I.J. Engineering and Manufacturing, **2016, 3, 37-52**
Published Online May 2016 in MECS (http://www.mecs-press.net)
DOI: 10.5815/ijem.2016.03.04

Available online at http://www.mecs-press.net/ijem

Survey on Word Sense Disambiguation: An Initiative towards an Indo-Aryan Language

Jumi Sarmah [a], Dr. Shikhar Kumar Sarma [b]

[a] *Department of Information Technology, Gauhati University, Guwahati, Assam, 781014, India*
[b] *Department of Information Technology, Gauhati University, Guwahati, Assam, 781014, India*

Abstract

Resolution of lexical ambiguity, commonly known as Word Sense Disambiguation (WSD) task is to distinguish the correct sense among the set of senses for an ambiguous term depending on the particular context automatically. It plays the vital role as it acts as an intermediate phase to many Natural Language Processing (NLP) applications like Machine Translation, Information Retrieval, Speech Processing, Hypertext navigation, Parts-of -Speech tagging. Existing literature reveals that there are various approaches for lexical ambiguity resolution-Knowledge based, Corpus based. In recent years, many WSD systems is being developed in Indian languages like Hindi, Malayalam, Manipuri, Nepali, Kannada but no such automated system has yet emerged for the Indo-Aryan language- Assamese. Our future work aims to develop a model for the WSD problem which is fast, optimal and efficient in terms of accuracy and scalability. This paper presents a survey report made in this research topic discussing the WSD problem, various approaches along with their algorithms. Moreover it also list out the various NLP applications which would be efficient when disambiguation system is merged. Evaluation measures used to determine the WSD performance are also discussed here.

Index Terms: Assamese, Lexical Ambiguity, Natural Language Processing, Word Sense Disambiguation.

1. Introduction

Language is the primary means of communication used by human. It shapes a thought, has a structure and carries meaning. There is some kind of representation in our mind of the content of language. Automatic processing of Natural language i.e., NLP is concerned with the development of computational model of aspects of human language processing. Fig 1. Will give us a clear understanding of the Word Sense Disambiguation concept. Contexts are received either in a spoken form or in a text form. Say E.g. sentences are S1: Boy ate the burger, S2: Boy the burger ate, S3: Burger ate the boy. The collection of words combines to form a proper sentence if it follows a syntactic structure which is analyzed by the Syntactic Processing phase. The parse tree formed indicates that S1 and S3 sentences are syntactically correct but S2 sentence is syntactically wrong. But, even though the sentences are syntactically correct say like sentence S3 it should also provide some correct sense which is processed by the Semantic phase. It is easier for a machine to identify semantic correctness of a

sentence when all words are annotated using their appropriate sense. But, there are certain words which have more than one sense. Lexical Semantic ambiguity is the starting point of semantic analysis. It is such a case where the lexicon is associated with two or more different meanings. To overcome this situation a Word Sense Disambiguation model is required. The last level of phase is pragmatic processing which analyzes the meaning of the whole context or discourse.

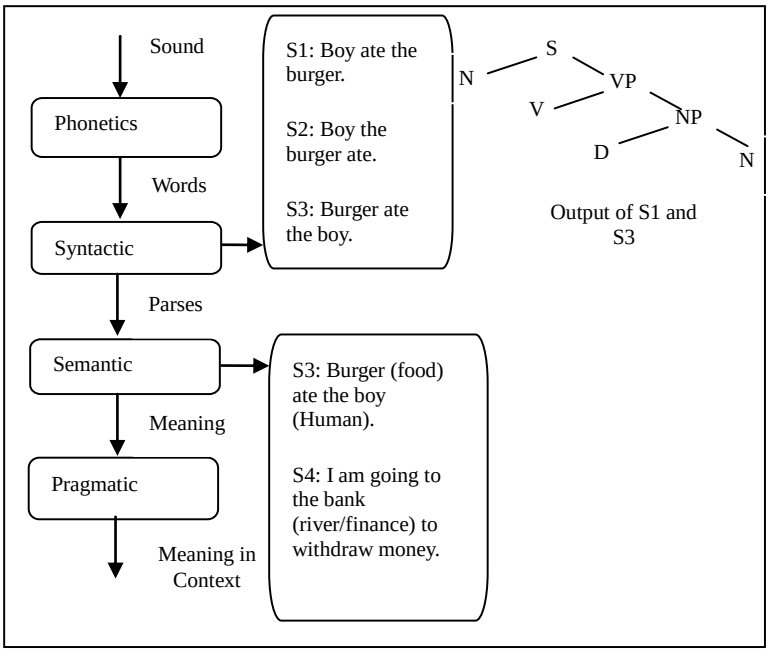

Fig.1. Basic Steps of Natural Language Processing

In any natural language there are certain words which have a number of meanings. The meaning of a word is different depending upon the context, but it is necessary that the proper meaning of a word should be identified based on the relevant context. This perception is known as WSD and is shown in example sentence S4. WSD is an important problem in the field of Natural Language Processing (NLP). Language is the only media through which an individual communicates with another verbally and in a written manner. To a human overcoming the ambiguity of a natural language is not a big deal but to a computer? The term ambiguity concentrates on those words which have different senses. Sense here denotes one of the possible meanings of the word in a text. To computationally determine the correct word sense of an ambiguous word is a key problem to many NLP applications. The task of Word Sense Disambiguation can be said to be a "*categorization*" problem as the possible word sense is selected from a set of predefined categories by Fujii, 1998.

