International Journal of Computational Intelligence Research
ISSN 0973-1873 Volume 12, Number 2 (2016), pp. 169-186

http://www.ripublication.com

A Novel Automatic Approach For Extraction and Classification of Noun Phrase Collocations

[1]C.Gnana chithra and [2] Dr.E.Ramaraj

[1]Equity Research Consultant, Angeeras Securities, Chennai, India.
Email: chitbalu2000@yahoo.com
[2]Professor, Department of Computer science and Engineering,
Alagappa university, Karaikudi, India.
Email: eramaraj@rediffmail.com

Abstract

In 2016 it is estimated that 2.4 billion users access the Web and people spend time on net invariably. Search engines are queried for millions of facts, and it retrieves the resultant web pages relevant to the query. The degree of high precision Information retrieval is based on the Query Phrase. A best noun phrase collocation extraction algorithm targets the combined strategic goal of achieving high retrieval efficiency and annotating the web semantically.Two methods discussed in this paper helps in classifying collocations. Initially the web pages are taken as input followed by pre-processing and POS tagging. In the first method, the rule based algorithm tries to identify the n-gram collocates of arbitrary order using the Ontologies, web and dictionaries together. Since this study focus on noun Phrase collocation, filters are applied to the results to fetch only the Noun Phrase Collocates(NPC). The correlation measure between the NPC is measured using the probabilistic association measures techniques such as chi-square test, Log-likelihood test and Point wise Mutual Induction(PMI). In the second method, the phrase is classified as Noun Phrase Collocate and Non-noun Phrase Collocates using optimal threshold value using Naïve Bayes and C4.5 for supervised classification and simple k-means for unsupervised learning.The proposed algorithm is tested with web pages from different corpus and the experimental results prove that the resultant noun phrase collocates has high degree of association between them. The hits during search with highly ranked Noun Phrase collocates is greater than other POS Phrase collocations. The F-score of the proposed algorithm on dataset is 82.3% .

Keywords: Noun Phrase Collocation, Probabilistic association measure, Collocation extraction, Noun Phrase filter, POS tagging.

INTRODUCTION

Wikipedia [1] defines collocation as "In corpus linguistics a collocation is a sequence of words or terms that co-occur more often than would be expected by chance".

Humans possess infinite energy and power to generate any number of sentences in the natural language by combining words and within the boundaries of syntax and semantic rules. Not all words can change their regular structure. Some fixed phrases made up of sequential words and of complex rules with restrictions, and the continuous words which occur often, are called as Collocations. Collocations are also called as Multi Word Expressions(MWE). Detection and extraction of collocation in Natural Language Processing is very important in query formulation for retrieving the document. Noun Phrase collocations to a larger extent are domain dependent. Only a small number of idiomatic collocations are available in the dictionaries.

Collocations can be classified into Proper Nouns(e.g. Sachin Tendulkar, Jammu and Kashmir),idiomatic phrases(e.g. Kick the bucket, red tape), Phrasal verbs(e.g. take care, switch on), lexically limited words(Strong tea) and domain dependent terms (bull market). Detection of Collocations of very high accuracy is not achievable due to the fact that, they cannot be classified alone with the help of syntax rules

When there is no free will for combinability, the sequence of noun words are often termed as Noun Phrase Collocates. In assigning metadata during automated semantic annotation of the webpages, Noun phrase Collocates play a major role.

DC(Dublin core) metadata such as Title, creator, Publisher, Contributor may need the proper noun phrases for annotating the documents.

According to Manning and Schutz[2],Collocations are characterized by at least three properties. They are Non or limited compositionality, Non or limited substitutability and Non or limited modifiability. Due to the limitation of the three properties, the collocates cannot be easily translated from one language to another.

Compositionality of collocation can can be described as the total meaning of the Noun phrase is derived from the meaning of the individual words in the phrase and the grammatical relationship which connects the words. For e.g. Lemon soda is drink which can be understood that soda contains Lemon. Both Lemon and Soda are individual meaningful units. But for e.g. Hot dog burgers, the f decomposition provides a different meaning. Dog and burgers are objects where as hot cannot be termed as an adjective for Dog. It does not mean that the dog is hot but means a variety of food. Hot dog can be cited as an best example for Non compositionality. The meaning assigned to the phrase is not straightforward.

