International Journal of Latest Trends in Engineering and Technology
Special Issue SACAIM 2017, pp. 462-467
e-ISSN:2278-621X

ANALYSIS OF WORD SENSE DISAMBIGUATION ALGORITHMS IN DATA MINING

Sowmya[1] & Seema[2]

Abstract- In this paper we display the strategy for Word Sense Disambiguation (WSD) in light of the domain information. Domain implies an arrangement of words in which there is a solid semantic connection among the words. The words in the sentence add to discover the domain of the sentence. We show the unsupervised way to deal with Word Sense Disambiguation utilizing the WordNet areas. The model finds the domain of the objective word and the sense corresponding to this domain is taken as the right sense. WSD is testing undertaking of Natural Language Processing (NLP). Despite the fact that there are bunches of calculations for WSD accessible, still little work is completed for picking ideal calculation for that. There are three methodologies accessible for WSD specifically Knowledge-based approach, supervised approach and unsupervised approach. Likewise one can utilize mix of given methodologies. This paper will break down these three methodologies and diverse strategies identified with each approach.
Keywords- Natural Language Processing, word sense disambiguation, text analysis, WordNet, Unsupervised Approach.

1. INTRODUCTION

In all the real dialects around the world, there exist considerable measures of words which signify distinctive implications in various contexts. Word Sense Disambiguation is a method used to locate the correct meaning of an uncertain word in a specific context and to consequently relegate a sense, chose from an arrangement of pre-characterized word detects. WSD is recognizing which sense of a word (i.e. importance) is really utilized as a part of a sentence, when the word has various implications. The answer for this issue impacts other PC related composition, for example, discourse, enhancing importance of search engine, anaphora determination, coherence, induction and so on.

Word Sense Disambiguation (WSD) is the way toward settling the meaning of a word unambiguously in a given normal dialect context. Given a polysemous word in running content, the assignment of WSD includes looking at relevant data to locate the expected sense from an arrangement of predetermined applicants. WSD is task of arrangement in which the senses are the classes, the context gives the confirmation and every event of the word are relegated to at least one of its conceivable classes in light of proof.

The distinguishing proof of the particular implying that a word expect in the context is just clearly basic. The purpose for this is human cerebrum is fit for deciding the right importance of the word in the given context precipitously. The computational distinguishing proof of importance for words in context is called word sense disambiguation (WSD).

The paper is composed as takes after: In Sections II we delineate the current methodologies in WSD. In area III we present the correlation of these depicted three methodologies. In section IV we talk about the conclusion and research audit of WSD approaches.

2. EXISTING APPROACHES

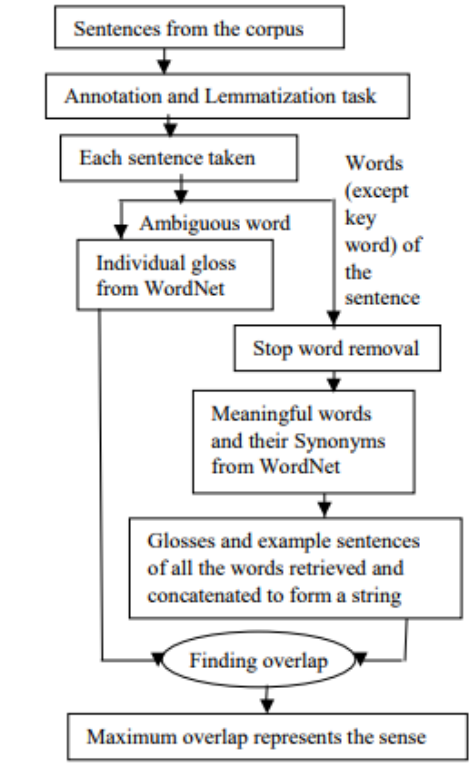
Word Sense Disambiguation Approaches are characterized into following three primary classifications-

a) Knowledge based approach,
b) Supervised approach and
c) Unsupervised approach.

