Journal of Computer Science

Review

Word Sense Disambiguation: Survey Study

[1]Ahmed H. Aliwy and [2]Hawraa A. Taher

[1]Department of Computer Science, Computer Science and Mathematics Collage, Najaf, Iraq
[2]Department of Computer Science, Faculty Computer Science and Mathematics, Najaf, Iraq

Article history
Received: 15-05-2019
Revised: 05-07-2019
Accepted: 22-07-2019

Corresponding Author:
Ahmed H. Aliwy
Department of Computer Science, Computer Science and Mathematics collage, Najaf, Iraq
Email: ahmedh.almajidy@uokufa.edu.iq

Abstract: The process of identifying the correct sense of a given word in a particular sentence is called Word Sense Disambiguation (WSD). It is complex problem because it involves drawing knowledge from various sources. Significant amount of effort has been put into resolving this problem in machine learning since its inception but the toil is still ongoing. Many techniques were used in WSD and implemented on different corpora for almost all languages. In this paper, WSD algorithms were classified to three categories as Knowledge-based, supervised and unsupervised techniques. Each category will be studied in details with explanation of almost all the algorithms in each category. Hence work examples for each method were taken with the used language, the used corpora and other factors. The benefits and drawback of each method were recorded. Some of these techniques have limitations in some situations, therefore this work will helps the researchers in the field of natural language processing to select the suitable algorithms to solve their particular problem in WSD. The novelty of the work can be seen in the comparison of the used works and the used algorithms. From this work, it was concluded that (i) some methods give high accuracy for language but low for other, (ii) the size of the used data set affects the performance of the used algorithm, (iii) some of these approaches can be run fastly but with limitation of the accuracy and (iv) most of these approaches are implemented for many languages successfully.

Keywords: Word Sense Disambiguation, Applications of WSD, Classification of WSD Techniques

Introduction

In all major languages around the world there are many words that refer to different meanings in different contexts. These multi-sensory words are called “ambiguous words” and the process of extracting the true meaning of a mysterious word in a given context is known as “Word Sense Disambiguation”. Therefore; Word Sense Disambiguation (WSD) can be defined as a task of selecting the right sense from a predefined set of word senses according to the context.

WSD has long history where it was reinstated in the 1970s as part of Artificial Intelligence Research (AI) to understand the full natural language. A turning point in clarifying the meaning of words was introduced in 1980s. From this date till now the researchers try to use new algorithms for solving this task.

WSD is an important task and open problem in Natural Language Processing (NLP) field. WSD is necessary for many real world applications such as “Machine Translation (MT), Semantic Mapping (SM), Semantic Annotation (SA) and Ontology Learning (OL)”. It, also, can be helpful for improving the performance of many applications such as Information Retrieval (IR), Information Extraction (IE) and Speech Recognition (SR) (Zhou and Han, 2005) and many other applications. For example the English word "bank" has different senses such as financial institution and river side and reservoir etc. any sentence contain this word can not be translated to other language without knowing the correct sense in the sentence. Knowing the exact sense for this word, according to context, is not easy task therefore many algorithms and techniques were introduced for solving this important task (Pal and Saha, 2015).

Applications of WSD

There are a lot of applications needs to WSD which assist in understanding the components of text. Some of them will be mentioned in this section:

- **Machine Translation (MT):** It uses WSD for solving the ambiguity in word meaning in the sentence for getting exact translation. For example,

in the English sentences, (He scored a goal) and (It was his goal in life), can not be translated correctly without extracting the correct sense for the word (goal) because it has different senses in these sentences (Pal and Saha, 2015)

- **Information Extraction (IE):** WSD is used in IE and text mining for the accurate analysis of text. In general, a semantic analysis is very useful in IE because senses and synonyms play main role in it (Pal and Saha, 2015)
- **Content Analysis (CA):** WSD is very important phase in content analysis and can help to categorize data according to user requirements and solve many problems (Giyanani, 2013)
- **Information Retrieval (IR): IR** is one of the main real world applications for WSD. It is used for retrieving a set of documents that are semantically linked to a particular user query. The WSD help to increase accuracy IR (Sarmah and Sarma, 2016)

