Research Journal of Applied Sciences 11 (10): 1115-1121, 2016
ISSN: 1815-932X

Textual Features Extraction and Clustering using Semantic Analysis

Ghaidaa A.Bilal and Rasha N. Shalaan
InformationTechnology College, University of Babylon, Hillah, Iraq

Abstract: A set of customers' reviews about restaurants has been analyzed syntactically and semantically for deducing syntactic, contextual and semantic features to leverage the textual similarity metrics. In this study an approach for rule based extracting semantic features from customer's reviews have been proposed. The features were extracted based on the external knowledge base (Word Net), co-occurrence and distributional similarity among the reviews' aspects and descriptors and then an algorithm has been created for grouping the aspects naturally by basing on the computed similarity features. The proposed system has applied on the Yelp academic challenges dataset and the results have shown encouraged performance.

Key words: Semantic analysis, textual aspec, descriptor, aspects, context

INTRODUCTION

The concept of aspects mining is one of the attempts to extract aspects and analysis its sentiment using the pair aspect-sentiment or aspect-descriptor (Lin *et al*., 2015). Online review websites such as Yelp had provided a way for information seekers for browsing user reviews and opinions about variousaspects of service at and restaurants (Gupta *et al*., 2015). However, such sites typically have contained a huge amount of opinionated text that are not always easily deciphered in blogs. The average human reader have difficulty identified relevant sites and accurately summarizing the information (Witten *et al*., 2011). Therefore text clustering considered a useful technique that aims to organize large collections of document into smaller manageable and meaningful groups, an essential role in information retrieval has been played by text clustering. Usually traditional clustering algorithms are based on the BOW (Bag of Words) approach (Holzinger *et al*., 2014). The disadvantage of BOW is the ignoring the semantic relationship among words so that it cannot represent the documents meaning accurately. As text documents growth rapidly, the textual data become variety of vocabulary, high dimensional, as well as it has contained semantic information. Therefore, it is possible that the theme of documents could be represented correctly by text clustering techniques and improved clustering performance where recently a number of semantic-based approaches have being developed. Word Net (Miller, 1995) which is one of the most commonly used thesauruses for English, has been extensively used for improving text-clustering quality with its semantic relations of terms (Amine *et al*., 2010; Bouras and Tsogkas, 2012). However, several problems exist in the clustering results.This study has attempted for solving these problems by considering semantic relation (synonyms and hyponyms) among the extracted aspects and leveraging these relations for implementing the aspect and its' synonyms and hyponyms to indicate to one element. This process has adopted in feature extraction steps and in the clustering procedure.

Literature review

Related work: Recently the problem of detecting semantic similarity in text has led the researchers to give the opinion analysis much attention. Lin *et al*. (2015) the researchers tried to extract opinion lexicons from reviews and identify the sentiment polarities of the words based on a word vector and matrix factorization. The Term Frequency- Inverse Document Frequency TF-IDF feature and Cosine function was utilized as similarity metrics. The researchers missed the semantic analysis in their research; in which the identification of the relations among the vocabularies might be strengthen the similarity process. While, Li *et al*. (2015) leveraged the terms' relations using word co-occurrence and TF-IDF method to identify a set of hierarchical relations among terms. They tried to employ the keywords as concepts source to build text taxonomy. The researchers in Hoang *et al*. (2009) exploited the features extraction approach by using normalizing (Pointwise) Mutual Information to categorize the Association Measures (AMs) into two groups, rank equivalence had been used to group AMs with the same ranking performance. In addition, many researches had given their attention for text clustering techniques. Where for dealing with text clustering a huge number of techniques have been proposed. Clustering similarity

Corresponding Author: Ghaidaa A.Bilal, InformationTechnology College, University of Babylon, Hillah, Iraq