2.1 Knowledge based approach –

This approach is based on different knowledge sources like machine readable dictionaries or sense inventories, thesauri etc. WordNet is the popularly used machine readable dictionaries in this research field. Generally there are four main types of knowledge-based methods are used.

[1] Department of Master of Computer Application, Aloysius Institute of Management and Information Technology, Mangalore, Karnataka, India
[2] Department of Master of Computer Application, Aloysius Institute of Management and Information Technology, Mangalore, Karnataka, India



2.1.1 LESK Algorithm:

This is the first and for most machine meaningful lexicon based calculation worked for word sense disambiguation. Lesk algorithm relies upon the overlap of the lexicon meanings of the words in a sentence. In this approach, at first a short phrase is chosen from the sentence. At that point, lexicon definitions (glosses) for the distinctive senses of the uncertain word and the other significant words exist in the expression are gathered from an online Dictionary. In the subsequent stage, every one of the glosses of the catchphrase is contrasted with the glosses of different words. The sense, for which the most extreme number of overlap happens, speaks to the coveted sense of the vague word.

function SIMPLIFIED LESK(word,sentence) **returns** best sense of word
    best-sense <- most frequent sense for word
    max-overlap <- 0
    context <- set of words in sentence
    for each sense **in** senses of word **do**
    signature <- set of words in the gloss and examples of sense
    overlap <- COMPUTEOVERLAP (signature, context)
    if overlap > max-overlap **then**
    max-overlap <- overlap
    best-sense <- sense
end return (best-sense)

2.1.2 Semantic Similarity:

It is said that words that are connected, share basic context and subsequently the proper sense is picked by those implications, found inside littlest semantic separation. This semantic component is fit to give agreement to entire talk. Different likeness measures are utilized to discover how much two words are semantically related. At the point when more than two words are there, this technique likewise turns out to be to a great degree computationally serious.

2.1.3 Selectional Preferences:

Selectional preferences decide data of the feasible relations of word types, and mean common sense utilizing the information source.

For instance, Modelling-dress, Walk-shoes are the two words with semantic relationship. In this strategy improper word senses are discarded and just those senses are chosen which have concordance with common sense rules. The fundamental aim

behind this approach is to check how often this type of word combine happens in the corpus with syntactic connection. Senses of words will be distinguished from this check.

2.1.4 Heuristic Method:

In this approach, the heuristics are assessed from various semantic properties to decide the word sense.

Three types of heuristics utilized as a pattern for deciding WSD system: Most Frequent Sense, One Sense for each Discourse and One Sense for each Collocation.

The Most Frequent Sense works by deciding every presumable sense that a word can have and it is fundamentally right that one sense happens regularly than the others.

One Sense for each Discourse says that a word will keep its importance among every one of its events in a given content. Lastly, One Sense for every Collocation is same as One Sense for each Discourse with the exception of it is expected that words that are closer give solid and steady flags to the sense of a word.

2.2 Supervised approach –

The supervised approach connected to WSD systems make utilization of machine-learning in procedure from physically made sense-explained information. Training set will be utilized for classifier to learn and this preparation set comprises of illustrations identified with target word. Every one of these labels is physically made from word reference. Fundamentally this WSD calculation gives well outcome than different methodologies. Techniques in Supervise WSD are as per the following:

2.2.1 Decision List:

A decision list is fundamentally an arrangement of "if-then-else" rules. Training sets are settled on use in decision list to initiate the arrangement of highlights for a given word. Utilizing those principles a few parameters like element esteem, sense, score are made. In view of the diminishing scores, last request of standards is made, which makes the decision list. At the point when any word is viewed as, first its event is figured and its portrayal as far as highlight vector is utilized to produce the decision list, from where the score is ascertained. The most extreme score for a vector speaks to the sense.

2.2.2 Decision Tree:

A decision tree is utilized to demonstrate classification rules in a tree structure that recursively partitions the training data set. Interior node of a decision tree means a test which will be connected on element esteem and each branch demonstrates a yield of the test. At the point when a leaf node is achieved, the sense of the word is entitled (if conceivable). A case of a decision tree for WSD is depicted in the figure underneath. The noun sense of the equivocal word "bank" is grouped in the sentence, "I will be at the bank of Krishna River in the morning". In the figure, the tree is made and navigated and the determination of sense bank/RIVER is made. Exhaust estimation of leaf node says that there is no determination accessible for that component esteem.