Related Works

There are many works in Word Sense Disambiguation; some of them give acceptable accuracy for different languages but little works was done for surveying these approaches. In this section, the related works will be shown. Zhou and Han (2005) summarized the different knowledge sources used in WSD as well as classified the current WSD algorithms according to their techniques. They also discussed the rationale, tasks, performance, knowledge resources used, computational complexity, assumptions and appropriate applications for each class of WSD algorithms. Haroon (2011) presented a survey on WSD which helps researchers and users to choose the algorithms and methods to solve their problems and in their specific applications. Pal and Saha (2015) presented a survey of different approaches adopted in different research works and the technical situation in the performance in this domain, these works focused in different Indian languages and finally a survey in Bengali language. Sarmah and Sarma (2016) discussed the task of WSD and the different approaches as well as their algorithms. They also explained various NLP applications that will be effective when integrating a deformation system. It also discusses evaluation measures used to determine performance WSD. This paper provides users with a general knowledge of the choice of WSD algorithms for their specific applications or to solve their problems.

WSD Approaches

Word Sense Disambiguation Approaches are classified into three main categories knowledge-based, supervised and unsupervised.

Knowledge-Based WSD

The Knowledge-based approaches are simpler compared to machine learning Methods where Machine learning require more training corpora instead of knowledge base and this approaches using external lexical resources such as (dictionaries, thesaurus, wordnet, etc.) (Haroon, 2011). Knowledge-based WSD consists of lots of methods where LESK algorithm is the most widely used method for solving WSD problems. **Benefit**: Knowledge-based WSD techniques can be used for solving the complex phenomena in the languages. It also provides practical resources for WSD from different concepts. **Drawback**: It needs to hard work of experts in linguistics, Not reflect the exact phenomena in practical text (corpora) and Inability to achieve expected performance results for some methods.

LESK Algorithm

LESK algorithm identifies simultaneously the correct senses for all words in the context using definition overlap. It was explained and implemented, by many researchers, for different languages. This method was suggested by Lesk (1986) where the meanings of each word are compared with the expressions found in the phrase. The original Lesk algorithm "measures overlap among sense definitions for all words in the text and identify simultaneously the correct senses for all words in the text" (Sarmah and Sarma, 2016; Haroon, 2011). Basile *et al*. (2014) applied this algorithm, using WordNet, Wikipedia and BablelNet, on English language. Pal *et al*. (2017) applied this algorithm using Bengali WordNet on Bengali language where synset hierarchy in the Bengali WordNet is not available. Bakhouche *et al*. (2015) applied this algorithm, using the data corpus of Elwatan, on Arabic language. **Benefit**: "A major advantage the Simplified Lesk algorithm over the original Lesk algorithm is that it is much faster to run, as it has a significantly lower computational time complexity. Simplified Lesk algorithm is also more accurate in disambiguating word senses". **Drawback**: are it needs a huge knowledge base and the original method cannot be used practically.

Semantic Similarity

And one of the used Knowledge-Based methods to solve WSD. And much less between text segments consisting of two or more words. The most researchers is focused on word-to-word analogy result from the availability of resources that define relationships between words or concepts, for example (WordNet) (Jiang and Conrath, 1997). Semantic similarity measures can be used to perform many tasks such as clarifying ambiguity and checking patterns for consistency or coherence. In general, all measures can be based into four categories: Information content, features, path length and hybrid measures (Meng *et al*., 2013). Jiang

and Conrath (1997) applied this method using WordNet on English language and they insure that IR application would benefit more from the semantic similarity measures when both document and query are short. Mihalcea *et al*. (2006) used this method using Corpus-based Measures, (i) point wise mutual information and (ii) latent semantic analysis, also on English Language. Pal *et al*. (2017) applied this algorithm using Bengali WordNet on Bengali Language. Karthikeyan and Udhayakumar (2015) applied this method using private data on English Language. **Benefit:** Semantic Similarity is the ability to provide harmony for the whole discourse. The smallest distance between two words mean they are to semantically related. When more than two words are considered, the approach will be intensive in arithmetic. **Drawback:** Semantic similarity of any two concepts with the same path length is the same uniform distance problem.

Selectional Preferences

It is a knowledge based method works by finding the information about potential relationships of the types of words and refers to common sense using the source of knowledge. Selectional preferences is antother method used to solve WSD problem by many researchers. Selectional preferences are given in terms of semantic classes rather than simple words (Pal and Saha, 2015). The methods of Selectional Preferences is represent some restrictions on the semantic type and impose the meaning of the word on the target word, which is collected through the grammatical relationship in the sentence (Sreenivasan *et al*., 2018; Sarmah and Sarma, 2016). Agirre and Martinez (2001) applied selectional preferences using Machine Readable Dictionary (MRD), light-weight ontologies or hand-tagged corpora. **Benefit**: From advantage of the selectional preference is avoiding a heavy cost in human time for manual tagging or computer time for unsupervised training. **Drawbacks**: It is difficult to determine the grammatical relationship between words in the selected text.

