==========
Language in India www.languageinindia.com**ISSN 1930-2940 Vol. 18:2 February 2018**
India's Higher Education Authority UGC Approved List of Journals Serial Number 49042
==========

A Survey on Word Sense Disambiguation

B. H. Manjunatha Kumar, B.E., M.Tech.
Sri Siddhartha Institute of Technology

==========

Abstract

In natural language processing (NLP), word sense disambiguation (WSD) is an automatic process carried out by a machine to sense the appropriate meaning of a word in a particular context or in a discourse. Natural language is ambiguous, so that many words may be interpreted in multiple methods depending on the context wherein they occur. The computational identification of which means for words in context is known as word sense disambiguation (WSD). In this paper, we will discuss the ambiguity of the words in the languages and the essential measures to deal with the ambiguous words.

Keywords: ambiguity, word discrimination, supervised ambiguity, unsupervised ambiguity.

1. Introduction

Word sense disambiguation is a necessary leap for many language translation applications for major human languages spoken across the world. It provides inspiration and motivation to many international research organizations to work on word sense disambiguation. As till date, no accurate system has been developed for WSD [1] which could achieve the current state-of-the-art accuracy rate of 60-70%. Though, WSD is a most significant problem in natural language processing most of the approaches and techniques presented till date to tackle WSD are not stand-alone techniques. To carry out the word sense disambiguation we are in a need of a very large amount of word knowledge. Ambiguity in language means words that carry multiple meanings in different settings and also changes the way the same sentence is interpreted.

For example, English noun bark can imply sound made by a dog, a sort of sailing vessel with 3 or greater masts or the outer- maximum layers of stems and roots of woody plants; similarly the Hindi phrase हल (hal) can mean जमीन जोतने का एक उपकरण (a tool used to plough discipline) or

==========

समाधान/नबटारा (answer). Human beings are pretty apt in determining the perfect sense of the phrase, but for machines that is a very hard challenge. The challenge of computational identification of the proper sense of a word in a given context is known as Word Sense Disambiguation (WSD).

Hindi Wordnet [4] is an important lexical resource which was developed at IIT Bombay, India. The Indian languages had evolved from four different families- Indo-Aryan, Austro-Asiatic, Dravidian, and Tibeto-Burman. The Kannada language belongs to the Dravidian [6] family. Kannada is a very free order word language which exhibits a rich morphological system, which includes Inflection, Conflation, Compounding,and Derivation. Unlike English, prepositions concept are not present in the Kannada language. By successively adding inflections before the noun of the phrases English prepositions are translated in Kannada. The choice of the appropriate word depends on the information of the Wordnet synset of the head noun [7]. A Prepositional Phrase (PP) is a word group that contains an object of a preposition, a preposition,and a modifier.

For example,

Father brought a shirt with tiny stars.

ಅಪ್ಪ ಚಿಕ್ಕ ನಕ್ಷತ್ರಗಳನ್ನು ಹೊಂದಿರುವ ಅಂಗಿ ತಂದಿದ್ದಾರೆ.

Yuvan writes a letter with pen.

ಯುವನ್ ಪೆನ್ನಿನಿಂದ ಕಾಗದ ಬರೆದನು.

In this example the preposition ‘with’ has been translated in the first case as ‘hondiruva’ but in the second case as ‘inda’. This kind of ambiguity will have adverse effects by drastically changing the semantics of the sentence.

Providing innovative technology to clear up this hassle can be one of the predominant demanding situations in language engineering to get right of entry to advanced knowledge technology systems.

1.1 Some basic types of Ambiguity

i. **Lexical Ambiguity**: This ambiguity occurs when a word carrying several meanings appears in the context. e.g.: bank.