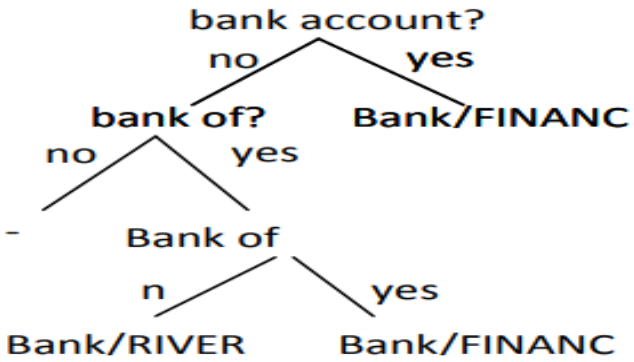


Figure 1. A example of a Decision tree

2.2.3 Naïve Bayes:

Naïve Bayes classifier is a probabilistic classifier in light of Bayes Theorem. This approach group's text documents utilizing following two parameters: the conditional probability of each sense (Si) of a word (w) and the features (fj) in the context. The greatest esteem assessed from the equation beneath speaks to the most proper sense in the context.

$$\hat{S} = \underset{S_i \in Sense_D(w)}{\operatorname{argmax}} P(S_i \mid f_1, \dots, f_m) = \underset{S_i \in Senses_D(w)}{\operatorname{argmax}} \frac{P(f_1, \dots, f_m | S_i) P(S_i)}{P(f_1, \dots, f_m)}$$

$$= \underset{S_i \in Senses_D(w)}{argmax} P(S_i) \prod_{j=1}^{m} P(f_j | S_i)$$

Here, the number of features is spoken to by m. The probability P(Si) is found from the co-occurrence recurrence in training set of sense and P(fj | Si) is resolved from the element within the sight of the sense.

2.2.4 Neural Networks:

In the Neural Network based computational model, artificial neurons are made utilized for information handling utilizing connectionist approach. Contribution of this learning program is the sets of info highlights, and objective is to separate the preparation context into non-covering sets. In the subsequent stage, to deliver a bigger enactment these recently build pairs and connection weights are slowly balanced. Neural systems are utilized to speak to words as nodes and these words will actuate the plans to which they are semantically related. The sources of input are engendered from the input layer to the yield layer through all the middle of the intermediate layers. The inputs can without much of a stretch be proliferated through the system and can be controlled to touch base at a yield. It is exceptionally hard to process a reasonable yield from a system where the associations are spread every which way and shape loops. Feed forward systems are normally a superior decision for those issues which are not time subordinate and foresee a diverse scope of utilizations.

2.2.5 Exemplar-Based or Instance-Based Learning:

This supervised algorithm frames arrangement model from cases. This model stores cases as point in feature space and new illustrations are considered for arrangement. Every one of these cases is continuously added to the model. The k-nearest neighbour algorithm is the one which depends on this strategy. In this methodology, as a matter of first importance a specific number of illustrations are gathered; after that the Hamming separation of a case is controlled by utilizing k– NN algorithm. This separation decides the closeness of the contribution as for the put away cases. The k >1 entitles the dominant part sense of the yield sense among the k-nearest neighbours.

2.2.6 Support Vector Machine:

Support Vector Machine based algorithms make utilize the hypothesis of Structural Risk Minimization. The primary objective of this approach is to isolate positive cases from negative cases with most extreme edge and edge is the separation of hyper plane to the closest of the positive and negative illustrations. The positive and negative illustrations which are nearest to the hyper plane are called as support vector. The SVM based algorithms are utilized for the characterization of couple of cases into two unmistakable classes.