Heuristic Method

Heuristic Method is classified as knowledge-based methods, Where the linguistic properties is the core of this method, used to obtain the correct meaning of the ambiguous word. The conclusion of the linguistic properties is evaluated and three types of inference are used to estimate the system of WSD where are (i) more frequent sense, (ii) one sense per collocation, (iii) one sense for each speech. The heuristics of different linguistic properties is estimated to determine the exact meaning. A word will preserve its meaning among all its cases in a text within the meaning of the discourse category. One sense for each grouping is the same one sense in speech except that the closest words provide strong and consistent references to the meaning of the word (Sreenivasan *et al*., 2018). Pal *et al*. (2017) applied this algorithm using Bengali WordNet on Bengali Language. **Benefit**: It is used for conduct usability testing to further examine potential issues.

Drawback: It requires knowledge and experience to apply the heuristics effectively. It is expensive for designers.

Walker's Algorithm

Walker's algorithm is a thesaurus based approach. This algorithm starts by finding the synonyms to which this meaning belongs and then calculates the result for each sense by applying words of the context words. It will add 1 to the sense of whether the synonyms of the word are identical to the meaning of this sense (Sarmah and Sarma, 2016; Haroon, 2011). **Benefit**: This method gives high resolution because it relies on synonyms. **Drawbacks**: It is difficult to identify synonyms that help solve the problem of ambiguity of the word.

Supervised WSD

Supervised methods are machine learning technique based on manually created sense-annotated data. A training set, consists of target word related examples, will be used for the classifier. The main task is to construct a classifier that classifies the new cases correctly and accurately based on the context of their use Giyanani, 2013). A supervised WSD approach includes two parts: "(i) converting each training instance of an ambiguous word into a feature vector and (ii) applying a supervised learning algorithm after encoding all training examples in feature vectors" (Liu *et al*., 2001). Many of well-known supervised Methods of WSD will be explained in this section. **Benefit**: almost all the used algorithms are language independent. **Drawback**: supervised WSD algorithm has limitation to the learning data and hence unknown work or unknown sense for this word will be problem.

Decision List

Decision List was widely used in many applications and tasks where one of these task WSD. The decision list is a set of rules (if-then-else) in an ordered list format. It works by representing the concepts as lists of decisions and then applying series of tests on each vector; if the test succeeds, The sense associated with this test will be returned; if the test fails, The next test will be applied in the sequence and, It continue until the end of the list. The list of decisions includes learning the classification and arrangement of individual tests depending on the characteristics of the training data (Rivest, 1987). Chatterjee (2012) implemented Decision List using the SENSEVAL corpus on English language. Liu *et al*. (2001) applied Decision List algorithm, using data Sense-Tagged Corpora Extracted from the Clinical Data Repository and the MEDLINE Abstracts, on English Language. Agirre and Martinez (2004) implemented Decision List using multiple corpora as Web corpus

(Semcor bias), UNED, Web corpus (Autom. bias), Kenneth Litkowski-clr-ls, Haynes-IIT2 and Haynes-IIT, on English Language. **Benefit**: The method can be used in almost all classification problems including disambiguation of the word that has more than the meaning and give good results because they apply a series of tests until reach the correct result. **Drawbacks**: It suffers from over fitting problem.

Decision Tree

One of the most well-known used technique, in classification, It is used for WSD by selecting the desired concept using Yes-No tree. A decision tree is a binary tree where each internal node is categorized by a variable, each sheet is classified with 0 or 1. The depth of the decision tree is the longest path length from the root to the leaf (Rivest, 1987). The training examples are divided by using the highest-gain feature and the process is repeated to get good DT. Lee and Ng (2002) implemented DT and other approaches using SENSEVAL-2 and SENSEVAL-1 data on English Language and found that SVM performs best without feature selection, while the NB performs best with some feature selection. AL_Bayaty and Joshi (2014) applied Decision Tree using senseval-3 data on English Language. They found few words provide accurate results and hence the overall accuracy of this approach is very low (45.14%). **Benefit**: DT is an effective method and Robust approach to filter data if the size of the tree is small. And Easy to understand and interpret. People can understand decision tree models after a brief explanation. **Drawback**: DT is difficult and complicated process in case of data maintenance. They are unstable, which means that a small change in data can result in a significant change in the optimal decision tree structure.