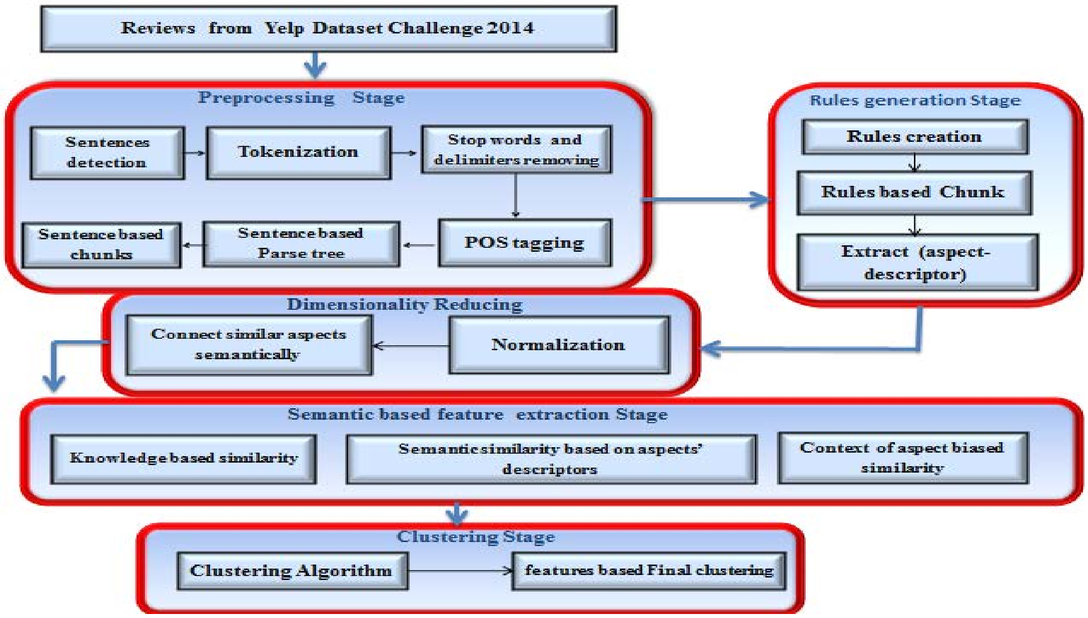

Fig. 1: The system framework

measures may depend on WordNet asknowledge resources (Wei *et al*., 2015). Guo *et al*. (2009), the words had grouped into a set of concepts according to their context documents by using latent semantic association model, then product features had categorized based on latent semantic structures of that words. In asymptotic manner to the proposed approach, researchers by Wu *et al*. (2009) had identified noun and verb phrases as aspect and opinion expressions. This work encouraged our approach for developing syntactic rules to extract aspects and their describing adjectives. Consequently, the researchers by Agarwal *et al*. (2015) adopted a method in which the aspects and its descriptors for a set ofreviews extracted based on syntax rules and clustered based on three features distributional similarity, co- occurrence and knowledge base. We have modified that method by building a new syntax rules which have used to extract aspect-descriptor pairsin a height accuracy and in two directions forward propagation and backward propagation. Where the proposed systemhas aluminized the lack of the previous way for extraction the aspect, for example the aspect was extracted as "wooden pool table", the proposed forward exploration has enhanced the identification and extraction for the aspect and its descriptor to be "pool table" as aspect and "wooden" as descriptor. The proposed system has applied semantic approach in co-occurrence feature where the exactly matching, synonyms and hyponyms of each aspect have treated as one aspect. As well as our approach has tried to reduce dimensionality by in clustering process where each aspect with its synonyms and hyponyms have considered as one dimension to implement initial clustering. The results we have achieved proved that our method helped the users to perform semantic search for better understanding to the reviews content.

Semantic based features extraction: We have adopted an approach to leverage Semantic-Syntax relations based customers reviews for getting the most importing terms and it's descriptors while gaining features that have been helped in analysis the reviews content semantically as showing in Fig. 1.

MATERIALS AND METHODS

Preprocessing stage

Sentences detection: The input to this stage is set of online customers reviews each review document has segmented into sentences .The reviews were filtered in which any unrelated information with the customers opinion was removed. This process has taken into account sorting the sentences according to their reviews.

Tokenization: Each sentence for each review has separated in to set of tokens "terms". The separation process has adopted the segregation each sentence into set of single tokens (uni-gram) in which each token may identify the sentence and the review that belongs to.