This calculation identifies a hyper plane in the middle of these two classes, so that, the partition edge between these two classes ends up plainly greatest. The arrangement of the test illustration relies upon the side of the hyper plane, where the test case really lies in. The input features can be mapped into a high dimensional space likewise, however all things considered; some portion capacities are utilized to diminish the computational cost of the preparation and the testing technique in high dimensional space. A regularization parameter is put forth use in defence of non-distinguishable preparing illustrations.

The default estimation of this parameter is constantly considered as 1. Henceforth this regularization methodology controls the exchange off between the vast edge and the low training error.

2.3 Unsupervised approach -

Unsupervised WSD approaches don't rely upon outside learning sources or sense inventories, machine readable dictionaries or sense-commented on informational collection. These strategies by and large don't allocate significance to the words rather they separate the word implications in view of data, found in un-explained corpora.

This approach has two types of distributional strategies, initial one is monolingual corpora and other one is interpretation equivalence in light of parallel corpora. These systems are additionally arranged into two types: type based and token-based approach. The type based approach disambiguates by clustering occasions of an objective word. Token-based approach disambiguates by clustering context of an objective word. Principle methodologies of unsupervised are:

2.3.1 Context Clustering:

Context clustering method depends on clustering methods where first context vectors are made and afterward they will be gathered into groups to distinguish the meaning of the word. This technique utilizes vector space as word space. Its measurements are words as it were. Likewise in this strategy, a word which is in a corpus will be shown as vector and how

frequently it happens will be counted inside its unique circumstance. Next, co-event system is made and comparability measures are connected. Presently separation is performed utilizing any grouping method.

2.3.2 Word Clustering:

This technique is identical to context clustering as far as discovering sense yet it bunches those words which are semantically indistinguishable. This approach utilizes Lin's strategy for clustering. It inspects indistinguishable words which are like target word. Also, closeness among those words is resolved from the highlights they are sharing. It can be gotten from the corpus. At the point when words are comparable they share same type of reliance in corpus. At that point, clustering algorithm is connected to separation among senses.

On the off chance that a list of words is taken, first the likeness among them is found. At that point those words are requested by that likeness and a closeness tree is made. At the beginning stage, just a single node is there and for each word accessible in the list, cycle is connected to add the most indistinguishable word to the underlying node in the tree. Finally, pruning is connected to the tree. Thus, it produces sub-trees.

The sub-tree for which the root is the underlying word which we have taken to discover sense, gives the senses of that word. One more technique to this approach is grouping by panel. As specified before, the word grouping is a type of context clustering, this clustering by panel takes after comparable advance, first the comparability network is made, so that, lattice contains match astute comparable data about the words. In the following stage, normal connection clustering is connected to the words. The separation among words is performed utilizing the closeness of centroids.

For every panel, there exists one centroid. Thus, as per the similitude of the centroid, the objective word gives the individual advisory group. In the last advance, includes between the panel and the word are expelled from the first word set, so in next emphasis, distinguishing proof of senses for same word which are less continuous, is permitted.

2.3.3 Spanning tree based approach:

Word Sense Induction is the way toward distinguishing the arrangement of senses of a vague word in a computerized way. These strategies decide the word senses from content with a thought that a given word conveys a particular sense in a specific context when it co-happens with the same neighbouring words. In these methodologies, right off the co-event graph (Gq) is built. After that the accompanying arrangement of steps are executed to decide the correct sense of a vague word in a specific context: a) First, every one of the nodes whose degree is 1 are wiped out from Gq. b) Next, the most extreme traversing tree (MST) TGq of the graph is determined. c) After that, the base weight edge e∈TGq is killed from the graph one by one, until the point when the N associated segments (i.e., word bunches) are framed or there remains no more edges to dispense with.

3. EXPERIMENT AND RESULT

Looking at and assessing different WSD systems is extremely troublesome, as a result of the diverse test sets, sense inventories or machine readable word references, and learning assets required. Prior to the association of particular assessment campaigns most systems were assessed on in-house, frequently little scale, informational indexes. Keeping in mind the end goal to test one's calculation, designers need to invest their energy to clarify all word events. Furthermore, looking at strategies even on a similar corpus isn't qualified if there exist diverse sense inventories.