Naïve Bayes

Naïve Bayes approach is applied successfully for many application and task in many fields because of its efficiency and ability to combine evidence from a large number of features. It can be applied if the workbook depends on a series of features. Naive Bayes chooses the category, in our methodology sense, with the highest probability. It can be works by gathering information from the surrounding words of the target word (Kalita and Barman, 2015). Naïve Bayes Method is the simplest representative of probabilistic learning algorithms therefore it can be used to classify the ambiguous words. Liu *et al*. (2001) applied Naïve Bayes algorithm, using data "Sense-Tagged Corpora Extracted from the Clinical Data Repository and the MEDLINE Abstracts, on English Language". It chooses the row with the highest back-end probability. Lee and Ng (2002) implemented NB using SENSEVAL-2 and SENSEVAL-1 data on English Language. El-Gamml *et al*. (2011) implemented Naïve Bayes Classifier using very small Private data (lexical samples of five words) on Arabic language. **Benefit**: Naïve Bayes is Very simple, easy to implement and fast, Need less training data, Can make probabilistic predictions. **Drawback**: A problem occurs because of the paucity of data. To obtain any potential value for a feature, you need to estimate the potential value by an iterative method.

Neural Networks

Neural network is an approach from supervised methods which simulate interconnection of artificial neurons. Artificial neurons are used to process data using a genetic approach. The input of the learning program is the input feature. The goal is to divide the training context into non-overlapping groups. Neural networks are used to represent words by contract and these words will activate the ideas associated with them. Inputs are transferred from the input layer to the output layer through the middle layers. The input can be easily deployed over the network and manipulated to obtain outputs, but it is difficult to obtain a clear output (Sreenivasan *et al*., 2016). NN WSD uses two scoring components that contribute to the final result of a string (word sequence, document). The scoring components are calculated by two neural networks the first is that captures the local context and the second is global context. Huang *et al*. (2012) applied NN using WordSim-353 data on English Language. They concluded that new multi-prototype neural language model perform better than previous neural models on the new dataset. **Benefit**: Storing information on the entire network, Ability to work with incomplete knowledge and Parallel processing capability. **Drawback**: NN is hardware dependence and need to parallel processing unit.

Exemplar/Memory Based Learning

In this way, all examples are stored in memory and called memory Because of the addition of new examples, new forms are not created, but gradually added to the current model. The most common approach to apply this method is K-Nearest Neighbor (kNN). "It is one of the best performing Exemplar based learning in WSD". It uses distance to measure proximity If k is greater than 1, the resulting meaning is in the sense of the majority of the nearest neighbors (Chatterjee, 2012). Escudero *et al*. (2000) implemented Exemplar/Memory-Based using data sense–tagged corpus on English Language. Their results showed that Exemplar–based algorithm have generally better performance than Naive Bayes algorithm. Also they showed that value of k of the nearest neighbors have a significant impact on the accuracy of the model-based classifier. Ng (1997) applied this algorithm using WordNet on English Language. He concluded that the accuracy achieved by his improved exemplar-based classifier was comparable to the accuracy on the same data set obtained by the Naive-Bayes algorithm. **Benefit**: Develop long-term knowledge retention. **Drawback**: The poorer performance potential on the tests.

Support Vector Machines

Support Vector Machines were introduced at 1992. It depends on the idea of learning a hyperplane using set of the training data. The hyperplane separates positive and negative examples. It maximizes the distance between the closest positive and negative examples (called support vectors) (Chatterjee, 2012). The Support Vector Machine implement optimization to find a hyperplane that separates training examples. Lee *et al*. (2004) applied SVM, using SENSEVAL-3, on English Language. The evaluation results, on the English lexical sample, refer to that their method achieves good accuracy on this task. **Benefit**: "It has a regularization parameter, which makes the user think about avoiding over-fitting and it uses the kernel trick, so you can build in expert knowledge about the problem via engineering the kernel". **Drawback**: SVMs is the lack of transparency of results. Also, kernel models can be quite sensitive to over-fitting the model selection criterion.