1.1. Lexical Ambiguity

For lexical items two types of ambiguity are traditionally found known as *polysemy and homonymy*. In a piece of context, an individual can come across a polysemous term or a homonymous one. Dash, 2012 draws a clear line of distinction between the two ambiguities and noted that they differ not only in their nature but also in function and implication. In polysemy, a word is associated with more than one meaning which is traditionally called as senses and is distinct but related in some semantic way. In homonymy, a word has meanings which are distinct but not related in a manner. Like the term "bank" signifies *river edge* and *financial*

institution is a common example of homonymy. Whereas the term "mouth" signifies *mouth of the river and mouth of the man* is an example of polysemy. Similarly, in Assamese language the terms "কলা" (Kola) and "ঘৰ" (Ghar) are examples of homonymy and polysemy respectively. Below example sentences for both the terms which will illustrate us the concept of homonymy and polysemy precisely. Example sentences of homonymy:

S1: *কলা* সকলোৰে সাধ্যৰ ভিতৰত নহয়। (Artistic nature is not built upon everyone.)
Concept: the creation of beautiful or significant things
POS: noun

S2: প্রদীপে *কলা*-বোবা লোকৰ বাবে বধিৰ বিদ্যালয় খোলাৰ কথা ভাবিছে। (Pradip is thinking to open a school for the deaf and dumb people.)
Concept: lacking or deprive of the sense of hearing wholly or in part
POS: adjective
Example sentences of polysemy:

S3: এই ঘৰটোত পাঁচটা কোঠা আছে। (This house has five rooms.)
Concept: a dwelling that serves as living quarters for one or more families
POS: noun

S4: তেওঁ নিজৰ উত্তৰ বহীত ঘৰ বনাই আছে। (He is drawing geometrical figures in his answer copy.)
Concept: any artifact having a shape similar to a plane geometric figure with four equal sides and four right angles
POS: noun

The above example sentences of homonymy shows that although the sentences exhibit same spelling or orthographic form, have got unrelated meanings. In one of the sentence the term *কলা (kola)* means to be a deaf person and in other it means the artistic nature. They differ in both meaning and in etymology (tree structure of the word from where it evolved). But, the polysemous example shows that a particular word has its various senses depending on the context but these words basically have the same core sense. In natural language, we also find that single or individual words means different than the collection of words like phrases as for example:

S5: *"This task is a piece of cake for me!"* The phrase *piece of cake* actually means an "easy task" but the individual word means different. It is a difficult task to disambiguate.

1.2. Problem Overview

WSD task is to select the appropriate sense among the set of senses depending on the piece of context. Basically, there are two main variations to disambiguate the word sense of a word- All words WSD and Target word WSD. Supervised approaches are basically used to disambiguate the restricted set or target word WSD as the system can be trained for each of the target or restricted words using the manually sense-annotated data. But to disambiguate all the words in the context, Unsupervised or Knowledge based approaches are feasible. The WSD problem needs to initialize with the following phases:

- Sense repository- At first it to be decided from where (source) the appropriate sense is to be allocated to the target words before the disambiguation process. Many a time sense inventories like Machine readable

dictionaries, WordNet, Thesaurus etc. are used but sometimes possible senses for a target word is assigned temporally through the disambiguation process.

- Representation of the context- The raw context where the target words appear contains some unnecessary information. Based on the algorithm the context should be represented with some features like POS-tagging, Bag-of words, collocation etc.
- WSD Approach identification- To solve the word sense disambiguation task there are methods which can be classified mainly into two types- Machine learning and Dictionary based approaches. Those systems which are trained with some meaningful manual hand-coded data at first are Machine learning approaches and those approaches which uses external lexical resources like WordNet, dictionary, thesaurus etc. are Dictionary based. Three types of techniques for Machine learning approaches are- Supervised techniques, Unsupervised and Semi-supervised.

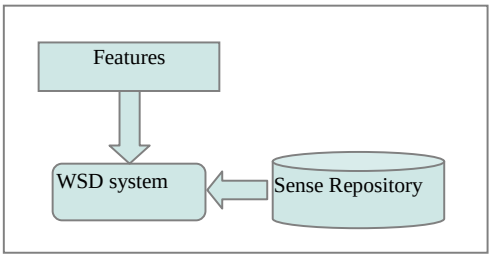

Fig.2. Logical Diagram of WSD

WSD is considered an AI-complete problem, that is, a task whose solution is at least as hard as the most difficult problems in artificial intelligence. Navigli, 2009 reported that Word Sense Disambiguation is an AI-complete problem. In Artificial Intelligence, the most difficult problem is known as the AI-complete or AI-hard which means that it is a task whose solution is at least as hard as the most difficult problem in Artificial Intelligence. AI-complete problems cannot be solved by computers alone but also requires human-computation and also it cannot be solve by a simple specific algorithm. As NLP is a subset of AI and WSD belongs to NLP hence WSD is AI-complete as well. One important reason for why it is termed as hardest problem in AI because it mainly depends on word knowledge. Without this factor it is impossible for both man and computer to process the disambiguation task. Moreover, the need of knowledge varies from domain to domain. Non-availability of knowledge resources considering all domains for all languages is another challenge which adds difficulty to WSD performance.

The morphologically rich Assamese language is spoken mainly by the people of North-East India. It is one of the less computationally aware Indian languages which belong to the Indo-Aryan language family. Nearly 14 million people of North-East region of India speak Assamese language. Unfortunately, this language has fewer number of computational linguistic resources compared to other Indian Language. But, recently some researchers have made a deliberate attempt to study Assamese language from technological perspective. Certain NLP tasks like Named Entity Recognition, POS tagging, Document Classification, Machine Translation, and Spell Checker are among their initiatives.

This paper highlights and discusses the various approaches to WSD problem proposed till date, its use in various NLP applications, various evaluation measures to access the WSD performance etc. The paper is concluded in Section5.

