HYBRIDIZING PARTICLE SWARM OPTIMIZATION WITH SIMULATED ANNEALING FOR WORD SENSE DISAMBIGUATION

WAFAA RIYADH MUSA

UNIVERSITI KEBANGSAAN MALAYSIA

HYBRIDIZING PARTICLE SWARM OPTIMIZATION WITH SIMULATED ANNEALING FOR WORD SENSE DISAMBIGUATION

WAFAA RIYADH MUSA

DISSERTATION SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIRMENTS FOR THE DEGREE OF MASTER OF COMPUTER SCIENCE

FACULTY OF INFORMATION SCIENCE AND TECHNOLOGY
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2018

PENGHIBRIDAN KAEDAH PENGOPTIMUMAN KERUMUNAN ZARAH DENGAN SIMULASI PENYEPUIHLINDAPAN BAGI PENYAHTAKSAAN MAKNA PERKATAAN

WAFAA RIYADH MUSA

DISERTASI YANG DIKEMUKAKAN UNTUK MEMENUHI SEBAHAGIAN DARIPADA SYARAT MEMPEROLEH IJAZAH SARJANA SAINS KOMPUTER

FAKULTI SAINS DAN TEKNOLOGI
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2018

DECLARATION

I hereby declare that the work in this thesis is my own except for quotations and summaries which have been duly acknowledged.

19 June 2018

WAFAA RIYADH MUSA
P75877

ACKNOWLEDGEMENT

First and foremost, praise be to Almighty Allah for all his blessings for giving me patience and good health throughout the duration of this research. I offer my sincerest gratitude to my supervisor, Dr Sabrina Tiun, who has supported me throughout my thesis with her patience and knowledge. I attribute the level of my master degree to her encouragement and effort and without her this thesis, too, would not have been completed or written. One simply could not wish for a better or friendlier supervisor. Finally, I must express my very profound gratitude to my parents, my husband and specially Mr Ali Al-saffar for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

ABSTRACT

Words sense disambiguation (WSD), is a task of assigning the most appropriate sense to words is used in a sentence, when the word has multiple meanings. WSD relies on the context of the target word to identify suitable sense. Selecting semantic evaluator by using an optimization strategy, is the intermediate objective to identify the set of suitable senses. Optimization methods are either based on population of solutions or single solution. However, achieving an effective balance between exploration and exploitation is a challenging task in the optimization process. Hence, this study aims to improve partial disambiguation of a sentence and find global meaning for a given text. Therefore, hybridizes a population based on an algorithm named Particle Swarm Optimization with a local search algorithm called simulated annealing algorithm (SA). PSO provides a global search of the problem space that can find various solutions of different qualities. While, the local search algorithm works on intensifying the search locally, where, promising solution is processed in this algorithm to be improved by searching its neighborhood. The hybridized method evaluates the solutions based on the semantic relation among the words. In this study, the semantic relatedness and similarity methods, which are Extended Lesk's algorithm(e-Lesk) and Jiang-Conrath algorithm (JCN), are combined. The designed model in this research was experimented based on semantic concordance corpus (SemCor). Specifically, 19 files from this dataset, which have been used in the related works, as a benchmark dataset. Some of the related works presented their results based on only noun part-of-speech, and thus, this study did a comparison on only noun part-of-speech. While, the other comparison was based on all part-of-speeches. The proposed method outperformed other methods regarding the noun part-of-speech with f-measure of 73.36% (with increasing 0.24%). On all part-of-speech, the proposed method outperformed only at the precision metric with the highest result of 67.44% (0.41% improvement). Hence, it can be concluded that the proposed method be able to provides a good WSD solution for noun part-of-speech especially, as well as for other part-of-speech when not all words are required to be disambiguated.