Non substitutability refers to the fact that the nearest synonym cannot be substituted for the collocation. For E.g. There are two kinds of wine. Red wine and white wine.

Since the color of the wine is pale yellowish white, it cannot be termed as yellow wine. Non-modifiability in collocation means that the collocates cannot be modified or supplemental with the additional lexical material or through the transformation in grammars.

Automatic identification of Noun Phrase Collocates helps in Machine Translation and in building semantic knowledge bases. Rule based Finite state Automata grammars and parse trees were used in earlier for Noun phrase detection and extraction. But due to the challenges in the detection process such as non-recursive phrase, ambiguity and unexpected rules statistical and machine learning methods came into existence.

Phrase recognition is the primary mission of Information retrieval Systems. During stage I, Heuristics are used to efficiently extract the collocation Phrases and in stage II it is classified into Noun Phrase entities. The classified Proper Noun Phrase Class such as PERSON, LOCATION, and ORGANISATION are to be used in the automatic semantic annotation systems. Some times the phrase may be unigram, bi-gram, trigram or n- gram. A clear unambiguous phrase fetches the correct expected results where as the other results are more irrelevant and noisy. Many researchers have tried solving the problem of ambiguity with different success ratios. Our new algorithm searches the corpus for new phrases and labels the data as well as iterates on the non-phrase data which has been classified wrongly by the classifier in the previous runs. Unigram and bi-gram produces improved results. Due to time restrictions only bi-gram is only considered for detailed study.

RELATED STUDIES

The concept of collocation was formulated by the English teacher Palmere [3]. In the 3rd century B.C Greek philosophers searched for the collocations. Robin [4] explains that Greeks believed that word meanings may change according to the collocation and the meanings of words does not exist separately. John R.Firth [5] introduced the new concept meaning by collocation in his research study in the lexis and to define the meaning of a single word.

Statistical techniques were used for identifying collocations [6]. Machine Learning Techniques and Data mining methods help in discovering collocates [7]. The application of machine learning techniques on the collocation algorithm identifies the best threshold either provided with or without the labeled data, The very first algorithm on Collocation extraction was presented in 1973 by Berry-Rogghe [8]. Many researchers like Church and Hanks [9], Shimohata [10] and Smadja [11] have made commendable contributions. Smadja in year 1993, in his research work proposed a new methodology to retrieve the collocations by using bigrams, when their co-occurrences are greater than the threshold. Smadja [11] used the concept of context window for extracting collocations and with the assumption that all of the relations involving a word w can be retrieved by examining the neighborhood of w when it happens within a span of five words around the window.

METHODOLOGY

Proposed Algorithm

This paper proposes a new algorithm for collocation detection and extraction, as Gries[12] has felt that, statistical association based collocation methods had deficiency since they use symmetrical measures.

The study of this paper is concentrated on two subjects. One is the extraction of collocation phrases and the other is filtering the Noun Phrase Collocates from among different parts of speech collocates.

The proposed study focus more on Noun phrase collocation extraction. So two variations of this algorithm is discussed in this paper. The first one is Collocations with Noun POS constraint and the second is Collocations without POS constraint. In the first constraint when all the POS components of the n-gram collocate are Noun then the collocate is extracted. But in the second case, POS tags are not considered for the collocation;

Collocation Phrase Extraction pipeline

The term Extraction pipeline in the collocation detection and extraction was first used by [13]. The flowchart for Phrase Collocation extraction pipeline is given in Figure-1. Webpages extracted by the search engines are the inputs for the study. The extracted web data is preprocessed using parsers and stemmers and the Phrases is the output at this stage. The algorithms for collocation extraction is applied to the phrases to find out the frequent combination of words and frequencies of the collocations words. Wordnet and Wikipedia ontology, BNC's Oxford Learners dictionaries and Collins parser dictionary, Geographic Gazateer and the Google search engines aids to find out the best and highly ranked collocations.