AdaBoost

AdaBoost Method is a way to create strong classifiers through the linear set of weak classifications. This method finds the cases that are incorrectly classified from the previous classifier so that they can be used for forthcoming work. The classifiers are learned from the weighted training group and at the beginning all weights are equal. At each step it performs some repetitions where in each repetition, the weight of the corrected work is increased so that the other two classifiers can focus on those incorrect examples .The basic idea of repetition is to give more weights to misclassified training examples, which makes the new classifier focus on those examples that are difficult to classify (Sreenivasan *et al*., 2016). Lee and Ng (2002) applied AdaBoost, using SENSEVAL-2 and SENSEVAL-1 data, on English Language. **Benefit**: AdaBoost can achieve similar classification results with much less tweaking of parameters or settings **Drawback**: it has large time complexity, it can be sensitive to noisy data and outliers and difficult to implement in real-time platform.

Unsupervised WSD

The unsupervised algorithms does not require a training corpus and not require long computing time and power (Zhou and Han, 2005). **Benefit** can be get unlabeled data from a computer easier than labeled data. and can be less complexity in comparison with supervised classification. **Drawback**: "It has worse performance than the supervised approach because it depends on less knowledge and the input data is unknown" (Fulmari and Chandak, 2013).

Context Clustering

Context clustering depends on the techniques of the groups. These groups is represented as either similarity matrix or context vectors which are created and then grouped into clusters to determine the meaning of the word. Purandare and Pedersen (2004) applied context clustering, using "sense–tagged instances of 24 SENSEVAL- 2 words and the well-known (Line, Hard and Serve sense–tagged) corpora", on English Language. They showed that, when smaller amounts of data were given for example "(SENSEVAL-2), second order context vectors and a hybrid clustering method like Repeated Bisections is perform better". **Benefit**: It can be used without any knowledge and then it can be used as a base for construction or annotated corpora. **Drawback:** Selection of proper features of the documents to be used in clustering.

Word Clustering

Word Clustering is a technique where the words are clustered according to the semantic similarity based on features. It verifies the identical words that resemble the word goal and the similarity between these words is calculated from the features they share. The identical words are seems to share the same type in the group. Hence, the clustering algorithm is applied to the distinction between the senses. If a set of words is taken, the similarity is determined first by the measures. The words are then arranged in order of similarity and create a similarity tree (Sreenivasan *et al*., 2016). The similarities between target word and Context words are determined based on their grammatical properties information (Sarmah and Sarma, 2016). Wanton and Llavori (2012) applied this method, using subset of SemCor 2.0 composed by all the documents of brown1 and brown2 corpora, on English language. "It contains a total of 192,639 words tagged with WordNet 2.0 senses. In the case of Senseval-3, they use the all-words corpus composed by 2081 words annotated with WordNet 2.0".

Co-Occurrence Graph

Co-occurrence Graph is a method based on a graph-based unsupervised learning. It can be used for detection of the meaning of a targeted word. The graph headers are words in the context with the target word (to be disassembled) and combined with an edge if they occur in the same paragraph (Sarmah and Sarma, 2016). Co-occurrence diagram "creates co-occurrence of the graph with the edge of E and the vertex V, where V represents the words in the text and the E is added if the words co-occur in the relationship according to the syntax in the same paragraph or text. For a specific target word, the graph is firstly created and the adjacency matrix is determined for the graph. After that the Markov assembly method is applied to find the meaning of the word" (Pal and Saha, 2015). Hassel (2005) implemented Co-occurrence, using two data set from Stockholm-Ume°a Corpus which are Swedish Parole corpus and WSD training set, on Swedish Language. Klapaftis and

Manandhar (2010) applied this method, using the SemEval-2010 WSI task dataset, on English language. They, also, applied HRGs to other related tasks such as taxonomy learning. **Benefit:** Co-occurrence can generate strong features by assembling weak features. **Drawbacks:** It does not include bookmarks that are annotated with single sense.

Conclusion

Interpreting the meaning of a particular word is a difficult task because it involves complete complexities of language and depends on unorganized text sources. This paper discussed the used methods in word sense disambiguation and discussed the previous papers that presents a survey on WSD approaches which helps the researchers in field of natural language processing to select the algorithms to solve their particular problem in WSD. From this survey, (i) any comparison cannot be accurate because each approach was applied on different data set with different sizes. (ii) some languages have eloquence phenomena, as in Arabic language, which affect the performance of the used algorithm. When numbers of papers were studied in WSD for doing this survey, the following points were concluded: Some methods give high accuracy for language but low for other, the size of the used data set affects the performance of the used algorithm, some of these approaches can be run fastly but with limitation of the accuracy and most of these approaches are implemented for many languages successfully. Finally we can construct a good WSD algorithm by taking in account the following points: The identical word meaning seems to have same neighbors, Some of the stop words can have affects in some circumstances therefore removing them can affect the accuracy, the position of the word for the ambiguous word can affect the meaning, POS is very useful in WSD. Finally, for surveying the selected algorithms and papers in this work, comparison tables were constructed as shown in Table 1 and 2. Table-1 shows comparison among all the mentioned algorithms in accuracy, the used language, benefits and drawbacks. Table 2 shows comparison among all the mentioned works which contains the author of work, the category of the used technique, the used method or algorithm, the used data set and the language.