ABSTRAK

Penyahtaksaan Makna Perkataan (PMP) adalah satu tugas di mana sesuatu perkataan diberikan makna yang paling sesuai pada sesuatu ayat, yang mana perkataan tersebut mempunyai banyak makna. PMP bergantung kepada konteks perkataan sasaran untuk mengenal pasti makna perkataan yang sesuai. Memilih penilai semantik dengan menggunakan strategi pengoptimuman adalah objektif perantaraan untuk mengenal pasti set makna perkataan yang sesuai. Kaedah pengoptimuman adalah sama ada berdasarkan kepada penyelesaian populasi (population of solution) atau penyelesaian tunggal. Walau bagaimanapun, bagi mencapai keseimbangan yang berkesan di antara eksplorasi dan eksploitasi adalah satu tugas yang mencabar dalam proses pengoptimuman. Oleh itu, tujuan kajian ini adalah untuk memperbaiki proses penyahtaksaan separa ayat dan mencari makna sejagat pada keseluruhan teks. Model yang dicadangkan menghibrid populasi berdasarkan algoritma Pengoptimuman Kawanan Separa (PKS) atau Particle Swarm Optimization (PSO) dengan algoritma carian setempat yang dipanggil algoritma Simulasi Penyepulihlindapan (SP) atau Stimulated Annealing (SA). PKS menyediakan carian sejagat pada ruang masalah yang dapat mencari pelbagai penyelesaian dengan kualiti yang berbeza. Sementara itu, algoritma carian setempat berfungsi untuk menggiatkan carian setempat dengan memperbaiki carian di kawasan yang berdekatan. Kaedah hibridisasi menilai penyelesaian berdasarkan hubungan semantik antara perkataan. Dalam kajian ini, kaedah kaitan semantik dan persamaan semantik, iaitu, algoritma Lesk yang diperluas (e-Lesk) dan algortima Jiang-Conrath (JCN), digabungkan. Model yang direka dalam kajian ini menjadikan korpus semantik konkordans (SemCor) sebagai bahan kajian. Secara khusus, sebanyak 19 fail dari set data tersebut, yang juga digunakan dalam kajian-kajian terdahulu, dijadikan sebagai set data penanda aras. Sebahagian dari kajian-kajian berkaitan melaporkan keputusan yang diperolehi berdasarkan golongan kata nama sahaja, dan dengan ini, kajian ini melakukan perbandingan pada golongan kata nama sahaja. Manakala, perbandingan yang lain adalah berdasarkan kepada semua jenis golongan kata. Metod kajian yang dicadangkan di dalam kajian ini melebihi jangkauan dari kajian-kajian terdahulu ke atas golongan kata nama, dengan ukuran-f 73.63% (peningkatan sebanyak 0.24%). Bagi semua jenis golongan kata, metod cadangan melebihi jangkauan metriks ketepatan dengan keputusan yang paling tinggi sebanyak 67.44% (peningkatan sebanyak 0.41%). Oleh itu, dapat disimpulkan bahawa kaedah yang dicadangkan berupaya memberikan penyelesaian PMP yang baik, khasnya golongan kata nama, dan juga pada semua jenis golongan kata, apabila tidak semua perkataan yang dikehendaki perlu dinyahtaksakan.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER I

INTRODUCTION

1.1 OVERVIEW

Words sense disambiguation is to identify the meaning of words in context in a computational manner (Agirre & Edmonds 2007). WSD is a vital issue in natural language processing (NLP) for years and it has been applied in various NLP tasks such as information retrieval, machine translation, and automatic summarization. The most important clue for WSD is the context of an ambiguous word. Feature words are selected from the context to determine the right sense of ambiguous word. Knowledge-based WSD usually selects the words in a certain length of window as feature words. Then, according to the relatedness between feature words and each sense of ambiguous word, the sense with max relatedness is selected as the right sense.

Two types of tasks can be distinguished in Natural Language Processing (NLP). The first one is the final tasks which perform for their own such as machine translation, information extraction and automatic summarization. The other type is the intermediate tasks, which perform to aid final tasks, such as part-of-speech tagging, identification of morphological root, parsing, and word sense disambiguation. As long as WSD is one of the intermediate tasks, so it will be useful for some final tasks such as machine translation and information retrieval.