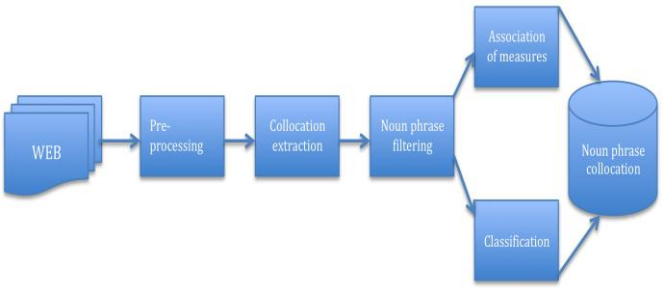

Figure 1: Phrase collocation extraction pipeline

During the filtering phase from the set of Collocates, noun collocates are marked and the rest of collocates are rejected for this study. The degree of independence between the collocate words are measured using the association measures on the Noun Phrase collocates. Words with high frequency of bonding are ranked, and selected as good candidate for Noun Phrase collocate or otherwise it is rejected. Machine learning algorithm uses the extracted feature set for training and testing the data. The candidate phrase with high collocation frequency above the threshold levels are classified as correct Noun Phrase colloctes which is used further in training the classifiers.

Unsupervised clustering algorithms identifies the clusters above the threshold as highly ranked noun phrase collocates and segregates the low ranked collocates.

Algorithm for Collocation Phrase Extraction

Input : List of Phrases or n grams extracted after pre-processing the web document.

Step 1: Take a phrase p1 from the list of phrases P= {p1,p2,p3..pn) in the collection.

Step 2: Compare the phrase p1 with Word Net super thesaurus. If phrase exists then add it to the potential collocation candidate (PCC) set. Go to step 7.;Otherwise goto step 3.

Step 3: Compare the Phrase p1 with the Wikipedia Pronoun ontology. The basic requirement is p1 should be in all capital letters. The result after the search is , if phrase exists it is the first element in the main body add to PCC. If it is a normal noun phrase it need be capitalized.If phrases exists then add to PCC. Go to step 7;
Otherwise goto Step 4.

Step 4: Perform Google search on the p1 and the Search engine result page (SERP) outputs results with ranking then, p1 above the threshold is added to the PCC . Go to step 7; Otherwise goto Step 5.

Step 5: Make a search for p1 in BNC dictionary. If phrase available then add to PCC. Go to step 7; Otherwise goto Step 6.

Step 6: Search Geographic Gazateer for Proper noun Phrase. If it matches add to PCC.

Step 7 : If the phrase cannot be classified as PCC through step 2 to step 6 then mark the phrase as REJECTED CANDIDATE and add it to rejected list.

Step 8: Increment the phrase to p2. Goto step 2 and proceed until the entire set is exhausted.

Step 9: Finally PCC contains the collocation phrases.

Figure 2 : Algorithm for Collocation Phrase Extraction

Preprocessing

The web pages are extracted from the web using Wrappers. Web data extracted from the web is treated as input data. The Audio, Video and the graphics are filtered and rejected
in the html pages. The html tags in the page are cleaned and the Plain text is stored for further processing. The Modified Mannings sentence boundary detection algorithm [14] with extended abbreviation detection and Geographical classification developed for our research on Automatic semantic annotation is applied here for sentence segmentation. Then it is tokenized.
The determiners are classified under stop words and all the stop words are rejected in the document parsing. Google [15] has given the stop list of words which are irrelevant for the queries.

Table 1: Ranked stop list of words by Google

I a about an are as at be by com for from how
In is it of on or that the this to was what when
Where who will with the www able about across
After before later through most truly under unlike