Bank can be sensed as a river bank or as a place where people keeps their money and other valuables. In the Hindilanguage, the word ‘aam’ has two meanings- one is ‘common’, e.g.:

"aamaadmiyon ka haq". The other meaning of 'aam' refers to a fruit 'mango'. e.g.: "garmiyon me aam logon kopasandaatahai".

ii. **Syntactic Ambiguity**: The syntactic ambiguity comes into play when a sentence or a collection of words can be given one or more grammatical structure where each of them will be having separate meaning.

iii. **Semantic Ambiguity:** Semantic ambiguity is when there is more than one way to read a sentence although containing no structural or lexical ambiguity it is called semantic ambiguity. e.g.:
ಎಲ್ಲಭಾಷಾಶಾಸ್ತ್ರಗಳುಸಿದ್ಧಾಂತವನ್ನುಆದ್ಯತೆನೀಡುತ್ತವೆ(ellabhasasastragalusiddhantavannuadyatenidutt ave)
In English, it means "all linguistics prefer a theory".

iv. **Pragmatic Ambiguity:** Pragmatic ambiguity occurs when a sentence has more than one meaning in the way it is uttered.
For example, every student thinks she is a genius.

1.2 Problem Description

Word sense disambiguation is a process of sensing the correct meaning of a word in a particular setting in a context. Now, if we consider a given text as T and we view the text by removing punctuations then, the text T would look like (w1,w2,w3,w4……wn) where, w1,w2,w3… etc. are the words in the text T. The mapping process assigns each word with multiple senses, Wi ϵ T, though the appropriate sense is determined, that is, to make out a mapping *A* from words to senses, such that *A*(*i*) ⊆*Senses D*(*wi*), where *Senses D*(*w*i) is the set of senses encoded in a dictionary D for word wi and A(i) is that subset of the senses of wi which are appropriate in the context T. The mapping A can assign more than one sense to each word wi∈ T, although typically only the most appropriate sense is selected, that is, | A(i) |= 1.

2. APPROACHES TO WSD

Basically, the WSD approaches can be categorized as:

i) Supervised WSD: these are basically machine learning techniques that are used by classifiers to learn from training sets.

ii) Unsupervised WSD: the unsupervised WSD is based on unlabeled-corpora and it does not provide sense-tagged corpus to extract a sense of the word.

These approaches can also be stated on the basis of another criterion:

a) Knowledge-based or Dictionary-based: in here, the WSD depends upon the utilization of external resources like word dictionaries, ontology, thesauri etc.

b) Corpus-based This system does not utilize any of the external resources like others.

2.1 Supervised WSD

The supervised WSD follows a machine –learning technique where it manually introduces a classifier from sense-annotated data sets. Basically, a classifier is concerned to process a single word and extract the proper sense of each instance of the word.

2.1.1 Decision List

It is an ordered rule set used to categorize the test instances. A decision list is viewed as a set of if-then-else rules. The training sets induce a feature set which results in the creation of rules of a kind (feature-value, sense, score). The ordering of rules on the basis of their decreasing score determines the decision list [2].

Consider a word sequence W is given and it is represented as a feature-vector, now, the decision list is checked and the maximum scoring feature vector which matches the input decides the sense of the word. It is defined as:

$$S = \underset{S_i \in Senses\ D(w)}{\text{argmax}} \; score\ (S_i)$$

According to Yarowsky [3], S_i is determined as the maximum score among all scored feature. Here, score of feature is calculated as-

$$score(S_i) = \max_f \log\left(\frac{P(S_i|f)}{\sum_{j \neq i} P(S_j|f)}\right)$$

2.1.2. Decision Trees

The decision tree is a predictive WSD model which is used to represent the rules of classification with a tree-like structure that can recursively partition the training set. Each node of the tree represents a feature value test and each branch depicts the outcome of the test. When the terminal node or a leaf is reached a prediction is made based on the test outcome.

2.1.3. Naïve Bayes

A Naïve Bayes is a classifier that is very simple yet probabilistic in nature and is based upon the Bayes' theorem applications. It lies upon determining the conditional probability of each sense for a given word and its feature in the context. A sense that maximizes the formula is considered as an appropriate sense.

$$\begin{aligned}\widehat{S} &= \underset{S_i \in Senses\ D(w)}{\operatorname{argmax}} P(S_i|f_{1,\ldots\ldots}f_m) \\ &= \underset{S_i \in Senses\ D(w)}{\operatorname{argmax}} \frac{P(f_1,\ldots..f_m|S_i)\ P(S_i)}{P(f_1,\ldots..f_m)} \\ &= \underset{S_i \in Senses\ D(w)}{\operatorname{argmax}} P(S_i)\prod_{j=1}^{m} P(f_j|S_i)\end{aligned}$$

Where m represents a number of features of the given word.