Table 1: Comparison of different algorithms and techniques on WSD

Algorithm	Language	Precision average	Benefit	Drawback
Lesk Algorithm	English, Arabic, Bengali	65%	A major advantage the Simplified Lesk algorithm over the original Lesk algorithm is that it is much faster to run, as it has a significantly lower computational time complexity. Simplified Lesk algorithm is also more accurate in disambiguating word senses	It needs a huge knowledge base and the original method cannot be used practically.
Semantic Similarity	English, Bengali	75%	It provide harmony for the whole discourse and useful when there are smallest distance between two words.	Semantic similarity of any two concepts with the same path length is the same uniform distance problem. When more than two words are considered, The approach will be intensive in arithmetic
Selectional Preferences	different languages	78%	It is used for avoiding a heavy cost in human time for manual tagging or computer time for unsupervised training	it is difficult to determine the grammatical relationship between words in the selected text
Heuristic Method	Bengali	74%	It is beneficial because it allows intuitive leaps to be made when solving problems	It requires knowledge and experience to apply the heuristics effectively and it is expensive for designers
Walker's algorithm	different languages	75%	This method gives high resolution because it relies on synonyms	It is difficult to identify synonyms that help of solving the problem of ambiguity of the word
Decision List	English	58%	It can be used in almost all classification problems including disambiguation of the word that has more than one meaning. It can give good results because they apply a series of tests until reach the good classifier	It suffers from over fitting problem
Decision Tree	English	60%	DT is an effective method and Robust approach to filter data if the size of the tree is small. It is Easy to be understand and interpreted. People can understand decision tree models after a brief explanation	DT is difficult and complicated process in case of data maintenance. They are unstable, which means that a small change in data can result in a significant change in the optimal decision tree structure
Naïve Bayes	Arabic, English	83%	Naïve Bayes is Very simple, easy to implement, fast and it need less training data and can make probabilistic predictions	A problem occurs because of the paucity of data. To obtain any potential value for a feature, you English need to estimate the potential value by using an iterative method
Neural network	English		Storing information on the entire network, Ability to work with incomplete knowledge	NN is hardware dependence and need to parallel processing unit
Exemplar/ Memory Based Learning	English	74%	Developing long-term knowledge retention and sincerely reproduce all training data	The poorer performance potential on the tests And very slow specially if the there are huge number of examples
Support Vector Machines	English	67%	It has a regularization parameter, which makes the user think about avoiding over-fitting and it uses the kernel trick, so you can build in expert knowledge about the problem via engineering the kernel	SVMs is lack of transparency of results. Also, kernel models can be quite sensitive to over-fitting the model selection criterion
AdaBoost	English	65%	AdaBoost can accomplish similar classification results with much less tweaking of settings	It has large time complexity, sensitive to noisy data and outliers and difficult to be implemented in real-time platform
Context clustering And Word Clustering	English	70%	It can be used without any knowledge and then it can be used as a base for construction or annotated corpora, and can be capture the relations among those words with different part of speech tags	Difficult selection of proper features of the documents to be used in clustering
Co-occurrence Graph	English	80%	Co-occurrence can generate strong features by assembling weak features	It does not include bookmarks that are annotated with single sense