Table 1: The part of speech tags meaning

Tag	Meaning	Tag	Meaning	Tag	Meaning
CC	Coordinating conjunction	JJS	Adjective, superlative	SYM	Symbol
CD	Cardinal number	NN	Noun, singular or mass	TO	to
DT	Determiner	NNP	Proper noun, singular	VB	Verb, base form
NNS	Noun, plural	PDT	Pre-determiner	VBD	Verb, past tense
VBG	Verb, gerund or present participle	PRP	Personal pronoun	FW	Foreign word
IN	Preposition or subordinating conjunction	RB	Adverb	VBN	Verb, past participle
VBP	Verb, non¬3rd person singular present	RBR	Adverb, comparative	JJ	Adjective
JJR	Adjective, comparative	RBS	Adverb, superlative	VBZ	5Verb, 3rd person singular present

Stop words and delimiters removing: Stop words are words with less weighting in the reviews with no specific rules to be considered for identification those words. The researchers themselves could select list of words that candidate to be stop words according to their work domain. There are many copies of stop words such as a stop word list that provided by website of Journal of Machine Largening Research; it consists of 571 words. In this study, the adopted dataset has suffered from noisiness including of spelling errors, informal expressions, abbreviations and improper punctuations.Hence, the list of stop words and delimiters was modified to get rid of the above-mentioned noises and to filter the sentences tokens. A sample of the words that were considered as stop words is: “ bla” , “\n” , “gooooood”, “where” ,”a”, “the”.

POS tagging: In corpus linguistics a part-of-speech Tagger is the operation of encoding the text words as corresponding to a particular part of speech. The tagging operation is contingent on both the words definition and context, i.e. relationship among adjacent words in a phrase, sentence, or paragraph. Taking the following example: “John often gives a book to Mary” that has tagged as: John/NNP often/RB gives/VBZ a/DT book/NN to/TO Mary/NNP. The Table 1 shows some of the tags symbols and it’s meaning.

It is clear from the example above that a tagger is basically a classifier where it considers text as input and returns the parts of speech for all its tokens (words) classified as verb, adverb, adjective and noun … etc. The online version of Stanford parser that has used in this research is available at

Chunking: It is a technique widely used in natural language processing for sentence analysis and constituents (noun groups, verbs, verb groups, etc.) identification. However, it neither specifies their internal structure nor their role in the main sentence. It is similar to the concept of lexical analysis in computer languages translators. A unigram chunker simply assigns one chunk tag to each POS tag where in The IOB representation every token is in a chunk or Out of a chunk.

Rules generation stage: A set of custom syntactic rules has been built to identify and extract pairs of aspects and descriptors for each review based on generated syntactic rules. It follows two directions in generating the rules: backward exploration and forward exploration.

Backward exploration: In this part of the algorithm, choosing the candidate pair starts from detecting the noun phrase in each sentence to select the elected aspect. Next backward search go back to the beginning of the sentence looking for any nominee descriptors that is located in front of the aspect. As described in the mentioned algorithm, the descriptors should be an adjective or past participle, e.g. “ They had prepared a delicious chicken”, where the extracted pair is (chicken, delicious).Several studies haven’t focused on extracting aspects in high accuracy as the rules are proposed in this research. For example “pool table” and “wooden pool table” most likely refer to the same aspect and they used Jaccard similarity metric, while this is a time consuming as compared with our approach wherefrom the aspect they had extracted “wooden pool table” our presented rules have extracted a pair of aspect- descriptor as (pool table, wooden) where this has leaded to exact matching with the otheraspect “pool table” and reduced time consuming .

Forward exploration : In this direction, the tokens located behind the elected aspect are tracked to extract new descriptors. This part of work has leveraged the fact of nouns usually are followed by adjectives so that it can say the discovered adjective would be the elected descriptor of that noun. With in this direction of rule generating many sub rules have described as the following:

If there is an auxiliary verbs in the sentence then it will be followed by a descriptor e.g. “ The waitress was rude”. This rule has extracted the aspect-descriptors pair as (waitress , rude) .

If there is a conjunctive in the sentence such as but while and thenthe sentence is separated into sub sentences each of them is treated as a new sentence. The identification and detecting process of (aspect-descriptor) pair have repeated on every sub sentence. For instance "nice texture but the service was bad".This rule has identified the conjunctive "but" then the sentence has divided into two sub sentence "nice texture " and "the service was bad ". Then the extraction process of aspect-descriptor pair for each sub sentence has started and the created rules have extracted the aspects-descriptors pairs as (texture , nice) and (service , bad).

