© 2018 IJSRSET | Volume 4 | Issue 11 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section : Engineering and Technology
DOI : https://doi.org/10.32628/IJSRSET1841120

Handling WSD using Hierarchical Clustering Algorithm with sentences

Mohana Priya K[1], Pooja Ragavi S[2], Krishna Priya G[3]

[1] ME CSE, Jansons Institute of Technology,Coimbatore,Tamil Nadu, India

[2]MBA, GRG School of Management Studies, Coimbatore,Tamil Nadu, India

[3]Assistant Professor, Department of CSE, Jansons Institute of Technology,Coimbatore, Tamil Nadu, India

ABSTRACT

Clustering is the process of grouping objects into subsets that have meaning in the context of a particular problem. It does not rely on predefined classes. It is referred to as an *unsupervised learning method* because no information is provided about the "right answer" for any of the objects. Many clustering algorithms have been proposed and are used based on different applications. Sentence clustering is one of best clustering technique. Hierarchical Clustering Algorithm is applied for multiple levels for accuracy. For tagging purpose POS tagger, porter stemmer is used. WordNet dictionary is utilized for determining the similarity by invoking the Jiang Conrath and Cosine similarity measure. Grouping is performed with respect to the highest similarity measure value with a mean threshold. This paper incorporates many parameters for finding similarity between words. In order to identify the disambiguated words, the sense identification is performed for the adjectives and comparison is performed. semcor and machine learning datasets are employed. On comparing with previous results for WSD, our work has improvised a lot which gives a percentage of 91.2%

Keywords : NLP-natural language processing, POS-part of speech, sentence clustering, K-means.

I. INTRODUCTION

Unsupervised learning approach are appealing for Natural Language Processing(NLP). In many Natural Language Processing (NLP) tasks Clustering algorithms are used. It organizes the documents to improve searching and in information retrieval. Many clustering algorithms have been proposed and are used based on different applications. In general, it is classified into hard clustering and soft (fuzzy) clustering. Sentence clustering is one of the best clustering technique when compared with other type of clustering. The work here deals with the ability to capture such relationships and scope of sentence clustering to solve it. It solves the problem of content overlapping. For each sentence cluster, membership values are assigned which shows the degree to which object represented by that node belong to each respective clusters. The assumption of measuring similarity within sentences based on word co-occurences lead to many of the sentence similarity measures. Similarity is determined by invoking one of the several measures like Jaccard, Manhattan, Euclidean, Jiang Conrath etc,. Several techiques like classification, clustering,c means, k means, hierarchical cluster, sentences cluster, document clusters are available. Classification means categorizing the given data into number of classes.

The main goal of classification is to identify the class to which a new data will fall under. Clustering means partitioning a set of data into meaning full

subclasses,which helps to understand the structure in dataset. K means clustering is used when you have data without defined categories or groups. The result of k means clusters are k clusters, which is used to label the new data. Hierarchical clusters have predetermined order from top to bottom, which builds a hierarchy of clusters. There are two types of hierarchical clusters they are divisive and agglomerative.. Sentence clustering is used to find the purity of clusters. sentence clusters are used for text summarization, topic detection, tracking etc. Document cluster means which consists of clusters that contain candidate words for classifying documents. WSD (word sense disambiguation) is used for identifying which sense of a word is used in a sentence, when the word has multiple meanings.

II. METHODS AND MATERIAL

A. Related Work

LinglingMeng.et al [1] observed that related sentences in a text tend to use the same or similar words.The goal of this paper was to produce a measure of semantic similarity which is a good predictor of "relatedness" between sentences, with the ultimate goal of assessing the coherence of an essay. In this paper they use four different measure for predicting the semantic similarity such as shortest path based measure and Information content based measure, Feature-based Measure, Hybrid Measure. For finding path based measure they use different measures such as shortest path based measure , Wu & Palmer's Measure, Leakcock& Chodorow's Measure, Li's Measure. For finding content based measure they use Resnik's Measure, Lin's Measure, Jiang's measure. Different semantic similarity measures have different characteristic.

Finally they compare all the measures and evaluate the efficiency. This paper reviews various state of art semantic similarity measures in Word Net based on is-a relation. They analyses the principles, features, advantages and disadvantages of different measure. Further more, they present the commonly used IC metric in information content based measures. Finally they discuss how to evaluate the performance of a similarity measure. In fact there are no absolute good performance measures. Different measures will show different performance in different applications.