2. Applications of WSD

Resolution of ambiguity commonly lexical is such that it occurs when a single word is associated with multiple senses. Solving this issue will help in improving the quality of various Natural Language Processing

applications which are mentioned below-

- Machine Translation

Statistical MT system is recognized as one of the main beneficiaries of WSD system as a single word in source language is frequently associated with multiple translations in a target language.

S6: আমেৰিকাত দুবছৰ গৱেষণা কৰাৰ পিছত ৰাম ঘৰলৈ ওভতি আহে। (After doing research for two years in America Ram return to his own place/country)

S7: ৰামৰ ঘৰ লক্ষীমপুৰ জিলাৰ কলবাৰীত। (Ram's home is in Kalibari of Lakhimpur district)

On assigning the correct sense of the ambiguous word we will retrieve correct output of machine translation. Many researchers Carpuat and Wu, 2007, Chan and Ng, 2007, Vickrey et al., 2005 have shown that WSD can improve the performance of SMT. WSD system's duty is to properly identify the sense in respect to the context where it is applied. Brown et al., 1991 describes MT Oriented WSD methods.

- Information Retrieval

Information Retrieval (IR) system is affected by lots of noisy words which have multiple senses and so the results given by a IR system when input a query is not efficient. The words in a text should be sense-annotated so that it could retrieve correct results up to certain extent. If every document words are indexed with the sense-tag instead of the single word, the search results will be very much efficient. Hyperlex described by Jean, 2004 is a very good graph-based example where WSD is successfully used for IR. Say, if a user searches a query - "বল"(Bol) in Assamese language than the search engine would retrieve results of "playing object" or "Energy". The query "Java" if indexed with the sense /language than the search engine would retrieve web-pages related to "Programming Language" rather than "Type of Coffee", or "location" is mentioned in the paper by Fukumoto and Suzuki, 1996. Moreover, if a user query is expanded with the proper synset with its proper sense annotation than the documents retrieved would be better and efficient than the original un-expanded queries retrieved documents

- Question Answering

WSD plays a pivotal role in Questing Answering domain. If a question arise like "What is Zidane's role in the play?" The answer would be related to "The Footballer" or "The Character of Final Fantasy". If the question was annotated with proper sense relative to the context than the answer would be appropriate to the user. Thus WSD plays a significant role in Question Answering system.

- Text Categorization

Document/ Texts categorization is the assignment of documents to its respective category automatically. If the words (or keywords) in a document are indexed with proper sense than the documents/texts would be correctly assigned to its respective category. Here WSD plays an important role.

- Speech Processing

Homophonic words like "sealing" or "ceiling" when pronounced the same way but spelled differently requires WSD interpretation.

- Named Entity Classification

The term অপূর্ব (Apurva) in Assamese is ambiguous as it has two senses- significant or may be a name of a person. If the WSD system detects that whether it has a sense of person's name/ significant than NEC may easily categorize it.

- Cross Lingual Information Retrieval (CLIR)

It is the process of retrieving relevant results when query given by the user is provided in one language and

results obtained in another language. As for example, if the Assamese term "*কল*" is indexed with the sense /যিজনে শুনা নাপায় (cannot hear) than English documents related to "deaf" will be retrieved or-else if indexed with sense /কৌশল(skill in arts) than arts related documents would be retrieved. Hence, WSD is an important application to CLIR.

3. WSD Approaches-Machine Learning and Dictionary based

Resolution of WSD means assigning a meaning to an ambiguous word suitable to that particular context. Say an input sentence when contains polysemous word(s), most WSD systems first pre-process the input document to extract a set of features or some clues for the disambiguation process. This preprocessing typically involves features like morphological/syntactic analysis (the parts-of-speech of words), semantic features (Animate/Inanimate). McRoy, 1992 identified syntactic tags, morphology, collocations, and word associations as the most important sources of information for the purpose of WSD. Thereafter, the system interprets polysemous word(s) by selecting a single plausible word sense. Certain systems interpret only one polysemous word in the input text and some systems simultaneously interpret all polysemous words appearing in the input. Wilks and Stevenson, 1997 call this second task type as "word sense tagging". The previous related work says that Machine Learning and Dictionary based approaches are two widely used approaches to solve the lexical ambiguity resolution. These two learning methodologies along with some techniques are briefly discussed below:

3.1. Machine Learning

In this learning methodology, a system is trained or learned with certain features and based on those features the system builds hypothesis and outputs results to unseen input (which are not trained up before). It basically follows three main points: Relies on corpus evidence, Train a model using tagged or untagged corpus, may be a probabilistic, rule-based or statistical model. Prior work reports that most word sense disambiguation tasks uses two types of features- collocation and co-occurrence feature. Collocation feature basically includes the root word (target), two words to left and right of the target word. And the co-occurrence feature consists of data about neighboring words. The neighboring words are considered to a fixed size window. Extracting the features from the input text along with the target word(word to be disambiguated) and then feeding those features to a learner or classifier builds them into a learned model and later the classifier output senses to unseen examples. For example, a machine learning system could be trained on email messages system to learn and distinguish between spam and the non-spam messages. After learning, it can then be used to classify new email messages into spam and non-spam folders. Three types of machine learning based approaches are: supervised, semi-supervised and unsupervised.

