsupervised approaches that discover meanings of words from raw text based on contextual hypothesis or *distributional hypothesis* (Harris, 1968). This method is based on that *similar terms appear in similar contexts*. This hypothesis indicates a clear way of comparing words: by comparing the contexts in which they occur.

In this thesis, we take a corpus–based machine learning approach to achieve sense disambiguation. Our approach first learns a set of common word patterns observed in the context of ambiguous word in samples of text, and then discriminates given target word using clustering algorithm. The clustering algorithm which we used is the partitional clustering (*K-means* clustering algorithm) that automatically groups together the words using similar patterns in their contexts. The word patterns selected for making such distinctions are referred to as features. Thus, the output of a sense disambiguous system shows clusters of given text instances such that the words grouped in the same cluster are contextually more similar to each other than they are to the words grouped in the other clusters. As the instances in the same cluster use the target word in similar contexts, we can presume that they all have the same meaning. Thus, each cluster presents a distinct meaning or sense of that ambiguous word.

1.2 Objectives

The objective of this thesis is to develop a highly portable and easily adaptable methodology that is to automatically group occurrences of a word into clusters based from raw text. Each cluster consists of occurrences having same meaning. Thus, instead of using information from a dictionary or thesaurus, we refer to an available corpus of electronic text, and then automatically identify which words tend to occur together very often. According to the contextual hypothesis, words observed in similar contexts are semantically related.

Our contextual representation scheme is mainly based on the vector space model, which is introduced in the field of information retrieval. One major advantage of using this model is that any clustering algorithm can be applied to solve the problem. Although our appraoch is related to the vector model to create context representation which has been employed by Schütze (Schütze, 1992; Schütze, 1998), our work uses log-likelihood scores on the recorded co-occurrence frequencies which makes somewhat differ from Schütze's work. The Schütze's technique is to create a word co–occurrence matrix that employs frequency counts.

1.3 Scopes and Assumption

In this thesis, our method disambiguates senses of two Thai words and an English word. The two Thai words are *หัว* /hua4/ which is a noun and *เก็บ* /kep1/ which is a verb respectively. The English word is *interest* word. The outcome which we are interested in studying in this thesis work is the ability to group word occurrences having the same meaning into clusters and the ability to disambiguate the sense of interested words.

1.3.1 Scopes

The scopes of this research are:

1. In this thesis, we disambiguate senses of *หัว* /hua4/ which is a noun and disambiguate senses of *เก็บ* /kep1/ which is a verb. All other parts of speech ambiguity of *หัว* /hua4/ and *เก็บ* /kep1/ are excluded from this thesis since they can be resolved by a part of speech (POS) tagger.

2. In this research, we will generally use the term polysemy to refer to both lexical ambiguity types, because the focus of this research emphasizes on disambiguating polysemous words which appear in sentences.

   2.1 For Thai polysemous words, the sense that derives from the word form directly is considered. For example, *หัวหงอก* can be considered as two adjacent units, that is *หัว* 'hair" and *หงอก "gray"* or it can be considered as one unit *หัวหงอก* "old man" depending on its surrounding context. We will consider that the word form which are two adjacent units such as *หัว* 'hair" and *หงอก "gray"* has the sense of *หัว* (Kanokrattanukul, 2001).

   2.2 For English polysemous word, English corpus-based which is distributed by the Computing Research Laboratory (CLR) (crl.nmsu.edu) is used in this work. The data set consists of sentences from the ACL/DCI Wall Street Journal corpus that contains *interest* noun word.

1.3.2 Assumption

1. In this thesis, Thai dictionary of "Thai Royal Institute" is used for word sense analysis to create sense-tagged Thai corpus and Longman Dictionary of Contemporary English (LDOCE) (Procter 1978) for sense-tagged English corpus respectively. Both sense-tagged Thai corpus and sense-tagged English corpus are used only

in the system evaluation process. We use sense-tagged corpus to evaluate the maximum accuracy of discovered sense groups when we perform the experiments with test data.

1.4 Contribution

This thesis aims to develop a methodology of word sense disambiguation in Thai by using purely knowledge lean unsupervised learning techniques that do not rely on any knowledge intensive resources like sense–tagged text or dictionaries. The thesis approach is a novel since the work focuses on Thai language which has not received much attention in the Natural Language Processing literature. The methodology is a new highly portable and easily adaptable. The sense clusters are occurrences of a word that have been grouped into clusters based from raw text where each cluster consists of occurrences having same meaning. This thesis work also benefits for the further development of word sense disambiguation for Thai language.

1.5 Overview of the Coming Chapters

Chapter 2 describes theoretical background that is used in this thesis. The evaluation criteria of Word Sense Disambiguation are provided.

Chapter 3 gives an overview of previous work done in the field of Word Sense Disambiguation. By approach or strategy, we refer to the primary resource used to extract information about the different senses of words.

Chapter 4 presents our proposed method, Distributional Semantics approach which is based on the distribution hypothesis. It is the corpus-based approach to solve the problem of word sense disambiguation. The distribution semantics method is an unsupervised learning.

Chapter 5 presents the experimental setting and results.

Chapter 6 we conclude the research work and present some future research directions.