Patheja.et al [2] discussed about part of speech tagging. They showed what are the classification techniques in part of speech tagging and how it works. There are two types of techniques. First one is supervised POS tagging and second one is unsupervised POS tagging. Supervised pos tagging are classified into two categories such as rule based and stochastic. Unsupervised pos tagging are classified into two categories such as rule based and stochastic. Supervised Classification mainly comprise of two phases i.e. training and prediction. Supervised Technique use a pre-tagged corpora (structured collection of text) which is used for training to learn information about the tagset, word-tag frequencies, rule sets etc. Unsupervised POS tagging models do not require pre-tagged corpora. Operates by assuming as input a POS Lexicon, which consists of a list of possible POS tags for each word. Rule based techniques use contextual and morphological information to assign tags to unknown or ambiguous words. These rules are often known as context frame rules. Stochastic taggers have advantage of reducing the need for manual rule construction and possibly capture useful information. Then they use different models for supervised and unsupervised technique. The models are Decision Tree Model, Condition Random Field Model, Hidden Markov Model, Maximum Entropy Model, Clustering Model, Prototyping Model, Bayesian Model and Neural Networks. supervised technique had shown good performance results in terms of accuracy yet it suffers from the problem of data sparcity. They compare all the models finally they concluded that CRF based model attains good performance results as compared to Maximum Entropy Model.

Liang Wen1 .et al[3] discussed about traditional WSD methods based on supervised learning. They consider only the word, position, part of speech and some other superficial morphological and syntactic features. Based on a knowledge base, they have proposed a novel approach to extract contextual semantic features of ambiguous words. The representation is based on hierarchical network concepts (HNC) theory. The HNC theory classifies all concepts into abstract concepts and concrete concepts. They shows how word extraction can be done, extraction of features in left and right sides of the sentence and extraction of the features within the sentence. They selected some polysemous words whose different senses usually appear in characteristic contexts which imply different domain information. Their experiment shows that the percentage of accuracy with semantic features and accuracy without semantic features. They compare two methods for finding accuracy such as knowledge based and unsupervised methods. By comparison, they can find that training a classifier by simply using the semantic category of left and right word starting from the target polysemous word performed poorly and did not combine the advantages of knowledge based methods and supervised methods well. Experimental results show that thier approach could extract the contextual semantic features of a certain kind of polysemous words which are helpful for identifying the meaning of it.

Saha Diganta.et al [4] observed that, word sense disambiguation (WSD) in Bengali language has been done using unsupervised methodology. This work is consisted of two sequential sub-tasks. First one is grouping of Bengali sentence into a certain number of clusters where a particular cluster contains the set of similar meaning and second one is labeling the cluster with its inner meanings with the help of linguistic expert as these sense tagged clusters could be used as a knowledge reference for WSD task. In their work, type-based and token-based discrimination strategies have been adopted for sentence clustering. They have been passed through a series of manual normalization procedures. As, uneven number of spaces and new lines have been removed, i.e comma, hyphen, underscore, colon, semicolon, slashes, tilde and other punctuation symbols etc., After the normalization, the text data have been lemmatized. Their feature selection is done by calculating the term frequencies of the individual keywords present in the document. The remaining words have been selected for future vector, the threshold value for pruning is considered, which is manageable length of the feature vector to be handled by the system. After developing the feature vector, they prepared a matrix of vectors of the sentences. Next, the overall test data has been clustered using K-means algorithm. In this strategy, the lexical similarity between the global feature vector and the individual sentence vector is derived. Two challenges have been encountered. First, the term frequency of a feature in a sentence is decreased. And secondly, the intra cluster relations among the features, which represent a particular sense of a cluster, have not been established properly, which left a great impact on the accuracy of output. Finally, in an experimental basis, the algorithm is tested on a single data set.

Michael steinbach.et al[5], they present the result of an experimental study of some common document clustering techniques such as agglomerative hierarchical clustering and K-means. They use two metrics for evaluating cluster quality, which provides a measure of “goodness” for un-nested clusters or for the clusters at one level of a hierarchical clustering, and the F-measure, which measures the effectiveness of a hierarchical clustering. They use bisecting K - means algorithm for best performance. The bisecting step, for a fixed number of times and take the split that produces the clustering with the highest overall similarity. They use three different agglomerative hierarchical techniques for clustering document such as Intra cluster similarity technique, Centroid similarity technique, UPGMA scheme.