The research of this rule is to check if the aspect in the sentence has more than one descriptor, e.g., "the chicken was tough and hot"then this rule has extracted aspect- descriptors pairs as (chicken, tough) and (chicken, hot). Sometimes the people often express their opinions about an aspect by using past participles e.g. " I liked the fried fish" or by using present participles e.g. "I like the dish sizzling". This rule is responsible of identifying if there is a past participles or present participles as a descriptor in the sentence then this rule has extracted the aspect- descriptor pairs as (fish, fried) and (dish, sizzling), respectively .

Rules based chunk: In the previous step set of rules to extract aspect and it's descriptors has created. At this step the created rules have written in chunks format to be groups of chunks . The chunked rules have mapped with the given chunked sentence for identifying and extracting aspect descriptor pair. The algorithm 1 has illustratedthe chunks have usedto implement the created rules and the (aspect-descriptor) extraction process. The output of this stage is a list of aspect- descriptor pairs.

Algorithm1

Rules generating:

```
Name: Rules creation algorithm
Input: List of sentences parse trees
Output: Set of aspect-descriptor pairs
Begin
    While reviews have sentences
    For I= first token to the last token in sentence
if chunk of token [I] is "NP" and tag is ("NN" or "NNS" or "VBG") then
save token (Zheng Lin et al., 2015) as aspect
    For J=I-1 to 0  // Back Exploration
    If tag of token[J] ="JJ" or "VBN" or "VBG" and If it negative or
positive then save token[J] as descriptor
    I=I+1 //forward propagation
    check If token[I] is "JJ" or "VBN" or "VBG" and If it negative or
positive then save token (I) as descriptor
    Else
    If tag of token[I] is "PRP" then save it as "For business" to be aspect
and return to step 6
    Else if token[I] is one of Connectivity Tools such as "but" or "while"
etc. then treat it as new sentence and return to step 2
    End for
    End while
    End
```

Dimensionality reducing

Normalization: All the aspects had extracted from the review documents are lemmatized to reduce to their root form for enhancing the features extraction process in the dataset. In order to extract the features in high accuracy, all aspects must be transformed into uniform case. This mean all the "aspects " and "descriptors" must transform into capital letters or into lower letters. Hence in this stage if words have differed by the letters case small or capital, after the transformation has performed, they would be treated as same words, e.g., food and Food have transformed into food.

RESULTS AND DISCUSSION

Semantic based similar aspects connectedness: The previous step has used for reducing the dimensionality. As we have noted, extraction the nouns to be aspects may increase the dimensionality of the feature space. We need to seek a way to reduce the dimensionality while achieving clustering performance in comparison to using all the nouns. In this step, for each aspect we have extracteda subset of it's synonyms and hyponyms with the help of information from the WordNet ontology.Each subset has indicatedto one dimension in the clustering stage . This step has considered each aspect as a head of cluster, then the head of cluster has grouped with its' synonyms and hyponyms to be in a same cluster. The goal of finding out the representative terms and their relationships may represent the main theme of the topics in the customers reviews clustering .

The extracted features

Context of aspect based similarity: The liturture showed that the aspects which co-occur in the same context, are mostly related and belong to the similar group (Holzinger *et al*., 2014). It missed to consider the semantic occurrence of the compared aspects. Hence, in this study at first all synonyms and hyponyms for each aspect have identified and repetition among these synonyms and hyponyms have deleted. Then the context information for all sentences in the review have gathered into a context vector, that used for comparison the semantic co-occurrence of the aspects and their synonyms and hyponyms with the all other aspects have presented in the same review. The association strength for each two aspects in the context vector has measured by the Point wise Mutual Information(PMI), in which the frequency of the two aspects that appear in the reviews together has compared to their frequencies separately, as shown in Eq. 1.The computation process of the aspect's context similarity has presented in Algorithm 2.

$$\mathrm{PMI}(x,y)=\log\left(\frac{p(x,y)}{p(x),p(y)}\right) \tag{1}$$

Algorithm 2

Context similarity of the aspects:

Name : Context similarity of the aspects algorithm
Input : Record of Aspects
Output : Record of Aspects with PMI value
Begin
Step1 : For I= first Aspect to the last Aspect
begin
Take aspect [I] with all other aspects
For J= I+1 to the last Aspect
Convert aspects to lower case
find the synonyms and hyponyms for aspect [I]
Compute Probability of occurrence and co-occurrence of aspect[I] and it's synonyms and hyponyms with all other aspects.
Apply equation 2 of each pair for aspects and save in PMI
Step4 : Save result End if End for
End