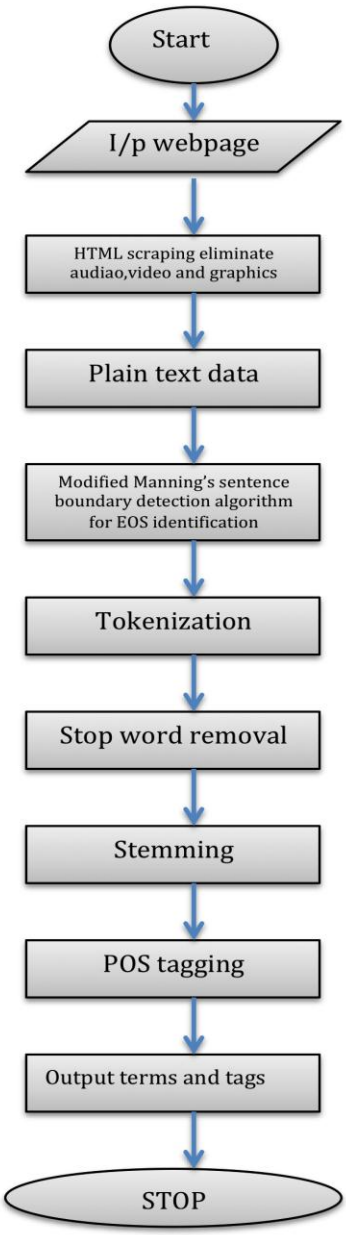

Figure 3 : Pre-processing pipeline

Morphological analyzer performs the study on words, investigates how words are formed and finds the relation between them. Porter's stemming algorithm is used for stemming the words to its root.

After the tokenization process the part-of-speech tagging module works. POS tagging is a lexical process of assigning tags to the words in the web corpora either manually, semi automatically or automatically. Generally for a single word there is one POS tag. POS tagging marks the nouns, pronouns, adjectives, determiners and verb. Some example POS tags are NN for single noun, NP- noun Phrase , VP -verb phrase, PRP-pronoun phrase. Implementation of POS tags is made by Penn Tree Bank Parser. Phrases or Named entities and its associated POS tags are stored in the database.

Table 2: Sample list of POS tags

POS Tag	Description
NN	Noun singular
NNS	Noun ,Plural
JJ	Adjective
DT	Determiner
VB	Verb
IN	Preposition
RB	Adverb
NNP	Proper Noun Singular
NNPS	Proper Noun Plural

Collocation Extraction

Collocation candidate data can be obtained in many different ways. All potential collocates in the corpus by can be found by using correlation of frequency between words, web and ontology. This study is restricted for bigrams but this algorithm can be extended for n-grams. Let is assume that a bigram is made up of two components elements the source word and the collocate word. The words together form the phrase. A context window of size 7 is considered for identifying the bi-gram phrase is represented in the Figure.3

n-3	n-2	n-1	n	n+1	n+2	n+3

Figure 3 : Context window pane of size 7

The most possible collocates lie within n-1 to n+1 range. The word n be prefixed or suffixed by the collocate. For example the word George may be suffixed by Bush which makes it a proper noun phrase. George may be prefixed with modifiers such as Poor George. Considering n-3, n-3, n+2, n+3 words for collocations, though may be not be very useful for bi-grams, but it can work wonders with tri-grams and 4-

grams.The contexts are studied to classify the related domain of its existence.

Table 4: Collocation Phrases examples with POS tagging

COLLOCATION PHRASE	POS TAG
Maiden voyage	JJ+NN
Ceasefire agreement	NN+NN
Dog barking	NN+VB
committing murder	JJ+NN
fully aware	RB+JJ
Subash Chandra Bose	NNP+NNP+NNP

The filter is set to extract only the noun and proper noun elements. In Table 3 the examples “Ceasefire agreement” and “Subash Chandra Bose” are Noun Phrase Collocates.

Filtering Noun Phrase Collocation.
The rule set for Noun phrase collocation identification is

The rule set for Noun phrase collocation identification is
R1 = [DETERMINER] HW;
R2 = [PREMODIFIER] HW;
R3 = HW [COMPLEMENT];
R4 = HW [POSTMODIFIER]
R5 = [DETERMINER] [PREMODIFER] HW [COMPLEMENT] [POST MODIFIER]

Figure 4 : Rule set for noun phrase collocation

A noun phrase can be termed as a phrase which plays the role of a noun. In the noun phrase, the head word can either be noun or pronoun. Head words (HW) are also called as catch words that does the syntax function of the entire phrase. Dependent words occur before the headword or after the headword. Determiners and premodifiers occurs before headword. Complements and post modifiers occur after headword

```
Algorithm  for Filtering Noun Phrase Collocation

Input: Set of Collocation Phrases S= {p1,p2…pn}
Output: Set of Noun Phrase Collocates NPC.
Procedure:
Step 1: Load the Collocation Phrases DB, Determiners
          DB, Premodifiers DB, Complements and
          DBPostmodifiers DB.
Step 2 : set Boolean NPCF:=false;
Step 2:  For each Collocation Phrase CP do
Step 3:   Find the noun HeadWord (HW);
            If prefix of HW is Determiner then
               GoTo step 4;
             Else
               if prefix of HW is premodifier then
               Goto step 4;
                Else
                  If  suffix of HW is Complement then
                    GoTo step 4;
                  Else
                    if prefix of HW is postmodifier then
                      Goto step 4;
                     Else set NPCF:= false;
                     Endif;
                  Endif;
               Endif;
            Endif;
Step 4:  set NPC ← CP; set NPCF ← true;
End for;
Step 5: return NPC;
```