Finally they concluded that, compared the two main approaches to document clustering, agglomerative

hierarchical clustering and K-means. Their results indicate that the bisecting K-means technique is better than the standard K-means approach and as good or better than the hierarchical approaches that they tested. In addition, the run time of bisecting K-means is very attractive when compared to that of agglomerative hierarchical clustering techniques.

Robert C. Moore.et al [6] , they present a new method of constructing tag dictionaries for part-of-speech (POS) tagging. Tag dictionaries are commonly used to speed up POS-tag inference by restricting the tags considered for a particular word to those specified by the dictionary. They clearly show that how tag dictionary works and tagging speed. A typical modern POS tagger applies a statistical model to compute a score for a sequence of tags t1, . . . , tn given a sequence of words w1, . . . , wn. The tag sequence assigned the highest score by the model for a given word sequence is selected as the tagging for the word sequence. To make tagging practical, models are normally defined to be factorable in a way that reduces the time complexity. They present a new method that reduces the average number of tags per token to about 1.5, with no loss of tagging accuracy. Then apply a simple variant of Ratnaparkhi's method, with a training set more than 4,000 times larger than the Penn Treebank WSJ training set. They introduce two additional modifications of Ratnaparki's approach. First, with such a large training corpus, they find it unnecessary to keep in the dictionary every tag observed with every word in the automatically-annotated data. So, estimate a probability distribution over tags for each word in the dictionary. Second, since their tokenized version of the English Gigaword corpus contains more than 6 million unique words, they have been reduce the vocabulary of the dictionary to the approximately 1 million words having 10 or more occurrences in the corpus. They have computed a probability distribution p(t|w) using unsmoothed relative frequencies. As noted above, treated all digits as indistinguishable in constructing and applying the dictionary. Finally, their method of constructing a tag dictionary is technically very simple, but remarkably effective. It reduces the mean number of possible tags per token by 57% and increases the number of unambiguous tokens by by 47%. This tag dictionary produces by far the fastest POS tagger reported with anything close to comparable accuracy.

B. Proposed Work

In our project, we have used unsupervised learning technique-Hierarchical sentence clustering algorithm.

We have done preprocessing by extracting individual words from the input text file, then stop words are removed from each sentence in the file by comparing with the predefined list of stop words. Stemming is done by using porter stemmer algorithm. POS Tagger[14] is used for tagging, which are rule-based these taggers try to assign a tag to each word using a set of hand written rules. This means that the set of rules must be properly written and checked by human experts. Similarly measure is done for formation of vector which collects distinct words from the entire input file and stores in a vector. Now vectors can be formed for all the individual sentences in the file with the distinct word count[9]. The distinct words are compared and marked as 1 if the vector with the sentence has the word across the sentence vector. WordNet dictionary is used for determining the similarity[13]. Clustering is performed by similarity measure value based on verbs[8]. This solves problems like complexity and sensitivity.

Here we use hierarchical cluster for cluster analysis. The endpoint is a set of clusters, where each cluster is distinct from each other cluster, and the objects within each cluster are broadly similar to each other. Hierarchical Clustering is a method of cluster analysis which forms a hierarchy of clusters. They are two types, Agglomerative and Divisive. Agglomerative is based on clustering with nouns and Divisive is based on clustering with verbs. Based on the newly refined clusters, sense of the word is identified[12]. K means

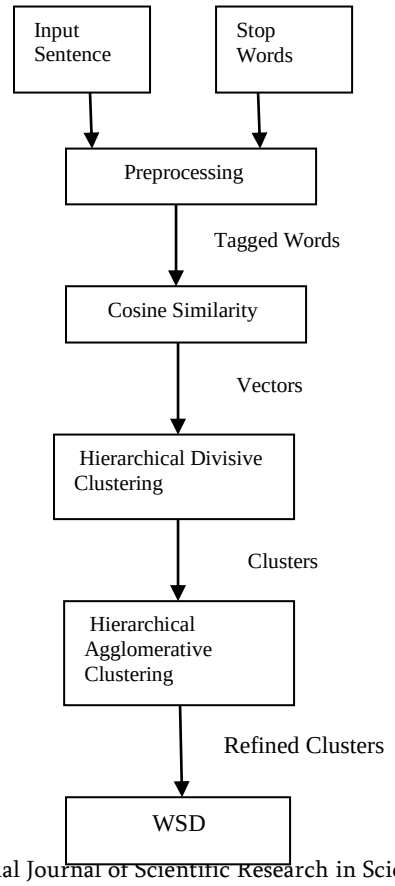
clustering algorithm is used because of its run time efficiency.