Table 2: Comparison of different works and methods on WSD

Author	Technique	Method	Data set	Language
1- Rivest (1987)	Supervised	Decision lists	Private	General
2- Ng (1997)	Supervised	Exemplar-Based	WordNet	English
3- Jiang and Conrath (1997)	Knowledge-Based	semantic similarity distance	WordNet (private)	English
4- Escudero *et al*. (2000).	Supervised	Naïve Bayes and Exemplar-Based	sense–tagged corpus	English
5- Agirre and Martinez (2001).	Knowledge-Based	knowledge types and actual information sources	-MRD, light-weight Ontologies, hand-tagged corpora	General
6- Liu *et al*. (2001).	Supervised	Decision Lists Naïve Bayes Exemplar-Based	Sense-Tagged Corpora Extracted from the Clinical Data Repository and the MEDLINE Abstracts	English
7- Lee and Ng (2002)	Supervised	knowledge sources, (SVM), NB, AdaBoost, and DT algorithms.	SENSEVAL-2, SENSEVAL-1	English
8- Purandare and Pedersen (2004).	Un Supervised	Clustering Contexts and Similarity Spaces	sense–tagged instances of 24 SENSEVAL-2 words, sense–tagged corpora.	English
9- Lee *et al*. (2004).	Supervised	SVM learning and multiple knowledge sources.	SENSEVAL-3	English
10- Agirre and Martinez (2004).	Supervised	Decision Lists	Web corpus (Semcor bias), UNED, Web corpus (Autom. bias),Kenneth Litkowski-clr-ls, Haynes-IIT2, Haynes-IIT1	English
11- Hassel (2005).	Un Supervised	Co-occurrence Graph	Ume°a Corpus	Swedish
12- Mihalcea *et al*. (2006).	Knowledge-Based	Corpus-based and knowledge-based measures of similarity	Corpus-based Measures	English
13- Klapaftis and Manandhar (2010)	Un Supervised	Co-occurrence Graph	SemEval-2010 WSI task dataset	English
14- Wanton *et al*. (2012).	Un Supervised	Word Clustring	SemCor 2.0 Senseval-3	English
15- Huang *et al*. (2012)	Supervised	Neural Network	WordSim-353	English
16- Meng *et al*. (2013)	Knowledge-Based	Semantic similarity	WordNet	English
17- Basile *et al*. (2014).	Knowledge-Based	Lesk algorithm	WordNet, Wikipedia, BabelNet	English and Italian
18- AL_Bayaty and Joshi (2014).	Supervised	decision tree	senseval-3	English
19- Karthikeyan and Udhayakumar (2015).	Knowledge-Based	Semantic Similarity	Private	English
20- Bakhouche *et al*. (2015).	Knowledge-Based	Lesk algorithm	corpus of Elwatan	Arabic
21- Pal *et al*. (2017)	Knowledge-Based	Lesk algorithm	Bengali WordNet	Bengali language
22- El-Gamml *et al*. (2011)	Supervised	Support Vector Machine (SVM) And Levenshtein Distance algorithm and Naïve Bayes Classifier (NBC)	Private(lexical samples of five Arabic words.)	Arabic anguage

Acknowledgement

The authors would like to thank the reviewers for their valuable comments and suggestions that contributed to the improvement of this work.

Author's Contributions

All authors equally contributed in this work.

Ethics

This article is original and contains unpublished material. The authors confirm that is no conflict of interest involved.

References

Agirre, E. and D. Martinez, 2001. Knowledge Sources For Word Sense Disambiguation. In: Text, Speech and Dialogue, Matoušek, V., P. Mautner, R. Mouček and K. Tašer (Eds.), ISBN-10: 978-3-540-42557-1, pp: 229-246.

Agirre, E. and D. Martinez, 2004. Unsupervised WSD based on automatically retrieved examples: The importance of bias. Proceedings of the Conference on Empirical Methods in Natural Language Processing, Jul. 1-3, ACL, Barcelona, Spain, pp: 25-32.

AL_Bayaty, B.F.Z. and S. Joshi, 2014. Empirical implementation decision tree classifier to WSD problem. Int. J. Adv. Technol. Eng. Sci., 2: 579-601.

Bakhouche, A., T. Yamina, D. Schwab and A. Tchechmedjiev, 2015. Ant colony algorithm for Arabic word sense disambiguation through English lexical information. Int. J. Metadata, Semant. Ontol., 10: 202-211.

Lesk, M., 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In Proceedings of the 5th annual international conference on Systems documentation (pp. 24-26). ACM.

Basile, P., A. Caputo and G. Semeraro, 2014. An enhanced lesk word sense disambiguation algorithm through a distributional semantic model. Proceedings of the Coling the 25th International Conference on Computational Linguistics: Technical Papers, Aug. 23-29,Dublin, Ireland, pp: 1591-1600.

Chatterjee, A., 2012. A survey of sense annotation, eye-tracking and wordnet linking approaches till June, 2012. Doctoral dissertation, Indian Institute of Technology, Bombay.

El-Gamml, M.M., M.W. Fakhr, M.A. Rashwan and A. Al-Said, 2011. A comparative study for arabic word sense disambiguation using document preprocessing and machine learning techniques. Proceedings of the Arabic Language Technology International Conference, Bibliotheca Alexandrina, (CBA' 11) Alexandria, Egypt.