3.1.1. Corpus based Supervised Approaches

Those approaches which are based on supervision are supervised approaches. Here, the learner is first trained with some collection of labeled data like in WSD some set of labeled senses and the output of the system is capable of specifying senses to new feature embedded input (target word). As a child is first trained up to read or write and later supervised by a teacher by conducting exam is an example of supervised approach. In the last years we have seen that many classifications of word senses was based on supervised approaches. Basically the main fact is that each supervised algorithm uses certain features associated with a sense for training. The training set is prepared with a set of examples where the examples are manually tagged with sense from some sense inventory. Some of the notable supervised WSD algorithms found in literature are discussed below:

3.1.1.1. Neural Networ

An artificial Neural Network is a network connected by many processing units or neurons. When the excited input is higher than a threshold value, the neuron is activated, and it outputs pulses to its respective neuron (sense). An artificial neuron is normally a multiple input and single output nonlinear unit mentioned by Yu et al., 2011, as shown in Figure1. There are many kinds of neural network like perceptron, feed-forward, recurrent networks. The Feed-Forward network basically consist of three layers- Input Layer, Hidden Layer and Output Layer. Training Phase: It is trained by learning the features of the words from examples, the threshold value of the hidden layer, and output layer neurons. The number of input layer neuron is the number of features of the word. The output layer neuron corresponds to a sense of w. Testing phase: It consists of a test example with its feature vector and the input feature if present in the feature vector are set to 1 and rest to 0. Correspondingly, a neuron in the hidden layer gets activated and fires an output neuron from the output layer which is the winner sense.

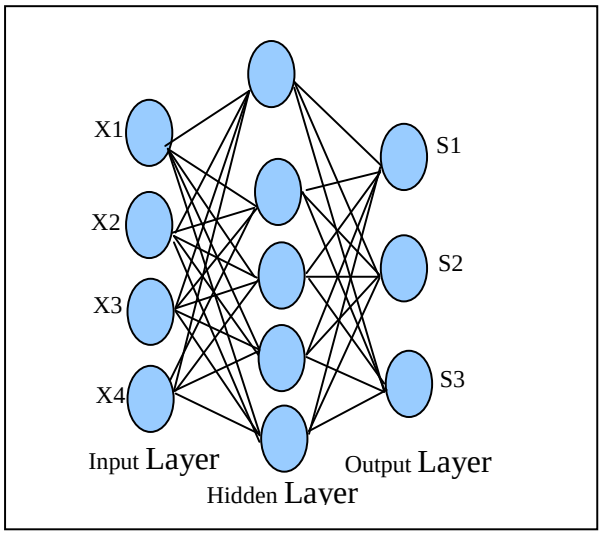

Fig.3. Architecture of Neural Network

Where x_i are inputs, w_i are weights, y_i are outputs (senses).

3.1.1.2. Decision List

A Decision List is a Probabilistic classifier consisting of a set of rules in an ordered list manner first described by Rivest, 1987. It is a set of weighted if-then-else rules and considered as one of the most efficient supervised algorithms. They describe the Decision List Classifier and notes that it basically follows three steps: Feature Extraction: Features are extracted from the set of training examples (sense tagged corpus). The derived feature vector may consist of the following: Parts-Of-Speech (POS) of w, Semantic features, Collocation vector (set of words around w), and Co-occurrence vector (no of times w occurs in a bag of word). Generation of Decision Lists: After the features are obtained from the training corpus, rules of the form (feature, value/score, and class/sense) are created and stored in a table list format. In the table, each tuple corresponds to a rule depicting each sense of the word. This table is sorted in decreasing order of their scores and finally forms the decision list. A separate classifier needs to be trained for each word. The score of a sense (say $sense_a$) is calculated by the below formula for word having two senses:

$$score(sense_a) = \log \frac{\Pr(sense_a \mid Collection_i)}{\Pr(sense_b \mid Collection_i)} \quad (1)$$

The Maximum likelihood also called Probability calculation is done by the below formula:

$$\Pr(sense_a \mid f) = \frac{C(sense_a \mid f)}{C(f)} \quad (2)$$

Where f=collection. $C(sense_a|f)$ means number of occurrences of $sense_a$ with the feature f and C(f) is the number of occurrences of the feature f in the training corpus. But, for words having more than two senses the below formula is implemented as shown in the paper by Sreedhar et al., 2012.

$$score(sense_a) = \log \frac{\Pr(sense_a \mid f)}{\sum_{j!=a} \Pr(sense_j \mid f)} \quad (3)$$

Executing the Decision List Classifier: The algorithm first scan the input with the features in the decision lists. The decision list is tested from the beginning of a tuple and if succeeded the sense associated with that tuple is returned. If failed then decision list is checked until it reaches the end and maximum among the score is assigned. The maximum among the score of senses in the entries is derived by the below formula:

$$s = \arg\max_{s1 \in S}(score(s1)) \quad \textit{where, S is the set of senses} \quad (4)$$

3.1.1.3. Decision Trees

The decision trees are prediction based model which was developed by Quinlan in 1986. Each decision tree has a root node, internal node, branches and leaf node. The internal node denotes a feature on which a test is conducted, the branches represents the outcome of the test result and the leaf denotes the sense label say "YES or NO". The feature vectors that can be used are the Syntactic Features (POS), Lexical Features. During tree construction, attribute selection measures such as Information Gain(for ID3 algorithm) or Gain Ratio(for C4.5 algorithm) are used to select the attributes that best partitions the tuples of the training data into distinct classes. Generation of decision tree:

- If all the tuples in the training data belongs to the same sense category than a leaf node with the same class label is created and the creation of the tree terminates.
- Otherwise considering the attribute having high Information Gain for each node, the sub-tree grows in a recursive manner and continues.
- If there are no attributes on which the tuples may be partitioned or no tuples for a given branch that is a partition is empty than majority voting is considered.