External knowledge base based similarity: The semantic similarity between two aspects has identified by the Word Net knowledge base which it has used several semantic relations such as synonymy, autonomy, hyponymy and so on. These relations can be used for word form relation or for semantic relation as a hierarchy structure for which the Word Net is regarded as a good tool for natural language processing. The word Net has provided four types of relations among nouns that may occur. The first relation is hyponym/hypernym relation that denoted as (is-a) relation, e.g., "Ali is a boy". The second one is meronym/part holonym relations(part-of) e.g. "battery is part of mobile". While, the third relation is expressed by (member-of) member meronym/member holonym relations that defines the relationship between a two terms one of them denoting the whole and the other denoting a part of, or a member of, e.g., the relation between the head and body. The last type of relation is the (substance-of) meronym/substance holonym relation which identifies how a word or phrase is used to stand in another word, e.g. "The pen is mightier than the sword". Word Net semantic similarity measures have been grouped in four classes types: path length based measures, feature based measures, information content based measures and hybrid measures. At this feature the shortest path based measure has adopted, where the sim(ci,cj) has considered the closeness ci and cj in the taxonomy as shown in Eq. 2:

$$\mathrm{Sin}_{\mathrm{path}}\left(C_1,C_2\right)=2\times \mathrm{deep\_max}-\mathrm{len}\left(C_1,C_2\right) \qquad (2)$$

Where, deep_max is a fixed value.The similarity between c1, c2 is len(c1,c2) from c1-c2. If len(c1,c2) is 0, simpath(c1,c2) gets the maximum value of 2*deep_max. If len(c1,c2) is 2*deep_max, simpath (c1,c2) gets the minimum value of 0. Thus, the values of simpath (c1, c2) are between 0 and 2*deep_max.

Distributional similarity of descriptors: The relationship of the aspects could be reflected by their descriptors, where descriptors may provide virtual contexts similarity to the unrelated aspects that neither have co-occurrence in the same contexts nor they have relation could be identified by Word Net. To extract this feature a word-to-word similarity-normalized PMI-metric that used by Holzinger *et al.* (2014) has been adopted to indicate the semantic similarity among the descriptors of two aspects in the all reviews as exhibited in Eq. 3. Some descriptors which are not reflect their aspects or consider as common words such as "good", "bad" and so on,were ignored from the comparison because they may effect negatively on the results:

$$\mathrm{Sim}\left(A_1,A_2\right)=\frac{1}{2}\left(\frac{\sum_{d\in A_1}\left(\max \mathrm{sim}\left(d,A_2\right)+\log\left(N/n_d\right)\right)}{\sum_{d\in A_1}\log\left(N/n_d\right)}+\frac{\sum_{d\in A_1}\left(\max \mathrm{sim}\left(d,A_1\right)+\log\left(N/n_d\right)\right)}{\sum_{d\in A_2}\log\left(N/n_d\right)}\right)$$

where, A_1 and A_2 are aspects, d is descriptor, N are the aspects' total number, n_d, the number of aspects that d appears with. The algorithm 3 below described the extraction of this feature.

Algorithm 3

Distributional similarity of descriptors:

Name : Distributional Similarity of Descriptors
Input : Set of aspects' descriptors
Output : Set of Aspects with Maximum Similarity values based on its' descriptors
Begin
Step1 : For I= first Aspect to the last Aspect
Take aspect [I] with all other aspects
For J= first Descriptor of Aspect [I]to the last Descriptor
For K= first Descriptor of Aspect [I+1]to the last Descriptor
Convert Descriptors to lower case
Compute Probability of occurrence and co-occurrence of Descriptor[J] and Descriptor[K]
Apply equation 4 on the Descriptors pair and Save the result
select max result from the previous saved results for each two aspects
Apply equation 5 on max result and save result to be similarity value of the pair of aspects
End

Features based final clustering: In this study, the extracted features from the last section are grouped in terms of similar aspects into set of clusters. the semantic similarity between features are taken into consideration to be the input properties for the clustering process. The initial clustering process has been performed in which each aspect has represented as a head of cluster. Then