Examination of these depicted three methodologies is appeared in underneath table:

Criteria	Knowledge-Based Approach	Supervised Approach	Unsupervised Approach
1. Dependency	Depends on sense inventories or machine readable dictionaries	Requires large sense-annotated data	No need of any external resources-works directly from raw unannotated corpora
2. Execution speed	The execution procedure becomes faster and second.	Low execution time	Fast execution time
3. Accuracy	Accuracy of 78 %	Accuracy of 60%-70%	Accuracy of 40%-60%
4.Disadvantage	Suffers from overlap	Not suitable for resource	Performance is always

4. CONCLUSION

From this exploration survey of WSD approaches, Knowledge-based approach is depend in light of sense inventories or machine readable lexicons which are have a tendency to restrict with their own translation of word, if word which isn't accessible in lexicon than it will be hard to locate its significance. Supervised approach gives preferred outcome over others yet it is vigorously relies upon manually made sense-explained text and to manually make these sense-clarified text requires a great deal of time and exertion. Unsupervised algorithm does not rely upon any sense-annotated on information or machine readable word references. It finds the sense on presumption that words that happen every now and again together have identical significance. Future work for unsupervised algorithm can be stretched out by utilizing diverse closeness measures and afterward applying clustering algorithm on that to discover exact sense on word.

5. REFERENCES

[1] D.Bikel and I.Zitouni, Multilingual Natural Language Processing Applications: From Theory to Practice, USA: IBM Press, 2012, pp. 400.

[2] D.Bikel and I.Zitouni, Multilingual Natural Language Processing Applications: From Theory to Practice, USA: IBM Press, 2012, pp. 286.

[3] S. Bandyopadhyay, S. Naskar and A. Ekbal. Emerging Applications of Natural Language processing, Hershey, PA, USA: IGI Global, 2013.

[4] Princeton University. (2012, Nov 10). WordNet: A Lexical database for English [Online]. Available: http://wordnet.princeton.edu/

[5] M. Song and Y. Wu, Handbook of Research on Text and Web Mining Technology, Hershey, PA, USA: IGI Global, 2009, pp. 194.

[6] B. Liu, Sentiment Analysis and Opinion Mining, USA: Morgan and Claypool, 2012.

[7] Nameh, M. S., Fakhrahmad, M., Jahromi, M.Z., (2011) "A New Approach to Word Sense Disambiguation Based on Context Similarity", Proceedings of the World Congress on Engineering, Vol. I.

[8] Navigli, R. (2009) "Word Sense Disambiguation: a Survey", ACM Computing Surveys, Vol. 41, No.2, ACM Press, Pp. 1-69.

[9] Seo, H., Chung, H., Rim, H., Myaeng, S. H., Kim, S., (2004) "Unsupervised word sense disambiguation using WordNet relatives", Computer Speech and Language, Vol. 18, No. 3, Pp. 253-273.

[10] Kolte, S.G., Bhirud, S.G., (2008) "Word Sense Disambiguation Using WordNet Domains", First International Conference on Digital Object Identifier, Pp. 1187-1191.

[11] Martín-Wanton, T., Berlanga-Llavori, R.,(2012)"A clustering-based Approach for Unsupervised Word Sense Disambiguation", Procesamiento del Lenguaje Natural, Revista no 49 septiembre de 2012, pp 49-56.http://rua.ua.es/dspace/bitstream/10045/23919/1/PLN_49_05.pdf date: 14/05/2015

[12] Kumar, R., Khanna, R.,(2011) "Natural Language Engineering: The Study of Word Sense Disambiguation in Punjabi", Research Cell: An International Journal of Engineering Sciences ISSN: 2229-6913 Issue July 2011, Vol. 1, pp. 230-238.

[13] S.G. Kolte and S.G. Bhirud, "Word Sense Disambiguation Using WordNet Domains," First International Conference on Digital Object Identifier, Pp. 1187-1191, 2008.