3.1.1.4. Na ïve Bayes

This probabilistic classifier used to disambiguate the ambiguous words by considering the words in the context window say c is described by Gosal, 2015. It relies on the fact that choosing the best sense output for the input vector means choosing the most probable sense for the ambiguous word. Calculation of the conditional probability of each sense of an ambiguous word along-with the words in the context window gives

the score to a sense say s_k. The sense which maximizes the below given formula (5) is considered to be the appropriate sense relevant to that context.

$$score(s_k) = score(s_k) \times log(\Pr(w_i \mid s_k)) \tag{5}$$

Where $score(s_k) = \Pr(s_k)$ and $\log(\Pr(w_i|s_k))$ is calculated by the below equation

$$\Pr(s_k) = \frac{C(s_k)}{C(W)} \tag{6}$$

$$\log(\Pr(w_i \mid s_k) = \frac{C(w_i, s_k)}{C(w_i)} \tag{7}$$

For all senses s_k of a ambiguous word 'm' and for all words w_i in the context window c for a particular sense. Finally the maximum score is determined by the equation:

$$s = \arg\max_{sk \in S}(score(sk)) \tag{8}$$

3.1.1.5. Memory Based Learning

This memory based learning developed by Ng, 1997 is basically based on learning from examples. The model store the examples constructed from the ambiguous words sentences along with its features. It is a memory based learning as when new examples are added then model is not trained again (with the new data) but they are simply added to the existing model. The most common method KNN (K-nearest neighbor) is used in this approach. Survey reports that it basically follows the following steps: Feature extraction: The features are extracted from the training corpus and the features may consist of POS of *w* as well as POS of neighboring words, collocations, Co-occurrence vector, any other Morphological features. Learning: Whenever a new example is added the model is not learned again with the previous examples but simply added. WSD based on KNN: This classifier (KNN) classifies the previous stores examples to clusters based on some similarity measure. Later a new test example is classified based on the previous k clusters. The similarity is measured by the hamming distance as mentioned in the below formula:

$$\Delta(x, x_i) = \sum_{i=1} m\partial(x, x_i) \tag{9}$$

Here, x is the new test example and x_i are previously stored examples. The distance is mathematically calculated with the m features of x_i (senses). The set of K closest clusters belong to a set say $Closest_k$. The new test example belongs to that cluster of $Closest_k$ which has the highest number of neighbors of x (test example) in it. In this way the Exemplar algorithm classifies sense to an ambiguous word but this algorithm will not work for unknown words which do not appear in the corpus.

3.1.1.6. Support Vector Machine

SVM is a binary classifier discussed by Boser et al., 1992 which is based on learning a hyper plane. The hyper plane is learnt from the training data (sense-tagged corpus). It separates the positive and negative examples. Basically hyper plane is located in the hyperspace which maximizes the margin between positive and

negative examples (support vectors). It follows the common training and testing phase by the bellow ways. Training Phase: Using a sense-tagged corpus, for every sense of the word a SVM is trained using the following features: POS of *w* as well as POS of neighboring words, Local collocations, Co-occurrence vector. Features based on syntactic relations (e.g. headword, POS of headword, voice of head word etc.) After feature extraction phase the SVM is trained with those features and the SVM treats the words by classifying them to two categories. But, WSD is a multi-class problem and there can be many distinct senses for a word. For the task of WSD the main problem is broken down into many binary class problems. Testing Phase: Given a test sentence, a test example is constructed using the above features and fed as input to each binary classifier. The correct sense is selected based on the label returned by each classifier. The label returned by the classifier is the maximum confidence score calculated as: $f(x) = w.x + b$. An instance is labeled as positive if $f(x)>=0$ otherwise negative. A geometric intuition will help us to explain better

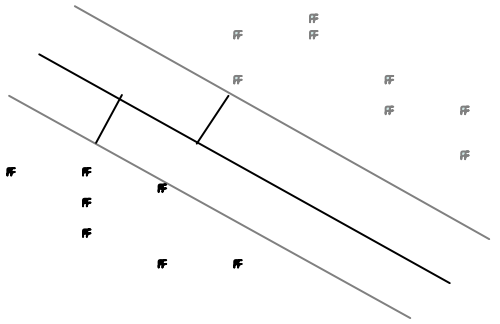

Fig.4. A Geometric View of SVM

Where, w is the vector perpendicular to the hyperplane and b is the bias which is the offset of the hyperplane from the origin. SVM has achieved better results than its baseline accuracy compared to other supervised approach.

3.1.2. Semi Supervised Approaches

This approach also known as minimally supervised algorithm uses small amount of tagged or labeled sense-annotated data and a large amount of untagged data. Basically Semi-supervised approach falls between unsupervised learning (without any labeled training data) and supervised learning (with completely labeled training data). As for example the process of taking food by a human being is a semi-supervised approach. Feeling hungry and knowing that taking food will stop our hungriness is an unsupervised approach and the way of eating food (with hands) is a supervised approach. A semi supervised approach is elaborated below:

3.1.2.1. Bootstrapping

The term “Bootstrapping” means to improve one’s position with own effort and so is the semi-supervised approach. It is based on using Decision Lists devised by Yarowsky, 1992. Two assumptions are mainly followed in this approach: one sense per collocation and one sense per discourse. Considering these assumptions, the algorithm first identifies a set of seed words which will be a set of disambiguating words. A decision list is prepared with these seed words. Using this decision list as many new words as possible are classified. Next, the entire sample set is classified using the Decision list generated previously. A below example will make clear: Suppose there is a topic named “plants” and then partition the data into two sets- “life” and “manufacturing plant”. Then, we received the sets with the words: life => animal, species, microscopic, Plant => equipment, employee and assembly. The rest are residuals (unclassified). Using these

new seed set the unclassified are partitioned. Those words which are labeled with the high confidence score are treated as the new seed data. The process is repeated (recursively) until the output converges to a steady state.

3.1.3. Unsupervised Approaches

According to Amruta and Pederson, 2004 the term unsupervised is itself ambiguous. The approach which doesn't use any sense tagged corpus is unsupervised. It may include approaches which may use the sense inventory created manually like WordNet, thesaurus other than sense tagged corpus. These are mainly clustering approaches. The neighboring words are clustered to one or more groups depending on some similarity strategy and each cluster corresponds to a sense. This approach requires no supervision. As a child is never taught to smile or laugh, they learn automatically is a perfect example of an unsupervised approach. According to Ted Pedersen unsupervised WSD are basically of two types:- discriminative and translation based. Discriminative approaches are based on monolingual untagged corpus whereas Translation based uses parallel corpus for disambiguation tasks. Manual annotation is required for such approaches.