Table 2: Aspect descriptors pairs

Descriptors	Aspect	The detected sentence
Rude	Waitress	The waitress was rude
Poor	Customer service	Very poor customer service
Stale	Plate	Everything on my wife's plate was stale
Tough	Chicken	The chicken was tough and the soup had no flavor
Had no flavor	Soup	
Not cleaned	Glass	The glass was not cleaned inside or outside for quite some time
Looked busy	Parking lot	The parking lot looked busy
Reasonable	Prices	The prices are reasonable, and the owners are very friendly
Friendly	Owners	

Table 3: The Aspects' features values

Aspect1	Aspect2	Word net	PMI	Distributional similarity of descriptors
Waitress	Food	0.0909090909090909	0	0
Waitress	Hostess	0.111111111111111	0	1
Food	Cashier	0.142857142857142	1.968448971231	0
Food	Meal	0.333333333333333	0	1
Hostess	Shrimp	0.166666666666666	0	0
Plate	Rib	0.333333333333333	0	0
Service	Restaurants	0	0	0.968576925045153
Service	Meal	0.0909090909090909	0	0.487239038084645
Chicken	Soup	0.111111111111111	0	0
Sauce	Patty	0.111111111111111	0	0
Meal	Owners	0	0	0

the head of cluster has grouped with its' synonyms and hyponyms to be in same cluster. The grouping process has been accomplishedbased on the distance among the aspects as It has compared the first head of cluster with all other heads of clusters based on the three extracted feature values. If any of the other heads of clusters has features values greater than a pre-fixed threshold, the chosen head of cluster with its' synonyms and hyponyms would be added to first cluster and so on . The merged clusters have removed from the set of clusters which need to be checked.After the clustering has been performed, aspects have clustered into natural groups. For example, in restaurant reviews, natural groups of aspects are summed up may be about food, some particular type of food, restaurant etc. It is done through aggregating aspects in terms of terms similarity and using the following features: Context or co-occurrence of aspects, External knowledge base based similarity, Semantic similarity based on aspects' descriptors(Algorithm 4).

Algorithm 4

Semantic features grouping:

```
Name: Semantic clustering algorithm S.C.A
 Input  : Set of  initial clusters
Output : Sets of  final clusters based on extracted semantic features
Begin
            For I = first initial cluster to the last one
   Begin
            If any of semantic features values of cluster[I] is
greater than threshold (1,2,3) then add cluster [I] to
            Give the aspect[I] an Cluster ID
            By using WordNet  find the synonyms and
hyponyms of aspect[I]
            For J=0 to last aspect
Begin
            If J=I then J++
            If aspect[J] is a synonym or hyponym of aspect[I]
then give it the Cluster ID of aspect[I]
End for
      End for
         End
```

The experiments and results: The experiments of this research have been implemented using Java platform on NetBeans IDE 8.0.2. This program has used Stanford POS tagger andWordNet Ontology for finding the relations (synonym, hyponyms) between the words and for providing semantic similarity among the aspects. The input of the proposed system is set of online businesses reviews from Yelp website "Yelp Dataset Challenge 2014"[1]. The reviews had been written in informal language where the customer didn't care to the language rules, e.g. one of the customers started his review with" bla bla bla" to indicate the food was bad and some of them didn't care to the spelling e.g. " gooooood ", abbreviations, improper punctuations such as " \nOpen " and some adjectives had returned as noun e.g." old" .POS tagger has considered all these outliers as noun. The reviews has considered to be input to the proposed system are 152 reviews, 1237 sentences have detected, while aspects have extracted are 1363 and the total descriptors for all aspects are 1475.The Table 2 shows sample of the extracted aspect- descriptor pairs.

The results have presented in Table 3 shows that the ability of generated rules to extract the aspect and descriptors in a high efficiency, where if there are more than one aspect or descriptor in the sentence the rules have detected them. A sample of the features extraction has shown in Table 3.