Supervised corpus-based WSD suffers from the knowledge-acquisition bottleneck and it is not practical to gather adequate manually tagged corpora for all possible domains. Also, the use of dictionary based methods was a pointer to move from supervised WSD approaches however, dictionary based methods suffers from the paucity in information to disambiguate all words using knowledge-based approaches. Hence, unsupervised WSD comprises the most promise solution for WSD task.

Particularly, this study tends to use unsupervised framework that exploit semantic relatedness and similarity methods to find the most suitable meaning for vague words. This framework is implemented using swarm intelligence search algorithm. This algorithm maximizes the semantic relatedness and similarity depending on various methods that measure the aforementioned criteria.

Combining PSO and SA, learning from other's strong points to offset one's weaknesses each other, the Simulated Annealing Particle Swarm Optimization can narrow the field of search and speed up the rate of convergence continually in the optimizing process. It has higher searching efficiency. It can also escape from the local minimums. These two algorithms are applied to several test functions optimization problem and simulation shows that PSO with SA algorithm is much better(Javidrad & Nazari 2017).

1.2 RESEARCH BACKGROUND

This chapter is dedicated to review the state-of-the-art WSD methods. Generally, there are four conventional WSD approaches, i.e., supervised, semi-supervised, knowledge-based, and unsupervised approaches. Also, WSD there are some methods that integrate two approaches to reinforce the process of word disambiguation. Specifically, the unsupervised methods keen to invoke knowledge-based assets to gain more necessary features to facilitate the classification process (Rigau et al. 1997; Sinha & Mihalcea 2007). Alternatively, knowledge-based methods have been generalized by using unsupervised scheme to search for the suitable sense for bag of words (Cowie et al. 1992). The proposed study falls in the last type, and hence this study focuses on search methods that use semantic similarity or relatedness.

Unsupervised approaches are capable to keep avoid the knowledge acquisition bottleneck (Gale et al. 1992), i.e. the lack of extensive resources which are manually labelled with word senses. in the context of unsupervised WSD, Co-occurrence graphs is a method that puts forward a different view of word sense discrimination, which has been recently explored with a certain success. The method is based on the concept of a co-occurrence graph, i.e. a graph G = (V, E) where vertices V correspond to words in a text and edges E connect pairs of words which co-occur in a syntactic relation, in the same paragraph or in a larger context. This graph-based algorithm for large-scale WSD (Navigli & Lapata 2010) is a method which has few parameters and does not need sense-annotated data for training. This method examines several measures of graph connectivity in order to identify those best suited for WSD. Mihalcea (2005) proposed a graph-based algorithm for sequence data labelling by means of random walks on graphs encoding label dependencies.

1.3 RESEARCH QUESTIONS

How to find best meaning for a given sentence, is the primary concern of this research. Regarding to this concern three research questions are arising as follows:

1. How to solve the ambiguity problem for bag of words using an automatic method?
2. How to balance between diversification and investigation in the search process to find the global meaning of bag of words?
3. How to evaluate the accuracy of the disambiguation process?

1.4 RESEARCH PROBLEM

The task of determining word meaning automatically in computational linguistics is denoted by Word Sense Disambiguation (WSD). This task relies on the context of the target word to identify suitable sense.

There are two types of word ambiguity; the first type is polysemous words which are the words that have multiple senses with subtle differences. The other type is a homonymy words which are the words that have multiple senses, each sense related to specific domain (Samhith et al. 2016). Consequently, the task of disambiguating homonymy words is easier than disambiguating polysemous words. Disambiguating any type of the aforementioned word ambiguity requires an accurate quantifier that can measure the semantic relation between any two senses (Srinivas & Rani 2016). Otherwise, large annotated corpus which is an expensive and time-consuming effort needed to do the disambiguation process.