Since clustering is used, a sentence can belong to multiple clusters in which each sentence can belong to more than one cluster. It assigns membership values based on the similarity value to clusters such that items in the same cluster are similar as possible, while items belonging to different clusters are as dissimilar as possible. Different similarity measures may be chosen based on the data or the application. Here we use cosine similarity measure.

$$\text{Similarity} = \text{COS}(\theta) = \frac{A.B}{||A||||B||}$$

We use sense identification in (WSD) Word-sense disambiguation for identifying the sense of a word (i.e. meaning). When the word has multiple meanings, the solution to this problem impact other computer-related writing , such as discourse, improving relevance of search engines, anaphora resolution, coherence, inference , et cetera.

Figure 1.1



III. RESULT

Our results on standard datasets proves that we outperform other methods because of sentence level clustering algorithm. Several results are compared from other algorithms where our result which is nearest to the maximum value 1. The standard datasets are used from the corpus- SemCor and Senseval.

Table 1.1

Data set	F1 - Measure
Sem Cor	0.694
Senseval - 2007 Task7	*0.761*

IV. CONCLUSION

A novel method of unsupervised technique is proposed where syntatic and semantic features are identified using the feature vectors of WordNet framework along with domain characteristics in a hierarchical level. Since sentence level clustering is a performed, the accuracy is much high compared to other methods. (i.e) inter cluster similarity is low and intra cluster similarity is much high. Word sense disambiguation for polysemous – hyponym words are handled using the domain features and its characteristics and sub clusters are formed based on noun and verbs which extracts and groups the contextual information together.

V. REFERENCES

[1]. Lingling Meng, Runqing Huang, JunzhongGu [2013]" A Review of Semantic Similarity Measures in WordNet" International Journal of Hybrid Information Technology . Vol. 6, No. 1,PP-1-12.

[2]. Patheja.P.S , Akhilesh A. Waoo ,RichaGarg [2012] ,"Part of speech tagging " International journal of computer science & information Technology (IJCSIT) Vol.3 No 4,PP..

[3]. Liang Wen1, Juan Li1, Yaohong Jin1, Yongjie Lu2 Kogilavani.A , "A Method for Word Sense Disambiguation Combining Contextual Semantic Features" c 2016 IEEE Vol.No.978-1-5090-0922-0/16/$31.00_

[4]. Saha Diganta, Alok Ranjan Pal, "Word Sense Disambiguation In Bengali: An Unsupervised Approach" ©2017IEEE Vol.No 978-1-5090-3239- 6/17/$31.00

[5]. Steinbach, M., Karypis, G., Kumar, V., "A Comparison of Document Clustering Techniques," University of Minnesota, Technical Report #00-034 (2000).

[6]. Robert C. Moore, "An Improved Tag Dictionary for Faster Part-ofSpeech Tagging" Conference on Empirical Methods in Natural Language Processing, pages 1303–1308,Lisbon, Portugal, 17-21 September 2015. c 2015 Association for Computational Linguistics.

[7]. Asim M. El Tahir Ali, Hussam M. Dahwa Abdulla, and Vaclav Snasel [2012] , "Overview and Comparison of Plagiarism Detection Tools "International journal of computer science& information Technology (IJCSIT) Vol 134,No.3,PP-161-172.

[8]. A. and Dr.P.Balasubrama,[2014] "Clustering and feature specific sentence extraction based summarization of multiple documents " International journal of computer science & InformationTechnology (IJCSIT) Vol.2, No.4,PP-99-111.

[9]. Clustering Algorithms for Sentence Level Text" International journal of computer trends and technology Vol 10 No2,PP-61-66.

[10]. Mujawar Nilofar Shabbir, Prof. Amrit Priyadarshi [June 2016]," Clustering Sentence Level Text using Hierarchical FRECCA Algorithm " International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 6,

[11]. Wordnet manual, A Lexical Database-Princeton university-WordNet 2.1 - https://wordnet.princeton.edu/download/current-version

[12]. POS tagger - stanford university https://nlp.stanford.edu/software/tagger.shtml