Figure 5 : Algorithm for Filtering Noun Phrase Collocation

Determiners occurs first before headword in noun phrase. Determiners include article, demonstratives, quantifiers, possessive determiners, interrogative words and numerals. The references made by the determiners may be definite, indefinite, possessive or demonstrative. For e.g In the sentence “This flat is the guest house”, “This” is the demonstrative reference and “the” is the definite reference whereas “house” is the headword.

Premodifiers also occurs before the headword in the noun phrase. It may be single nouns, noun phrases, single adjective or adjective phrases. Nouns in premodifiers specify the characteristic features of the noun. Adjectives also explains the features of the noun.

After loading the database with Phrase collocates, determiners, premodifiers,complements and postmodifiers check the ruleset to find whether the headword confirms the rules. It is classified as Noun Phrase Collocation(NPC).

Association measures

Probabilistic associative measures (PAM) describes how strong the n gram collocates are linked to each other. The basic measure used to measure the bindingness of the collocation is the frequency. Since frequency significance results in large number of false positives, statistical scores are used to measure the association between the words.

Let us assume that TW be the total number of words in extracted web document. A bigram collocate is made of two words. Let the first word and second word be represented by R_word and F_word respectively. The frequency of occurrence of the first word in the bi-gram is $Freq_{fw}$. The first word may also be called as the root word. The second word is termed as collocate and the frequency of occurrence of the same is represented by $Freq_{sw.}$ The number of pairs of First word and the second word occurring together is given by $Freq_{fw}$,$Freq_{sw.}$ Contigency table can be designed with the above data for analyzing and recording the relationship between the frequency of the first word and the second. The size of the context window is assumed to be made of 8 words.

The web page https://en.wikipedia.org/wiki/World_war when analysed with the analyser produces the Ngrams ranked by the frequency. Total number of tokens in this page is 2127 and the types is 1756. In this case let "World" be the first word and "War" be the second word. Figure 1 presents the list of first and second word containing WorldWar or World or War.

The sample noun phrases with high frequency extracted from the web pages are given in Table 5.

Table 5: List of Bi-gram word frequency in the phrase World War in Wikipedia

Bi-gram word	count	Frequency
World War	38	1.598
First World	10	0.470
Second World	10	0.470
World wars	7	0.327
Third world	5	0.235
world war	5	0.235
"world war"	2	0.094

Chi-square test

This is a statistical test to compare the expected frequencies with the observed frequencies and whether they differ from one another. Null hypothesis states that two variables are statistically dependent whereas alternative hypothesis suggest that the two variables are statistically independent of each other.

The formula for Pearson's Chi-Square test as given by

$$x^2 = \sum_{i,j} \frac{(0_{i,j} - E_{i,j})^2}{E_{i,j}} \ldots (1)$$

Where

$O_{i,j}$ is the observed Frequency
$E_{i,j}$ is the Expected Frequency
i represents the row index and
j represents the column index

The Observed, Expected frequencies are given in the contingency table Table 6.

Table 6: Contigency table for the Phrase Collocation "World war"

	$Freq_{fw,sw}$	Not $Freq_{fw}$	Total
$Freq_{sw}$	45	25	70
Not $Freq_{sw}$	7	2050	2057
Total	52	2075	2127

The values given in the contingency table is observed frequencies. Expected frequencies are calculated with row and column totals to find whether the word occurs independently of each other. Chi-square test results are more accurate than other statistical tests because it does not totally depend on the Observed frequencies but on larger data. We analyse the results such as when the Chi-square value is equal or greater than the critical value then we assume that the the probability of null hypothesis is very small and the null hypothesis is rejected. Otherwise null hypothesis is accepted which means that there is no significant difference between the variables.