2.1.4. Neural Networks

Neural networks depict words as nodes. The words activate to those concepts which are semantically connected and vice versa. The activation of nodes activates nodes connected by the exhibitory links and deactivates the nodes connected by inhibitory links.

2.1.5. Exemplar-Based

Instance-based or exemplar-based (or memory based) is supervised learning algorithm, where classification models are built from examples. This classification model contains the examples as points in the memory and new examples are progressively added to the model.

2.1.6. Support Vector Machines (SVM)

The method of SVM is based on the idea that separates the positive examples from the negative examples by learning a linear hyperplane from training sets. The SVM tends to maximize the gap between the examples and minimize the empirical classification error at the same time.

2.2. Unsupervised WSD

Unsupervised WSD is based on the fact that the same meaning of a word will be having similar words with it in the neighborhood. The unsupervised approach does not make use of any training text or resources like dictionaries, thesauri etc. The major disadvantages of unsupervised approach are they do not rely on any shared resources like dictionaries for word senses. There are basically three methods of this approach- context clustering, word clustering and occurrence graphs.

2.2.1. Context Clustering

In context clustering method each of the senses of the target word is depicted as context vectors. The idea behind the approach of context clustering is word space; i.e. space is considered as a vector where words are represented as dimensions. A word w is a vector whose j^{th} component counts the occurrence of w_j within a given context. Finally, a clustering algorithm is used to perform sense discrimination by grouping the contexts of the targeted words. This is called context-group discrimination. Another approach to sense the appropriate meaning of a word is agglomerative clustering. In agglomerative clustering, each instance constitutes a singleton cluster, then next cluster merges with the most similar pair and successively progress until the end or a stop.

2.2.2. Word Clustering

The method of word clustering focuses on grouping semantically similar words in a context to provide an appropriate sense to it. One of the approaches to this method is to identify one of the words (w_1,w_2,……w_k) which is similar to the target word w0. The similarity in w_0 and w_i is identified on the basis of information provided in the context of syntactic dependencies that occur in the corpus. More the dependencies between the words more are the content information.

2.2.3. Occurrence Graphs

It is based upon the notion of co-occurrence graph G= (V, E), where the vertices of the graph V correspond to the words in the context and the edges E connects the syntactically connected vertices in the same paragraph.

One of the approaches to the occurrence graph is called HyperLex. It is an ad-hoc approach proposed by V'eronis. In this approach, at first, a co-occurrence graph is being constructed where words in the corpus are represented by nodes of the graph and edges to the graph are added when the words occur in the same paragraph.

Mathematically, an edge $\{I,j\}$ is given and the weight of the edge w_{ij} is written as:

$$w_{ij} = 1 - max\{P(w_i|w_j), P(w_j|w_i)\}$$

Where $P\left(w_i|w_j\right) = \frac{f_{req_{ij}}}{f_{req_j}}$

$f_{req_{ij}}$ is frequency of occurrence of words and frequency of w_j within the given text is represented as f_{req_j} in the equation.

Now, an algorithm is implied to the graph occurrence and the nodes with the highest degree are selected as hubs. All the selected hubs form a set of hubs that represent the proper sense of the word.

An alternate graph-based word sense disambiguation technique is the PageRank algorithm. PageRank algorithm was developed to rank the web pages and now it is the major part of Google's search engine. In here the PageRank degree is given by the formula:

$$P(v_i) = (1-d) + d \sum_{v_j \to v_i} \frac{w_{ji}}{\sum_{v_j \to v_k} w_{jk}} P(v_j)$$

Where, $v_j \rightarrow v_i$ denotes that there is a vertex from V_j to V_i and w_{ji} is its weight, and d usually set to 0.85 is a damping factor which models the probability of a link to V_i or jumps to V_i randomly. In PageRank algorithm, the vertices of the graph are sorted according to their PageRank and the best-valued vertices are selected as hubs for the targeted words.

3. MERITS AND DE-MERITS

3.1 Disadvantages

Supervised approaches:

i. The exhaustive knowledge base is required.
ii. Generally, the dictionaries that are used are very small in size.
iii. Dictionaries do not always contain different senses of a word.
iv. The pronouns are not present in a corpus that can be matched for a clue.