Escudero, G., L. Màrquez and G. Rigau, 2000. Naive Bayes and exemplar-based approaches to word sense disambiguation revisited. Comput. Lang. Artificial Intel., 3: 421-425.

Fulmari, A. and M.B. Chandak, 2013. A survey on supervised learning for word sense disambiguation. Int. J. Adv. Res. Comput. Commun. Eng, 2: 2278-1021.

Giyanani, R., 2013. A survey on word sense disambiguation. IOSR J. Comput. Eng., 14: 30-33.

Haroon, R.P., 2011. Word sense disambiguation-A survey. Proceedings of the International Colloquiums on Computer Electronics Electrical Mechanical and Civil, (EMC' 11), ACEEE, pp: 58-60. DOI: 02.CEMC.2011.01.582

Hassel, M., 2005. Word sense disambiguation using co-occurrence statistics on random labels.

Huang, E.H., R. Socher, C.D. Manning and Ng, A.Y. 2012. Improving word representations via global context and multiple word prototypes. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long, Jul. 08-14, Jeju Island, Korea, pp: 873-882.

Jiang, J.J. and D.W. Conrath, 1997. Semantic similarity based on corpus statistics and lexical taxonomy. Proceedings of the 10th Research on Computational Linguistics International Conference, Aug. 5-7, Taipei, Taiwan, 19-33.

Kalita, P. and A.K. Barman, 2015. Word sense disambiguation: A survey.

Karthikeyan, R. and V. Udhayakumar, 2015. A web search engine-based approach to measure semantic similarity between words. Transact. Know. Data Eng. 23: 977-990. DOI: 10.1109/TKDE.2010.172

Klapaftis, I.P. and S. Manandhar, 2010. Word sense induction and disambiguation using hierarchical random graphs. Proceedings of the Conference on Empirical Methods in Natural Language Processing, Oct. 09-11, Cambridge, Massachusetts, pp: 745-755.

Lee, Y.K. and H.T. Ng, 2002. An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, (NLP' 02), Stroudsburg, PA, USA, pp: 41-48. DOI: 10.3115/1118693.1118699

Lee, Y.K., H.T. Ng and T.K. Chia, 2004. Supervised word sense disambiguation with support vector machines and multiple knowledge sources. Proceedings of the 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text SENSEVAL-3, Jul. 4, Barcelona, Spain, 137-140.

Liu, H., Y.A. Lussier and C. Friedman, 2001. Disambiguating ambiguous biomedical terms in biomedical narrative text: An unsupervised method. J. Biomed. Informat., 34: 249-261.

Meng, L., R. Huang and J. Gu, 2013. A review of semantic similarity measures in wordnet. Int. J. Hybrid Informat. Technol., 6: 1-12.

Mihalcea, R., C. Corley and C. Strapparava, 2006. Corpus-based and knowledge-based measures of text semantic similarity. Proceedings of the 21st national conference on Artificial intelligence, July 16-20, Boston, Massachusetts, pp: 775-780.

Ng, H.T., 1997. Exemplar-based word sense disambiguation: Some recent improvements.

Pal, A.R. and D. Saha, 2015. Word sense disambiguation: A survey.

Pal, A.R., D. Saha and A. Pal, 2017. A Knowledge based methodology for word sense disambiguation for low resource language. Adv. Computat. Sci. Technol., 10: 267-283.

Purandare, A. and T. Pedersen, 2004. Word sense discrimination by clustering contexts in vector and similarity spaces. Proceedings of the Eighth Conference on Computational Natural Language Learning at HLT-NAACL May 6-7, ACL Anthology, Boston, Massachusetts, USA, pp: 41-48.

Rivest, R.L., 1987. Learning decision lists. Machine Learn., 2: 229-246.

Sarmah, J. and S.K. Sarma, 2016. Survey on word sense disambiguation: An initiative towards an Indo-Aryan language. Int. J. Eng. Math., 6: 37-52.

Sreenivasan, D., M. Vidya, D. Sreenivasan and M. Vidya, 2016. A walk through the approaches of word sense disambiguation. Int. J. Innov. Res. Sci. Technol., 2: 218-223.

Wanton, M.T. and L.R. Berlanga, 2012. A clustering-based approach for unsupervised word sense disambiguation. Procesam. Leng. Nat., 49: 49-56.

Zhou, X. and H. Han, 2005. Survey of word sense disambiguation approaches. Proceedings of the FLAIRS Conference, May 1, pp: 307-313.