Selecting semantic evaluator is intermediate objective to identify the set of suitable senses finally. This is determined in our study using an optimization strategy. Several different meta-heuristics algorithms applied for WSD task (Zhang et al (2008); Hausman(2011); Abed et al(2015) and Abed et al(2016)) but they had not yet to discuss the PSO impact. Nevertheless, an efficient optimization model is supposed to lead the search process to global optimum solution which represented by global meaning of the sentence. Optimization methods are either based on population of solutions or single solution. Population-based algorithm explores a wide area of the search space, thus it is capable in the diversification procedure. However, these types of algorithms are not good in exploiting the search space in comparison to single solution algorithms. In consequence, it will be helpful to balance between exploitation and exploration in order to reach the global optimum (Ursem 2002; Alba & Dorronsoro 2005; Valizadegan et al. 2011; Mirjalili & Lewis 2016). However, achieving an effective balance between exploration (Particle Swarm Optimization) and exploitation (Simulated Annealing) is a challenging task in the optimization process. The model in this work is expected to achieve a satisfying result by using the combination of two meta-heuristic approaches; PSO (global search) and SA (local search) algorithms, in order to increase coverage of the hybrid PSO that means the balance of exploitation and exploration lead to global optimum of search space.

1.5 RESEARCH OBJECTIVE

The main aim of this research is to find the set of senses that carries maximum semantic relatedness or similarity for an ambiguous group of words. Hence, the following objectives have been determined:

1. To implement the efficiency of Particle Swarm Optimization (PSO) for word sense disambiguation.

2. To proposed improved hybrid PSO based SA that partial disambiguation of the sentence and find global meaning for the given text.

1.6 SIGNIFICANCE OF THE RESEARCH

In order to show the role of word sense disambiguation in Natural Language Processing (NLP), it is useful to differentiate the intermediate and final tasks. The latter are tasks which are carried out for their own usefulness (e.g. machine translation, automatic summarization, and information extraction); the former are tasks carried out to assist the final tasks (e.g. part-of-speech tagging, parsing, and the identification of morphological root and word sense disambiguation); these are tasks in which there is little interest in their results.

The benefits of intermediate tasks can be explored by looking at some of the final tasks with which they are likely to help. Particularly, WSD has been traditionally called to help with two tasks which are information retrieval and machine translation. Among the other tasks that WSD can help with are text summarization, question answering system and document classification.

1.7 RESEARCH SCOPES

The scope of this research is to develop a word sense disambiguation model based on unsupervised WSD. This approach combines unsupervised WSD with lexical based methods. This system has to identify the relation between each two words in a sentence. Thereafter, a combination of e-lesk and JCN as method for similarity measures and relatedness methods are proposed to measure the coherent score between each two related words. This score is augmented using WordNet domains. The designed model in this research was experimented based on semantic concordance corpus (SemCor). Finally, this system applies a population based meta-heuristic algorithm to maximize the coherent score (similarity or relatedness) for each sentence. In meta-heuristic algorithm hybridized a population-based algorithm named PSO with a local search algorithm called simulated annealing algorithm. In addition, exploring the mechanism of each algorithm and how they can work together for achieving high quality search method.

1.8 RESEARCH ORGANIZATION

This research is laid in five chapters including current chapter. The first chapter presents the introduction of this research that describes the research background, research problem, research questions, research objectives and significance of the research. Chapter two, in section (2.1 and 2.2) provide a review of the literature and explanation for approaches that were used to solve the WSD task. In section 2.3, several methods of the similarity between two terms had been explored. in section 2.4, related works of meta-heuristic for WSD were given in one table with its comprising. In addition, at the end of this chapter by section 2.5 discussed the hybrid PSO and SA algorithms from earlier works.