3.1.3.1. Discriminative based WSD

3.1.3.2. Context Clustering

The set of target words (ambiguous) in a text is selected initially. Each context of the ambiguous word is represented by a small vector (feature vector) of first order comprising of morphological features. It may also include the POS of the surrounding words, co-occurrence features including the three most frequent words in the bag of context window. All the feature vectors are represented by a N×M matrix. An N×N dissimilarity matrix is created from N×M matrix in which (i, j)th value is the number of differing features in ith and jth context. Then these contexts with the similar value are clustered until a specific number of clusters is reached. Later the clusters are labeled with an appropriate a sense. Various clustering methods like Ward's agglomerative clustering, Expectation Maximization can be applied but McQuitty average link clustering algorithm performs best among these approaches.

Word Clustering

Lin, 1998 specifies that two words clusters if they share the same syntactic relationship. More the similarity in their relation more close the words belong to a cluster. Say, the context words are w_1, w_2, w_3 and target word say w, the similarity between w and w_i is determined based on information of their syntactic features. Suppose say an example "The *facility* will employ 100 new employees" and from here the term *facility* is to be disambiguated. Here the sense of the term facility (install, proficiency, readiness, adeptness) and the four topics of the employ (org, plant, company, industry) in the corpus are determined in terms of log-likelihood measure. The output is the install sense for facility with the highest log-like likelihood measure.

Co-occurrence Graphs

Hyperlex, a graph-based unsupervised way proposed by Jean, 2004 is an approach meant for detecting the context with the relevant sense of the target word. According to this algorithm, the vertices of the graph are the words in the context along with the target word (to be disambiguated) and they are joined with an edge if they co-occur in the same paragraph. The edge weights are determined by the below formula:

$$w_{ij} = 1 - \max\{\Pr(w_i, w_j), \Pr(w_j, w_i)\} \quad (10)$$

Where Pr(wi|wj)= Frequency of co-occurrences of words wi and wj / Frequency of occurrence of wj.

After the weights are determined, the node with the highest degree (connections) is considered as the hub. Then the neighbors of this node become the candidates of this hub. The hubs are then linked to the target word by determining the MST (Minimum Spanning Tree) for the resultant graph. Each node in the MST is assigned a score vector and it is determined by the below formula:

$$s_i = \frac{1}{1+d\left(h_{i,v}\right)} \text{ if } v \text{ belongs to component } i \tag{11}$$

$$s_i = 0 \text{ otherwise} \tag{12}$$

$d(h_{i,v})$is the distance between root hub h_i and node *V* in the tree. The score vectors of all words are added and the component with the highest score becomes the winner sense.

3.1.3.3. Translation based WSD

Diab and Resnik, 2002 describes Translation based WSD in their research paper. This approach disambiguates target-words by translation process which uses un-tagged word-aligned parallel corpus in two languages. It is based on translational equivalence and relies on the fact that the different senses of a word in a source language may translate to completely different words in a target language. These approaches have got two attractive properties.

- Automatically derive a sense inventory that makes distinctions that are relevant to the problem of machine translation.
- A sense-tagged corpus can be automatically created and used as training data for traditional methods of supervised learning.

The algorithms works in the below steps: Step 1: Words in the target corpus and their corresponding translations in the source file are identified. Step 2: Target sets are formed by grouping the words in the target language. Step 3: Within each of these target sets {w1, w2, w3}, all the possible sense tags for each word are considered. Considering the syntactic, semantic features of the word in the text, final sense tag is selected determining their score. Step 4: Finally sense tags of words in target language are mapped to the corresponding words in the source language.

3.2. Dictionary Based Approaches

Knowledge based methods use lexical and semantic knowledge such as Machine Readable Dictionaries (MRD), thesaurus. Assamese WordNet developed in the year 2009 by Sarma et al., 2010 is the widely used MRD in Assamese NLP and used for many developing applications like MT by Barman et al., 2014, Document classification by Barman et al., 2013. But this hand-made thesaurus is available for only some language as creating this is expensive and time-consuming. WordNet was developed by Princeton University. Like an ordinary dictionary it contains definitions—glosses—of words; however, its distinctive feature is semantic relationships which form hierarchical structures of words. Basic building block of a WordNet is the synset i.e. synonym set—which represents a single concept. Conceptually for the words to be disambiguated the senses are retrieved from the dictionary in this approach. They may use some grammatical rules for disambiguation. Some of the known knowledge based applications are discussed below:

3.2.1. WSD using Selection Preferences

Selection Preferences described by Mihalcea, 2006 restricts the number of meanings of a target word (word to be disambiguated) occurring in a context. Selection preference approaches is some constraint on semantic type such that a word sense is imposed on the target word where it combines usually through grammatical relationship in sentence. Such that an example sentence considering the ambiguous term "*employs*":

- The facility will employ new employees. ("\to hire")
- The committee employed his proposal. ("\to accept")

To be more precise, the term employ (a) restricts its subject and object nouns to those associated with the semantic features HUMAN/ ORGANIZATION and HUMAN (animate), respectively. On the other hand, employ in (b) restricts its subject and object nouns to those associated with the semantic features HUMAN/ORGANIZATION and IDEA (inanimate), respectively. Consequently, given employees as the object, the sense \to hire is selected as the interpretation of employ in (a), and the sense \to accept is ruled out. One may notice that selection restriction can also disambiguate polysemy of verb complements (the subject and object). For example, facility in (a) has multiple senses, a sample of which are "\installation", "\proficiency" and "\readiness". However, the selection restriction imposed for the subject of employ (\to hire) can correctly select the sense \installation as the interpretation of facility. It should be noted that the polysemy of both facility and employ are theoretically disambiguated simultaneously. However, considerable human effort and large amount of knowledge base is required to describe large-scaled selection restrictions.