Log likelihood

Pearson's coefficient computes ranking on the distance between the observed frequencies and the expected frequencies in the contingency table whereas Log-likelyhood computes ranking based on the sample distributions. This ratio was first proposed by Wilks[16] and later by Dunning's[17] is deemed to be very complex statistical measure when compared to Chi-square and t-test.

Dunning's refined model explains the hypothesis that noun words in the n-gram from the sample is distributed independent of each other or otherwise it may occur together due to the chance. G^2 log-likelihood ratio[1] measures the strength of association between the noun phrase collocations. In the bi-gram model the probability that collocates occur together is

P(R_word,F_word) = P(R_word) * P(F_word)

When the value is more higher, the F_word is the collocate of R_word.

For a tri-gram the G^2 ratio is
M1:P(R_wordF_wordS_word)/(P(R_word)*P(F_word)*P(S_word))
M2: P(R_wordF_wordS_word)/(P(R_word F_word)*P(S_word))
M3: P(R_wordF_wordS_word)/(P(R_word)*P(F_word S_word))
M4: P(R_wordF_wordS_word)/(P(R_word S_word)*P(F_word))

Four expected value models are built for tri-grams. In the first model (M1) the three words are dependent on each other and in the second model (M2) first and second word are dependent and third word is independent.In the third model (M3) first word is independent and second and third words are dependent on each other.

The resultant G^2 score is assumed as the degree of divergence between the observed and the expected values. When the G^2 score is high it means it less likely that sample is independent of each other and when G^2 score is Zero it can be interpreted as the good fit with no divergence between the observed and the expected values. For a tri-gram and 4-gram noun collocate the dimension of the Expected frequency grows in size. A threshold cutoff is established which decided the point at which all the N-gram Noun phrases above the threshold value are Noun phrase collocation and the values which are below the threshold is not a collocation.

Pointwise Mutual Information
This is a method to measure the dependency of the variables. In the bi-gram, the probability of finding the root word and the first word together to the probability that the two words are independent to each other are compared. The P(R_word,F_word) is always higher than the P(R_word)*P(F_word) which is given in equation 2.When the noun phrase collocates occur together and the frequency is lower, then Mutual Information will also be higher.
The formula is given by

$$\text{I(R\_word,F\_word)}=\log_2 \frac{P(R\_word\,,F\_word\,)}{P(R\_word\,)*P(F\_word)} \quad (2)$$

Using the Probabilistic associative measures ,the degree of dependency between the words in the collocation phrase is found and the results are discussed in the evaluation section.

Collocation by Machine Learning algorithms
Supervised and unsupervised algorithms are used for classification and clustering the collocations. Naïve Bayes classifier [18], C4.5 decision tree classifier[19] and K means clustering algorithm [20] are used to segregate the Noun phrase collocates and Non-noun phrase collocates. C4.5 classifier uses the concept of information gain for

splitting the classifier data. Both the continuous and the discrete values can be handled by the C4.5 classifier.
When the k centroid is predicted correctly and on having the cluster kept small the square error function is minimized.

Feature Vector for Noun phrase collocation
Using the Features, supervised learning algorithms are trained to classify the bi-gram data in the dataset as Noun phrase collocates.

1. If the POS tags are NN_NN, NN_NNP or NNP_NNP assign the weight 1. The weight is assigned a positive value because the head word can be noun or proper noun. If the first word and second word is both not noun or proper noun assign the weight 0. If the first word or the second word is NN or NNP then assign the weight 0.5.
2. If the bigram word contains all numbers then assign weight 0. If one word is number then assign the word with value 0.5. If bi-gram word contains integers in both words assign 0 or otherwise assign 1 for floating point numbers.
3. If the bigram contains special characters or symbols as in chi-square test, then assign the word with weight 0. For Partial noun assign weight 0.5.
4. If the two words in the bigram are less meaningful words such as Determiners or words in the stop word list then assign 0. If one word is stop word assign 0.5 or otherwise assign weight 1.
5. If both the words are in uppercase and if both the words are in lower case Then assign weight 0. If one word in upper case then assign weight 0.5. If the first letter of the words is in Uppercase or lowercase together then assign weight 1.
6. Using the results of the probabilistic statistical tests of chi-square, PMI and log likelihood , if the words are fully dependent on each other assign weight 1. If the degree of dependence of words is less than the threshold then assign weight 0 or otherwise assign weight 0.5.
   The feature set gets increased in size as the n-gram increases.

