Handling WSD using Fuzzy Hierarchical Clustering of Sentence Level Text

Mohana Priya K[1], Krishna Priya G[2] , Geetha Harini K R[3], Gowri S[4], Pooja Ragavi S[5]

[1] *Student,I ME CSE, Jansons Institute of Technology,Coimbatore,Tamil Nadu,India*
[2]*Assistant Professor, Department of CSE, Jansons Institute of Technology,Coimbatore, Tamil Nadu,India*
[3]*Research Analyst, Zinnov, Coimbatore,Tamil Nadu,India*
[4]*Online Tutor, Focus Edumatics, Coimbatore,Tamil Nadu,India*
[5]*MBA, GRG School of Management Studies, Coimbatore,Tamil Nadu,India*
[1] *Mohanapriyakm97@gmail.com*
[2]*Krishna.kcse@gmail.com*

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 Fuzzy 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, F measure.

1. 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 with fuzziness is one of the best clustering technique when compared with other type of clustering. The work here deals with the ability to capture such Fuzzy 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,fuzzy 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. Fuzzy C means clustering algorithm, which allows one piece of data that belongs to two or more clusters. This algorithm assigns membership to each data point that corresponding to each cluster. 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. Fuzzy clustering , in which each data point belongs to one cluster. 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.

International Journal of Research in Advent Technology, Vol.6, No.11, November 2018
E-ISSN: 2321-9637
Available online at www.ijrat.org

2. RELATED WORK

Abinaya.et al [1] proposed an approach which produces an extractive summary by selecting sentences from the documents cluster wise. All the relevant documents are grouped together into clusters by using threshold-based document clustering approach. Based on feature profile salient sentences from each cluster are identified and ranked according to their weights of importance. Based on the ranking of sentences, sentences are selected and ordered. The system then iteratively extracts one sentence at a time, until the required Summary is met for each cluster. They use different algorithms for their proposed work. They describe the use of Page Rank and use the Gaussian mixture model approach. Page Rank can be used within the Expectation- Maximization algorithm to optimize the parameter values and to formulate the clusters. Expectation is used to Calculate the Page Rank value for each object in each cluster and Maximization is used to updating the mixing coefficients based on membership values calculated in the expectation step. They use fuzzy relational clustering approach to produce clusters with sentences, where each of them corresponds to some content. This algorithm involves the following steps initialization, expectation and maximization. Hierarchical fuzzy clustering is used in their work, which is used to partitioning the data items into a collection of clusters. This algorithm is an extension of fuzzy relational clustering algorithm. To calculate similarity between the words, they use a common vector space representation for all sentences. The performance metrics used in this paper are Partition Entropy Coefficient (PE), Purity and Entropy, V-Measure, Rand Index and F-Measure. Finally comparison can be done in performance evaluation, the comparison is performed out for 6 numbers of clusters. They compare the performance of FRECCA algorithm with ARCA, Spectral Clustering, and k - Medoids algorithms to the quotations data set and evaluating using the external measures.

LinglingMeng.et al [2] 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 [3] 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[4] discussed about traditional WSD methods based on supervised learning. They consider only the word, position, part of speech and some other superficial morphological and

International Journal of Research in Advent Technology, Vol.6, No.11, November 2018
E-ISSN: 2321-9637
Available online at www.ijrat.org

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 knowledgebased 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 [5] 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 Kmeans 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.

Sadina Gagula Palalic.et al [6] discussed about fuzzy clustering models and algorithms for pattern recognition. Clustering is one of the most essential processes in pattern recognition, since it plays a key role in finding the structures in data. Considering that real world problems in pattern recognition require the stochastic data to be processed, fuzzy clustering method will be shown to serve best in this case. Fuzzy clustering can be divided into two basic groups, namely Fuzzy point prototype and fuzzy non-point prototype, based on fuzzy partitions. Fuzzy hierarchical clustering method based on fuzzy equivalence relations. Cluster validity is used to measure the quality of clustering result. The quality of clustering process can depend on several factors such as method of initializing cluster centers, chosen model for clustering, the choice of the number of classes etc. For measuring cluster validity in fuzzy clustering, some criteria from crisp cluster analysis have been adopted to fuzzy clustering. They use two different methods for clustering such as cost optimization function and fuzzy equivalence relations. In the first approach, they have seen that depending on the data type and the clustering tendency of the data different clustering models give different results.
If the data tends to cluster in a cloud like shape, where they search for the point to be a cluster centre, the best model to serve the clustering is fuzzy c-means model. Some extensions of this model are also presented, such as fuzzy possibility c-means model. On the other hand there are models which can determine the cluster centers as a line, elliptical or circular shapes. Another approach to data clustering is based on fuzzy equivalence relation. This model is definitely better approach in some cases, especially if the user has no idea on the data clustering tendency or the number of clusters, which needs to be specified in the first approach. However, any case chosen, clustering validity needs to be performed in order to check whether proper number of clusters is selected.

Robert C. Moore.et al [7] , 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

International Journal of Research in Advent Technology, Vol.6, No.11, November 2018
E-ISSN: 2321-9637
Available online at www.ijrat.org

restricting the tags considered for a particular word to those specified by the dictionary. They clearly shows 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.

3. PROPOSED WORK

In our project, we have used unsupervised learning technique-Hierarchical fuzzy 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 else fuzzy comparison[10] is done. 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. Fuzzy clustering is done in a hierarchical level[11]. 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]
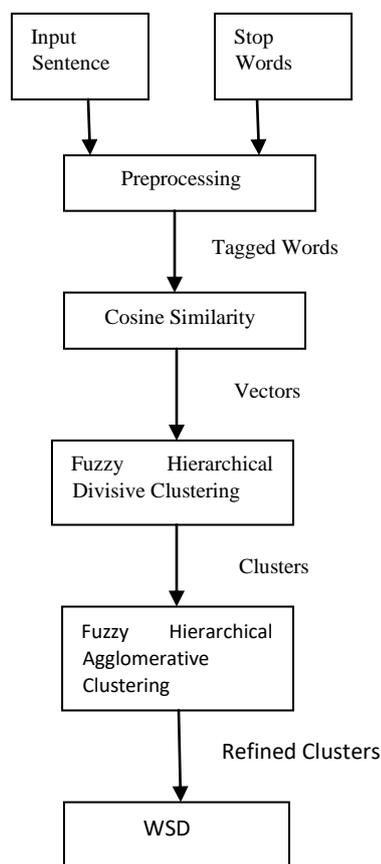
Figure 1

Since fuzzy 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 which is depicted in Eq(1).

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

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 , etc. The entire process is defined in Fig 1.

4. RESULT

Our results on standard datasets proves that we outperform other methods because of sentence level fuzzy 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.

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

TABLE 1

5. 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 fuzzy 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.

ACKNOWLEDGEMENT

REFERENCES

[1] Abinaya. V, M. Vennila, N. Padmanabhan [2014] "Sentence Level Text Clustering using a Hierarchical Fuzzy Relational Clustering Algorithm" International Journal of Communication and Computer Technologies Vol 2,No.10,PP-50-55.

[2] LinglingMeng,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.

[3] 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..

[4] 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_

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

[6] Sadina Gagula Palalic ,"Fuzzy Clustering Models and Algorithms for Pattern Recognition"

[7] 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.

[8] 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.

[9] 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.

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

[11] 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,

[12] G. Krishna Priya, G. Anupriya, "Clustering sentence level-text using fuzzy hierarchical algorithm "IEEE Conference ICHCI-IEEE-2013-6887778, September 2014.

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

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