3.2.2. Overlap based approach

This knowledge based approach generally requires a Machine Readable Dictionary (MRD). Various features of different senses of words (ambiguous) and context words are determined and then overlap is performed between the features. The maximum overlap is selected as the appropriate sense for the ambiguous term. The common Lesk's algorithm and Walker's algorithm are discussed here:

3.2.2.1. Lesk algorithm

The Lesk algorithm proposed by Lesk, 1986 uses WordNet, a huge lexical data-base. The approach is explained with the below example: "On burning *coal* we get *ash*." Here the term *ash* is to be disambiguated. For these two bags (context and sense) are determined. Sense bag say S contains the definition of the senses of the ambiguous word. Along with the definition it may contain features like Synonyms, Glosses, Example sentences, Hyponym etc. Context bag say C contains collection of the context words of the sentence. After the bags are prepared then overlap or intersection similarity is measured and the maximum common sense among the senses is the most probable sense. Say the term "Ash" has three senses and let this be in sense bag S:

- Trees of the olive family with pinnate leaves, thin furrowed bark and gray branches.
- The *solid* residue left when *combustible* material is thoroughly *burn*ed or oxidized.
- Strong elastic wood of various ash trees; used for furniture and tool handles and sporting goods such as baseball bats. Moreover, the context bag C contains: 1. A piece of glowing carbon or *burn*t wood. 2. Charcoal 3. A black *solid combustible* substance formed by the partial decomposition of vegetable matter without free access to air and under the influence of moisture and often increased pressure and temperature that is widely used as a fuel for *burn*ing. Here the sense b of word "Ash" is the winner as maximum intersection similarity is in this sense only.

3.2.2.2. Walker algorithm

Walker in 1987 described it is a thesaurus Based approach. Mainly follows the below two steps:

- For each sense of the target word find the thesaurus category to which that sense belongs.
- Calculate the score for each sense by using the context words. A context words will add 1 to the score of the sense if the thesaurus category of the word matches that of the sense. E.g. sentence. The money in this bank fetches an interest of 15% per annum. Say the word to be disambiguated is "bank". Words from the context are: money, interest, annum, fetch. The category finance of the thesaurus will be the perfect sense for the term bank. Paper by Kalita and Barman, 2015 implements the Walker Algorithm for Assamese WSD.

3.2.3. Conceptual Density

WSD using Conceptual Density is determined by selecting the proper sense depending on the relatedness of the sense to that particular context. Relatedness is measured in terms of conceptual distance i.e., how close the concept represented by the *word* and the concept represented by its *context words* are. It uses the hierarchical semanticnet WordNet to determine the conceptual distance. The common methodology is smaller the conceptual distance higher will get the conceptual density. If all words in the context are strongly related to a particular concept then that concept will have less conceptual distance and higher density.

4. Evaluation Measures

This section presents the common evaluation measures for accessing a WSD system. Resnik and Yarowsky, 1999 Presents some of the common evaluation methodologies. Basically, the main objective of WSD is to get embedded in applications like IR, MT, TC etc. and to improve their performance. Some of the common evaluation measures are Accuracy, Precision, Recall trade-off and F1 measure. Let us elaborate firstly the notion of Precision and Recall. In case of IR, recall is define as systems that retrieve as many documents salient to a user query as possible, while precision is define as systems that retrieves few irrelevant documents as possible. As can be seen, when all the documents are retrieved, recall is always 100%, sacrificing precision. Again in case of TC, recall defines systems that assign as many correct categories to each document as possible, while precision defines systems that assign few incorrect categories to each document as possible. In case of IR: Recall= correct answers provided/ correct answers to provide. Precision= correct answers provided/ answers provided. In case of TC: Recall= correct categories assigned to documents/ correct answer to provide. Precision= correct categories assigned to documents/ no of categorized documents.

In-order to integrate precision and recall F-measure is used which is defined as:

$$F= 2PR/P+R$$

As one may notice that as a type of categorization task, word sense disambiguation can equally be evaluated as performed for TC tasks. However, more than one category can be assigned to a document and most researchers assign a single sense to each word. Therefore, some of them seem to prefer accuracy as the evaluation criterion which is mentioned below:

Accuracy=no of correct decisions made/ total no of decisions made.

5. Conclusions

This paper summarizes the overall concept of an absolute problem in NLP which is WSD. It discusses the various lexical ambiguities and applications of WSD in IR, MT etc. Various supervised, unsupervised, semi-supervised, knowledge based WSD algorithms are surveyed and discussed here elaborately. Survey reports that evaluation measures basically Precision, Recall and F-measure are generally used as metrics in accessing WSD system. But, we need to come up with such an algorithm which is efficient, scalable and portable to other languages. We further need to analyze and experiment existing supervised, un-supervised and knowledge-based approaches to know about the strengths and shortcoming of the existing system. Improvement of earlier proposed systems and exploring other approaches should be also made. Assamese is the official language of North-East and is in the developing phase of Natural Language Processing. Surveying the approaches would enable the developers, researchers to build a WSD model for Assamese Language which will improve the accuracy of many developing Assamese NLP resources.