DATASET
Three different datasets was extracted from multiple domains. Dataset1(DS -1) contains randomly selected 500 sentences from Wall Street Journal(WSJ). Dataset 2 (DS-2) contains 1000 sentences extracted from web documents of different domains. Dataset 3 contains 500 sentences which was extracted from the Reuters news corpus.
The dataset are comprised of Noun+Noun collocations, Verb+Noun collocations, Adjective +Noun and Verb+preposition collocations, idioms and proverbs. The dataset is designed in such a manner to make it a gold standard data for training and testing the machine learning algorithms. 60% of data will be used as training set and 40% as the test set.

EVALUATION
The methodologies used in the Collocation extraction, Filtering of Noun Phrase collocations, Association statistical measures, Classification and Clustering

techniques are tested to find the precision, recall, F-measure. Smadja's sliding window size of 6 provided with good results in her research. In this research work, the size of the context window span is expanded to 7. The bi-gram phrase collocation accuracy is tabulated in Table 7. Phrase collocation accuracy decreases with increase in the window span size. The threshold for accuracy is 92%. Unigram achieves higher accuracy than bi-grams. 7 window span decreases in accuracy to 89.14%. Results are recorded in table 7.

Table 7: Sliding window phrase collocation accuracy

CONTEXT WINDOW SPAN	PHRASE COLLOCATION ACCURACY
1	93.24%
2	90.68%
3	89.42%
4	87.39%
5	86.88%
6	85.23%
7	89.14%

Google search has provided the highest recall of 0.93 due to its indexing nature. Wikipedia measures a high precision and recall due to its open architecture. WordNet and BNC are the worst performers because they were mere dictionaries. With idioms collocation phrase identification Wordnet and BNC corpus was the best. The results by Gazetteer is extremely poor and can be rejected. Only Geographic location Proper noun collocates can be extracted with Gazetteer. The results are given in Table 8.

Table 8: Collocation Extraction algorithm evaluation using precision and recall

CORPUS	PRECISION	RECALL	F-Measure
Word Net	0.87	0.23	0.36
Wikipedia	0.95	0.70	0.80
Google	0.96	0.93	0.94
BNC	0.85	0.20	0.32
Gazetteer	0.90	0.07	0.12

The results of the collocation phrase accuracy on three different datasets provide various results due to the composition of the dataset. Dataset 1 and Dataset 3 produces good results of precion and recall when compared to the documents extracted from the web. The recall on the Dataset 2 is very high. The precision and recall on datasets using the Collocation algorithm without the Parts of Speech constraint is recorded in table 9.

Table 9: Collocation phrase accuracy without using filtering algorithm

Dataset	Without POS restriction	
	Precision	Recall
DS-1	0.96	0.75
DS-2	0.94	0.76
DS-3	0.98	0.73

The proposed collocation phrase extraction and Noun Phrase filtering algorithm produces a combined effort to give high precision and recall on the datasets. The results prove that there is negligible difference between the Precision and recall values in Table 9 and Table 10

Table 10: Collocation phrase accuracy with using filtering algorithm

Dataset	With Noun Phrase POS restriction	
	Precision	Recall
DS-1	0.97	0.76
DS-2	0.95	0.76
DS-3	0.98	0.74

The final ranking of the noun phrase collocates are studied for the frequency and efficiency in retrieving information using the Noun phrase collocate as Query Phrase. In table 11 the top 1% of ranked candidates have high precision of 0.80 but the lowest recall rate of 0.07. In the 100% segment the precision stands at the lowest point whereas the recall is 1 which is absolute high.

Table 11: Collocation phrase accuracy with filters

Ranking of Noun Phrase Collocates	With POS restrictions		
	Precision	Recall	F-measure
1%	0.80	0.07	0.19
5%	0.50	0.25	0.33
10%	0.48	0.32	0.37
20%	0.30	0.53	0.37
50%	0.22	0.72	0.32
70%	0.18	0.86	0.28
100%	0.15	1	0.15

Among the statistical measures log likelihood emerges with a recall of 97.23% for the List size 40. But the precision is very poor among the bi-gram collocation phrase. The

overall precision of all the AM tools comes under the threshold due to the accuracy of classification. Chi-square test provides the second best solution. The results are tabulated in Table 12.