Naïve Bayes:

i. Suffers from data sparseness.
ii. As the scores are depicted by some probabilities, hence the score might get degraded if some weak feature is present.
iii. A large number of parameters are to be trained to get the proper sense.

Decision List:

i. A separate classifier is needed for every single instance of a word.

SVM:

i. A word sense-specific classifier which uses a separate handler for each instance.

Exemplar-Based:

i. It is a word-specific classifier.
ii. It does not work for any word that is not mentioned in the corpus.

Unsupervised Approaches:

i. Hyperlex (The hyperlex is a word specific classifier) algorithm would fail to recognize the finer senses of a target word.
ii. Always need a parallel corpus which is very difficult to get.

iii. Exceptionally large quantities of parameters are needed to be trained.

3.2 Advantages

Supervised Approaches:

i. Uses corpus instead of dictionary defined word senses.
ii. Can easily grab the clues of proper nouns as they do readily occur in the corpus.
iii. The decision list uses the most predictive features that overcome the demerits of Naïve Bayes approach.
iv. Exemplar-based approaches use a diverse set of features that includes noun-subject-verb pairs and morphological pairs.
v. SVM provides the highest accuracy of output from the baseline.
vi. SVM also uses a diverse feature set.

Unsupervised Approaches:

i. It combines the merits of supervised and knowledge-based approaches.
ii. As like supervised approach, it gathers information from the corpus.
iii. Unsupervised approaches do not need tagged corpus.
iv. Lin's algorithm has a broad coverage.
v. Lin's algorithm can work for words not mentioned in the corpus.
vi. Using a parallel corpus we can even distinguish between the finer senses of a target word.

4. CONCLUSION AND FUTURE WORK

In this paper, we have discussed many of the ambiguous approaches and how to deal with it. We can conclude from our study that,

a) The supervised WSD approaches have always yielded better results as compared to the unsupervised WSD approaches.
b) The neural networks have proven to be a better option to disambiguate the sentences but when it comes to a large quantity of data, tuning its parameters and training the data sets are the major disadvantages in the technique.
c) To use an SVM for WSD it is needed to be adapted to the proper senses of a target word or multiclass classification.
d) Relying on most predictive information increases the accuracy of the algorithm.

e) Word specific classifiers are good to work with in terms of accuracy but cannot be reused or recycled.
f) Completely depending on the dictionary senses are the major reasons of low accuracy in a knowledge-based approach.
g) HyperLex and Lin's [5] algorithm are proven better in case of disambiguating the Indian languages.

The above methods that are stated and discussed are to handle the problem of ambiguity in human languages. One method or approach is not alone enough to deal with the ambiguity properly and accurately. Hence, combining them together can fetch us a better result in the word sense disambiguation process in our near future.

==========

References

[1] Benjamin Snyder and Martha Palmer. "The English all-words task". In Proceedings of SENSEVAL-3, pages 41–43, 2004.

[2] Ronald L. Rivest "Learning Decision Lists", August,2001. http://prople.csail.mit.edu/rivest/Rivest-DecisionLists.pdf.

[3] David Yarowsky "Decision Lists for Lexical Ambiguity Resolution: Application to accent restoration in Spanish and French" http://acl.ldc.upenn.edu/P/P94/P94-1013.pdf.

[4] D. Chakrabarti D. Narayan P. Pandey P. Bhattacharyya "An Experience in Building the Indo-WordNet-A WordNet for Hindi." GWC- 2002.

[5] D. Lin "An Information-Theoretic Definition of Similarity." Proceedings of International Conference on Machine Learning, Madison, Wisconsin, July 1998.

[6] Kavi Narayana Murthy and G. Bharadwaja Kumar, "Language Identification from Small Text Samples," Journal of Quantitative Linguistics, vol. 13, no 1, pp. 57-80, 2006.

==========

[7] S.N. Sridhar, Modern Kannada Grammar, Manohar Publications & Distributors, 2007.

==========

B H Manjunatha Kumar B.E., M.Tech.,
Assistant Professor
Ph.D. Student
Department of Computer Science and Engineering
Sri Siddhartha Institute of Technology
Tumkur572105
Karnataka
India
bhm.cse@gmail.com