Chapter three, explained the research methodology of the study which has been adjusted to accomplish the objectives in section 3.1. In section 3.2, presents the WSD as numerical data to be solved using the swarm search algorithm. In section 3.3, focuses on the model design steps that includes (Reading Semcor files, Implementation of hybrid PSO, Data presentation, Perform the fitness function and Performance evaluation). In section 3.4 explored a common language resource called WordNet that used to find the taxonomic for English concepts in the model. in section 3.5, provided a description of the dataset used in the model. Section 3.6, mentioned to the measures in order to measure both the similarity and the relatedness for all part of speeches. This section (3.6) contains three subsections which are JCN measurement, extend Lesk's algorithm and the objective function that combines both methods. In section 3.7, The proposed method in this study consists of employing two types of meta-heuristic search algorithms (PSO and SA). Each type characterizes by special searching ability. This section explains the mechanism of each algorithm and how they can work together for achieving high-quality search method. In Section 3.8, explained the model evaluation of results from the related works on this benchmark dataset to use for the performance comparison. Finally in Section 3.9 concludes the chapter.

Chapter four presents and discusses the results of the proposed method that has been described in Chapter 3. Section (4.1 and 4.2), describes the dataset that used for evaluating the word sense disambiguation method. While Section 4.3 presents the experimental results of the proposed method. In Section 4.3, the results given in details to show the impact of the window size as well as to illustrate the effectiveness of the similarity methods. Also, the local search influence is highlighted by experimenting the proposed method with local search and without it other times. Finally, a conclusion of this chapter is given in Section 4.4.

Chapter five gives the conclusions of the research, contributions of the research and suggestions for future work.

CHAPTER II

LITERATURE REVIEW

2.1 INTRODUCTION

In a given text, the most appropriate senses are assigned to words with the help of word sense disambiguation (WSD) (Raganato et al. 2017). A predefined sense inventory is employed to assign senses in stand-alone WSD systems. This sense inventory has a collection of predefined senses installed to it that can be applied to different words of a certain language, i.e. a lexicon, a dictionary, or a wordnet. Even though there exist some criticisms and problems, e.g. for sense distinctions at the granularity level, the most popularly employed sense inventory is the Princeton WordNet (Navigli 2009) for English WSD. WordNet’s availability and coverage has made it a popular choice for use as a sense inventory in WSD (Bhingardive & Bhattacharyya 2017).

WSD is regarded as a core task of NLP (Ide & Véronis 1998); (Navigli 2009) since such systems have offered considerable benefits for many NLP areas such as paraphrasing (Rus et al. 2009), information retrieval (Zhong & Ng 2012), knowledge extraction (Hassan et al., 2006; Ciaramita and Altun, 2006; Navigli and Ponzetto, 2010; Hartmann et al., 2013), sentiment analysis (Rentoumi et al., 2009; Martın-Wanton et al., 2010; (Balamurali et al. 2011), knowledge extraction (Hassan et al. 2006); (Ciaramita & Altun 2006); (Navigli & Ponzetto 2010); (Hartmann et al. 2013), and semantic role labeling (Che et al. 2010)). In the early years of machine translation, the system faced performance issues regarding sense disambiguation. However, since then, the system has benefited in many ways with the more recent machine translation systems ; (Vickrey et al. 2005); (Carpuat & Wu 2007); (Chan et al. 2007).

Considering the widespread utilisation of WSD systems, additional enhancements to WSD performance can provide consequent advantages in many NLP areas.

Understanding the meaning of the words being used is often a prerequisite in understanding natural language meaning. For example, in the following sentences, consider the different meanings of the word ‘bass’:

- I sing bass in the choir.
- I went fishing and caught a bass.

In the first sentence, the word bass denotes a type of singing voice, while in the second, it is referring to a fish. Here, correct interpretation of the word bass is crucial to understand and represent the meaning of both sentences. This is especially important in tasks such as summarization and translation to transform the text itself based on its real contextual meaning.