Table 12: Measuring Collocation phrase accuracy using AM

Statistical measure	List Size	Recall in %	Precision in %
CHI SQUARE TEST	30	76.20	5.8
	60	80.23	3.2
LOG LIKELY HOOD	30	83.40	6.9
	60	97.23	4.0
MUTUAL INFORMATION	30	74.78	5.6
	60	76.93	3.1

The proposed Noun Phrase collocation extraction and POS filtering algorithm together measures a high F1-score. C4.5 decision tree classification with average F1 score of 73.34% proves to classify better than other machine learning algorithms.

Table 13. Comparative F1-score of algorithms on Datasets

Dataset	F1-Score			
	Collocation extraction + filtering algorithm	Naïve Bayes	C 4.5	K-means
DS-1	82.3%	53.56%	75.52%	65.70%
DS-2	78.9%	55.67%	70.96%	68.00%
DS-3	81.3%	56..43%	73.56	67.89%

CONCLUSION

We have evaluated several algorithms ranging from Phrase Collocation algorithm, Filtering algorithm, Statistical measures such as Chi-square test, PMI and Log likelihood. Supervised machine learning algorithm classifiers such as C4.5 decision tree and Naïve Bayes were used for learning Noun collocation phrases and K-means algorithm defined the clusters to classify the dataset into valid Noun Phrase collocates and invalid noun phrase collocates using the feature set.

Adding more statistical measures for testing and classification can extend further work on this title. This algorithm should not only be restricted to be used in English language corpus but also in other languages.

REFERENCES

[1] https://en.wikipedia.org/wiki/Collocation

[2] Manning and H. Schutze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press.

[3] Harold E. Palmer. *A Grammar of English Words*. Longman, London, UK, 1938.

[4] Robert Robins. *A Short History of Linguistics*. Longman, London, UK, 1967.

[5] John R. Firth. Modes of meanings. In *Papers in Linguistics 1934–1951*, pages 190–215. Oxford University Press, 1951.

[6] S. Ikehara, S. Shirai, and H. Uchino. A statistical method for extracting uninterrupted and interrupted collocations from very large corpora. In *COLING*, pages 574–579, 1996.

[7] S. Shimohata,T. Sugio, and J. Nagata. Retrieving domain-specific collocations by cooccurrences and word order constraints. *Computational Intelligence*, 15:92–100, 1999.

[8] [Godelieve L.M. Berry-Rogghe. The computation of collocations and their relevance in lexicalstudies In *The Computer and Literal Studies*, pages 103–112, Edinburgh, New York, USA,1973. University Press.

[9] Kenneth Church and Patrick Hanks. Word association norms, mutual information and lexicography. *Computational Linguistics*, pages 22–29, 1990.

[10] Sayori Shimohata, Toshiyuki Sugio, and Junji Nagata. Retrieving collocations by co-occurrences and word order constraints. In *Proceedings of the 8th conference on European chapter of theAssociation for Computational Linguistics (EACL)*, pages 476–481, 1997.

[11] Frank A. Smadja. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19:143–177, 1993.

[12] S. T. Gries. 50-something years of work on collocations: what is or should be. *International Journal of Corpus Linguistics*, 18(1):137–166, 2013.

[13] Brigitte Krenn. *The Usual Suspects: Data-Oriented Models for Identification and Representation of Lexical Collocations*. PhD thesis, Saarland University, 2000.

[14] Gnana Chithra .C, Ramaraj. E. Heursitic sentence boundary detection and classification. In Proceedings of the First International Conference on Recent Innovations in Engineering and Technology 2016, published in International Journal of Emerging Technoloies-IJET(online ISSN: 2249-3255).

[15] www.google.com

[16] S. S.Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *TheAnnals of Mathematical Statistics*, 9:60–62, March 1994.

[17] Dunning's [T. Dunning (1993). Accurate methods for the statistics of surprise andcoincidence.

[18] Bayes, T. 1763. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society*, 53:, 370-418.

[19] J.R. Quinlan, C4.5 programs for machine learningMorgan Kaufmann Publishers, livres Google.

[20] B. MacQueen (1967): "Some Methods for classification and Analysis of Multivariate Observations, *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability"*, Berkeley, University of California Press,1:281-2