It is noted that, the related aspects have a WordNet similarity value ranging from 0.1 and above, while aspects with no clear related, other features could be used to discover if there is a similarity among them e.g. “service” and” restaurants” the similarity between them has discovered by distributional similarity of descriptors feature. The case of PMI feature based on occurrence and co- occurrence for the aspects, there for most of its’ values are 0. While in clustering stage two clusters have resulted after applying the proposed semantic clustering algorithm, the clusters are:

- Cluster 1: {waitress, hostess, business, Owners}
- Cluster 2: { food, eating experience, place, plate, customer service, rib, lettuce wedge, parking lot, glass, service, lunch, Breakfast, soup, restaurants, chicken}

For evaluation of 120 reviews which have been clustered by our clustering method, we used the precision and recall measures, we have captured an encouraging results, the results were Precision 0.87 and Recall 0.92for the main dataset.

CONCLUSION

This study has introduced a developed approach for discovering the aspects and extracting the descriptors of these aspects by building a set of syntax rules have treated all the syntax cases of the sentences and extracted aspects-descriptors pairs purely from reviews and then these aspects and descriptors have analyzed semantically and the features values have extracted where the resulting values have given an overview of similarity between aspects as showing in Table 3, by basing on the computed features the aspects have been clustered by using the adopted S.C.Awhere the values of precision and recall have proved the performanceof our approach. In some sentences there is descriptor belongs to a pronoun. The future work is to discover the pronoun belongs to which aspect

REFERENCES

Agarwal, B., N. Mittal, P. Bansal and S. Garg, 2015. Sentiment analysis using common-sense and context information. Comput. Intell. Neurosci., 2015: 1-9.

Amine, A., Z. Elberrichi and M. Simonet, 2010. Evaluation of text clustering methods using wordnet. Int. Arab J. Inf. Technol., 7: 349-357.

Bouras, C. and V. Tsogkas, 2012. A clustering technique for news articles using word net Knowl. Based Syst., 36: 115-128.

Guo, H., H. Zhu, Z. Guo, X.X. Zhang and Z. Su, 2009. Product feature categorization with multilevel latent semantic association. Proceedings of International Conference on Information and Knowledge Management, November 2-6, 2009, ACM, China, ISBN:978-1-60558-512-3, pp: 1087-1096.

Gupta, P., S. Kumar and K. Jaidka, 2015. Summarizing Customer Reviews through Aspects and Contexts. In: Computational Linguistics and Intelligent Text Processing, Alexander, G. (Ed.). Springer, Berlin, Germany, ISBN:978-3-319-18116-5, pp: 241-256.

Hoang, H.H., S.N. Kim and M.Y. Kan, 2009. A re-examination of lexical association measures. Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications, August 6-6, 2009, Association for Computational Linguistics, Stroudsburg, Pennsylvania, USA., ISBN: 978-1-932432-60-2, pp: 31-39.

Holzinger, A., J. Schantl, M. Schroettner, C. Seifert and K. Verspoor, 2014. Biomedical Text Mining: State-of-the-Art Open Problems and Future Challenges. In: Interactive Knowledge Discovery and Data Mining in Biomedical Informatics, Andreas, H. and I. Jurisica (Eds.). Springer, Berlin, Germany, ISBN: 978-3-662-43967-8, pp: 271-300.

Li, S., Y. Sun and D. Soergel, 2015. A new method for automatically constructing domain-oriented term taxonomy based on weighted word co-occurrence analysis. Scientometrics, 103: 1023-1042.

Lin, Z., W. Wang, X. Jin, J. Liang and D. Meng, 2015. A word vector and matrix factorization based method for opinion lexicon extraction. Proceedings of the 24th International Conference on World Wide Web, May 18-22, 2015, ACM, Florence, Italy, ISBN: 978-1-4503-3473-0, pp: 67-68.

Miller, G.A., 1995. WordNet: A lexical database for English. Commun. ACM, 38: 39-41.

Wei, T., Y. Lu, H. Chang, Q. Zhou and X. Bao, 2015. A semantic approach for text clustering using Word Net and lexical chains. Expert Syst. Appl., 42: 2264-2275.

Witten, I. H., E. Frank and M.A. Hall, 2011. Data Mining: Practical Machine Learning Tools and Techniques. 3rd Edn., Morgan Kaufmann, USA.

Wu, Y., Q. Zhang, X. Huang and L. Wu, 2009. Phrase dependency parsing for opinion mining. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, August 6, 2009, Association for Computational Linguistics, Stroudsburg, Pennsylvania, USA., ISBN: 978-1-932432-63-3, pp: 1533-1541.