WSD systems face two main challenges. The first one is the sparsity of sense annotated data. Normally, annotation of word uses with senses is carried out to build an established sense inventory, such as Onto Notes (Hovy et al. 2006) or WordNet (Fellbaum 1998). A clear description of the meaning and sometimes syntactic information is provided through these inventories to help in understanding how the sense is used. However, since producing annotated examples involves many resources and a large number of unique words in a language, most words are typically linked with only tens of sense-annotated examples, where few words contain more than several hundred annotated instances. For example, SemCor (Miller et al. 1993), the largest sense-annotated resource, comprises only 234,157 annotated examples. Due to this sparsity, many WSD approaches include techniques such as unsupervised methods that disambiguate by utilizing knowledge from the sense inventory itself (Pedersen & Kolhatkar 2009) and (Navigli & Lapata 2010) or those methods that automatically generate additional examples of sense-associated features (Agirre et al. 2001)and (Zhong & Ng 2010). However, even with such additions, supervised methods marginally outperform the baseline in automatically selecting the most-frequent sense of a word (Navigli et al. 2007); (Pradhan et al. 2007), while unsupervised approaches most of the time perform poorer compared to supervised approaches.

The second challenge arises from the task difficulty level itself and in an effort to distinguish amongst senses. In the easiest setting, consider a homonym word that has unrelated meanings, such as the example ‘bass’ as in fish and vocal senses, while ‘bank’ with its senses related to financial institution and a sloping surface bordering water. Since these senses are typically unrelated, specific contextual cues associated with each (e.g. “fishing” or “music” for the senses of bass) can be exploited to easily disambiguate. Moreover, WSD systems can easily learn these cues due to distinctness of contexts where unrelated senses appear. However, there are many words that have senses that could be related in some or the other way. With reference to our earlier example, bass may be associated with both, as a singing voice as well as a musical instrument (e.g. a double bass). Two challenges stem due to this sense relatedness. First, both meanings would appear in similar music-related contexts due to their relatedness, which augments the difficulty of learning features that could differentiate between the two. In addition, the sparsity of sense-annotated examples could further compound this difficulty, which could have potentially distinguished key contextual differences amongst such senses. Second, the senses themselves could be indistinguishable from one another in some contexts. For example, consider the context ‘At the concert last night, the bass was too loud’. Here, the word ‘bass’ obviously refers to a musical entity and not the fish in any ways, but the context itself is not clear as to which musical entity is being referred. Also, this second challenge poses challenge for human annotators when creating sense-annotated corpora and could result in annotator disagreements, which results in increasing time and costs needed to build such corpora (Palmer et al. 2007).

Over the years, many systems have been proposed and developed for the WSD task. Through the late 1990s, a thorough review of state-of-the-art is presented in (Ide & Véronis 1998) and more recently in (Iacobacci et al. 2016). Various techniques have been employed to solve the problem ranging from unsupervised and supervised machine learning techniques to rule based/knowledge based approaches. This chapter emphasizes on the unsupervised optimization task, where disambiguation of all the content words in a document is required by the systems.

2.2 WSD APPROACHES

In computational linguistics, right from the 1950s up to recent years, WSD has been an active area of research (Ide & Véronis 1998); (Agirre & Edmonds 2007); (Navigli 2009). Most of the work on WSD has been on English language (Kilgarriff & Palmer 2000). The lack of appropriate resources, especially in the form of sense-annotated corpus data, has been one such factors affecting WSD research for other languages. WSD systems consider sense-annotated corpora as gold standards for training, evaluation and development. As such, a steady progress in the performance and development of WSD algorithms has not been a surprise for languages such as English, for which there are many large sense-annotated corpora, and considerably less on languages that have lesser availability of such corpora. All machine learning approaches commonly use corpora as knowledge sources; however, they differ in the exact task they perform. Many different approaches had been used to solve the WSD task. Those approaches can be categorized into knowledge-based approaches and machine learning-based approaches, of which the former is further categorized into supervised, unsupervised and semi-supervised approaches (Brown et al. 2014).