References

[1] Fujii Atsushi, Corpus-Based Word Sense Disambiguation PhD Thesis, Department of computer science, Tokto Institute of Technology, March 1998.
[2] Shekhar Dash Niladri, Polysemy and Homonymy: A Conceptual Labyrinth; Proceedings of the 3rd IndoWordNet Workshop; pp. 01-07, IIT Kharagpur, India, 2012.
[3] Navigli R, Word Sense Disambiguation: A survey; ACM Computing Surveys, Vol. 41, No. 2, Article10, 2009.
[4] Carpuat M, Wu D, Improving statistical machine translation using word sense disambiguation; Proc. of EMNLP-CoNLL, 2007.
[5] S Chan Y, T Ng H, Domain adaptation with active learning for word sense disambiguation, Proc. of 45th Annual Meeting of the Association of Computational Linguistics, pp. 49–56, Prague, 2007.
[6] Vickrey D, Biewald L, Teyssier M, Koller D, Word-sense disambiguation for machine translation; Proc. of EMNLP, pp. 771–778, 2005.
[7] Brown PF, Stephen A, Pietra D, JD Pietra V, Word-sense disambiguation using statistical methods, Proc. of 29th Annual Meeting of the Association for Computational Linguistics, pp.264-270. 1991.
[8] Jean V, Hyperlex: Lexical cartography for information retrieval, Computer Speech & Language, Vol. 18 No.3, pp. 223-252, 2004.
[9] Fukumoto F, Suzuki Y, An automatic clustering of articles using dictionary definitions; Proc. of 16th International Conference on Computational Linguistics, pp. 406-411, 1996.
[10] McRoy SW, Using multiple knowledge sources for word sense discrimination, Computational Linguistics, Vol. 18, pp. 1-30. 1992.
[11] Wilks Y, Stevenson M, Sense tagging: Semantic tagging with a lexicon; Proc. of ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?, pp. 47-51, 1997.
[12] Yu J, Huang L, Fu J, Mei D, A comparative study of Word Sense Disambiguation Of EnglishM odal Verb by BP Neural Network and Support Vector Machine, International Journal of Innovative Computing, Information and Control ICIC International, Volume 7, No. 5(A), 2011.
[13] L Rivest Ronald, Learning Decision Lists; Machine Learning, pp. 229-246, 1987.
[14] Sreedhar J, Viswanadha Raju S, Vinaya Babu A, Shaik A and Pavan Kumar P, Word Sense Disambiguation: An Empirical Survey, International Journal of Soft Computing and Engineering, Volume-2, Issue-2, pp. 494-503, May 2012.
[15] Pal Singh Gosal Gurinder, A Naive Bayes Approach for Word Sense Disambiguation; Published in IJARCSSE, Volume 5 Issue 7, 2015.
[16] Ng HT, Exemplar-based word sense disambiguation: Some recent improvements; Proc. of 2nd conference

on Empirical methods in natural language processing, pp. 208-213, 1997.
[17] Boser BE, Guyon IM and Vapnik VN, "A training algorithm for optimal margin classifiers", Proc. of 5th Annual ACM Workshop on Computational Learning Theory, pp. 144–152, 1992.
[18] Yarowsky D, Word-sense disambiguation using statistical models of roget's categories trained on large corpora, Proc. of 14th conference on Computational linguistics, pp. 454–460, Morristown, USA. 1992.
[19] Amruta P and Pedersen T, Word sense discrimination by clustering contexts in vector and similarity spaces, Proceedings of the Conference on Computational Natural Language Learning. Vol. 72. 2004.
[20] Lin D, Automatic retrieval and clustering of similar words, Proc. of 17th International conference on Computational Linguistics, pp. 768-774, Morristown, USA, 1998.
[21] Diab M and Resnik P, An unsupervised method for word sense tagging using parallel corpora, Proc. of 40th Annual Meeting on Association for Computational Linguistics, ACL '02, pp. 255–262, Morristown, USA, 2002.
[22] Sarma SK, Medhi R, Gogoi M and Saikia U, Foundation and structure of developing Assamese WordNet, Proc. of 5th international conference of the Global WordNet Association (GWC 2010), 2010.
[23] Barman AK, Sarmah Jumi and Sarma SK, Assamese WordNet based Quality Enhancement of Bilingual Machine Translation System, Proceedings of the Seventh Global Wordnet Conference, pp. 256-261, Estonia, 2014.
[24] Barman AK, Sarmah Jumi and Sarma SK, Automatic Assamese Text categorization using WordNet Proc. of International Conference on Advances in Computing, Communications and Informatics(ICACCI),pp. 85-89, Mysore, India, 2013.
[25] Mihalcea R, Knowledge based methods for WSD, Text, Speech and Language Technology, Vol. 33, pp. 107-132, Springer, Netherland, 2006.
[26] Lesk Michael, Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone, Proc. of 5th Annual International Conference on Systems Documentation, SIGDOC '86, pp 24–26, New York, 1986.
[27] Kalita Purabi, Barman AK, "Implementation of Walker Algorithm in Word Sense Disambiguation for Assamese Language"In Proceedings of IEEE IACC 2015, 14th -15th September 2015.
[28] Resnik P and Yarowsky D, "Distinguishing systems and distinguishing senses: new evaluation methods for word sense disambiguation". Nat. Lang. Eng., 5, pp. 113–133, 1999.

Authors' Profiles

Jumi Sarmah is a Research Scholar (PhD) in the Dept of Information Technology, Gauhati University, Assam, India. Her research of interest includes Natural Language Processing, Machine Learning.

Dr. Sikhar Kr Sarma is the Professor & HOD in the Dept of IT, Gauhati University, Assam, India.

How to cite this paper: Jumi Sarmah, Shikhar Kumar Sarma,"Survey on Word Sense Disambiguation: An Initiative towards an Indo-Aryan Language", International Journal of Engineering and Manufacturing(IJEM), Vol.6, No.3, pp.37-52, 2016.DOI: 10.5815/ijem.2016.03.04