One of the main advantages of unsupervised machine learning approaches is being independent of sense-annotated corpora as they employ non-annotated corpora to cluster word senses, which make them least affected by the knowledge acquisition bottleneck. However, the downside of this approach is the much difficulty involved in the evaluation of the word sense induction task when compared with the WSD classification task. The main reason behind this difficulty is the lack of clear criteria on judging the quality of word sense clusters (Navigli 2009). The next subsections will explain WSD approaches:

2.2.1 Supervised Machine Learning Approaches to WSD

According to Eggebraaten et al. (2014), the last twenty years have garnered interest within the NLP field on machine-learning based approaches relating to the utilization of automatic classifications for word definitions. (Wang et al. 2014) attributed this to the identification of an increasing number of supervised WSD approaches. According to (Chasin et al. 2014), the WSD approaches that are currently monitored have identified the biggest number of algorithms that are being used for disambiguation. A monitored WSD includes machine-learning techniques within the sense-annotated data in order to improve the classification of word meanings.

Supervised approaches to WSD make use of supervised machine learning methods to correctly assign senses to a word. The task at hand could result in a classification problem, where the class requiring prediction could be the corresponding word sense (from a given sense inventory). Since for each lemma, there is a difference in the sets of word senses as well as of classes to be predicted, for each lemma, classification and training of supervised WSD systems are performed separately (Ng & Lee 1996), (Veenstra et al. 2000), (Hoste et al. 2002), (Martínez 2007), and (Dinu & Kübler 2007). Classifying each word lemma separately is also referred to as word-experts (Berleant 1995).

To learn predicting corresponding word senses in the case of unseen words, supervised classification algorithms depend on corpora where words have already been annotated with senses through a given sense inventory. As such, each occurrence of annotated word is designated as an instance. A certain amount of sense annotations is required for training a supervised method on predicting correct senses even for unseen words; another set of sense annotations is employed to determine the performance of the automatic disambiguation prediction. The availability of sense-annotated corpora is generally limited as these have to be constructed manually and is quite an expensive process.

However, several word experts have assigned the intended meaning to a word. The algorithms training set is one which uses a manual process to target a word through the integration of a definition derived from the dictionary. This seeks to identify that algorithm under supervision by incorporating the target word process within the WSD. Individual algorithms utilize specific features that develop a correlation with a meaning pertaining to training (Wang et al. 2014). This enhances the development of a common thread of functionality within the algorithms under supervision. This has led to the identification of enhanced results from supervised systems compared to unsupervised systems, as indicated through experiments, together with the international evaluation exercises incorporated by Senseval.

The issue has to do with the features design, as the features seek to obtain the required information and knowledge related to the context in which the target words for disambiguation are used. The computational requirements of the learning algorithms combined with the availability of the information have several limitations as to the features that are to be examined. This results in the codification of the generalized elements within the word sense instances.

Normally, a difficult pre-processing step is incorporated to enhance the development of a features vector in relation to the individual context examples. The employment of a windowing scheme or a sentence-splitter is included in this step to assist in the choice of the necessary context. (Escudero et al. 2000) discussed in detail how the efficiency and accuracy of two WSD learning methods are influenced by the features representation. (Agirre & Martinez 2001) conducted a survey on the types of knowledge sources that may be relevant for the codification of training examples. As a result, the feature sets that are usually used in the supervised WSD are categorized into three main groups: local features, topical features and syntactic dependencies.

The contexts used for these annotated words offer linguistic clues that are specific to particular senses. To disambiguate between word senses, supervised WSD systems employ these features or clues. That is, a feature can be considered as a distinct bit of information for encoding linguistic clues through the context of a target word (McCarthy 2009), such as -occurring words or word classes, structural information from the sentence or morphological information for the target word. A set of features containing values is specific to the instance of represented word. Some of the forms of machine learning features are given below: