# Statistics based evaluation of English Multi-Word Expressions

**Rakhi Joon, Archana Singhal**

***Abstract*: The linguistic and statistical information extraction is an important aspect of text processing. The extraction of Multi Word Expression (MWEs) plays a key role in text processing as these are used to find correct meaning of a text phrase. MWEs are the lexical phrases consisting of two or more words conveying some different meaning together other than its constituent words. The linguistics in MWEs extraction is mainly related to the text information including the Part of Speech (POS) tags, grammar rules, related literature, and so on. It is important to extract the correct MWEs for a particular language as there exists variety and veracity in languages. The selection of MWEs are based on the statistical analysis of the MWEs extraction process. In the proposed work, the MWEs extraction is done for English dataset. Along with the existing statistical measures, i.e. Pointwise Mutual Information (PMI), Dice Coefficient (DC) and Modified Dice Coefficient (MDC), the additional measures, Lexical Fixedness (LF), Syntactic Fixedness (SF) and Relevance Measure (RM) are also been evaluated. The results are compared with the other existing approaches applied for English MWEs. The results shows that the proposed measures LF, SF and RM are more significant than existing measures to find the best statistics for the MWEs extraction process. The process model is generic in nature and not adhered to a particular language. It can also be applied for other languages by selecting POS tags for that particular language.***

***Keywords* : *English MWEs, linguistics, statistical measures, text processing.***

## I. INTRODUCTION

Multi-Word Expressions are the lexical phrases consisting of two or more words or the compound word exhibiting some semantic or pragmatic property [21]. The relation between the compounds and MWEs is based on various factors including the formal constructions and the functional categories [5]. The formal constructions mainly deals with the different compound terms like Noun+Noun, Adjective+Noun, Compound Adjectives, Compound verbs, and so on. The functional categories include the standard classification terms of MWEs like Named Entities, Idioms, Abbreviations, proverbs and so on. The classification of MWEs into lexicalized and institutionalized phrases [1] is related to the sematic idiosyncrasy and the statistical idiosyncrasy. The sematic idiosyncratic terms mainly comprised of Adjective-Noun pair, Compound nouns, words with spaces, idioms, verb-particle combinations, light verb constructions and so on, for example 'black book', 'ring-ring', 'chair car', etc. While the statistical idiosyncratic terms include Adjective-Noun pairs, Compound nouns, ordered noun sequences, verb-object pairs and so on for example 'warm up', 'tear off', 'come across', etc. MWEs are most commonly used in spoken languages and also important for written languages. The usage of MWEs in many languages are dependent on many other factors including domain, dataset, language, etc. The MWEs are very important for Natural Language Processing (NLP) because of the usage in information retrieval, ontology building, text alignment and machine translation.

In this paper the main focus is on the formal construction of the MWEs [5], thus the classes of formal constructions are considered for experiments using the movie review dataset and the results are compared with the existing work[1] . Earlier, only PMI, DC and MDC statistical measures were evaluated which are further extended in the proposed approach for LF, SF and Relevance Measure. Based on the results of the proposed measures it is observed that RM is the most significant among the proposed ones, while earlier PMI was considered as the best one.

## II. RELATED WORK

The extraction of MWEs had been discussed by various authors for many languages. The extraction of MWEs can be carried out by following various techniques. The word alignment methods were discussed by the authors in [20]. The statistical methods were discussed by authors in [1] and some new methods were discussed in [29] and [30]. The hybrid methods were explained in [4,11]. Since various types were discussed for extraction but the suitable method for extraction was not suggested. In [2] and [16], the authors used latent semantic analysis for the comparison of MWEs and their component words. The semantics based on the lexical substitution and replacing similar meaning word was used for extraction of MWEs [9], the distributional characteristics of collocations were discussed. In [6], the authors worked on the designing of Indo WordNet which was useful for finding the reduplicated words in Hindi corpus and later extended for Bengali language [7]. In [20], the authors evaluated various features of the idiomatic expressions and automatic word alignment techniques. In [23], the authors proposed a methodological framework for the MWEs extraction, while in [26], the authors discussed about the alignment methods for extracting MWEs from parallel corpus. The generalized collocation extraction was also done in [13]. The noun+verb MWEs [31], noun compounds [14], verb noun [10] and verb-noun pair MWEs [17] were some of the specific types

Revised Manuscript Received on October 15, 2019
* Correspondence Author
**Rakhi Joon**, Department of Computer Science, University of Delhi, Delhi, India, rjoon30@gmail.com
**Archana Singhal**, Department of Computer Science, IP College for Women, University of Delhi, Delhi, India.

discussed by the researchers. In [31], maximum entropy model was used to measure the compositionality of noun+verb MWEs, while the authors in [8] classified the verb-noun MWEs on the basis of idiomatic and literal usage. In [1] and [12], the statistical measures, PMI, DC and MDC were evaluated for the English MWEs. The comparative results were also shown for the best statistical measures for these type of the MWEs.

The authors have evaluated various measures for Hindi dataset [29, 30] and the best statistical measures for various categories have been identified. In the proposed work, the MWEs extraction is done for the English Movie review dataset. The additional measures, LF, SF and RM are evaluated with the existing statistical measures, PMI, DC and MDC. It has been analyzed that the proposed measures, LF, SF and RM are more significant than existing measures to find the best statistics for the English MWEs extraction process.

## III. DATASET

A corpus of 100 document is taken from the movie review dataset used for the experimental purpose written by general users. Five fold cross validation method is used to select the training data and test data. 70 text documents are selected for training and 30 text documents for testing purpose. The training set of 70 documents is divided into seven sets containing 10, 20, 30,…, 70 text files correspondingly. When the value of the statistical measure becomes constant then it is selected as the threshold value. It is observed from the results that after 40 documents the values for all the measures stay almost constant.

## IV. PROPOSED METHODOLOGY

The proposed methodology mainly deals with the evaluation of existing and additional proposed measures for the English dataset. Two main aspects of proposed methodology are the linguistic behavior and the statistical properties of the MWEs in English. In the proposed work 2-grams MWEs are evaluated on the basis of linguistic and statistical properties. The process model is implemented for the multiple threshold method of extracting the MWEs in which each linguistic pattern is evaluated for the different threshold values. The process model is shown in the figure 1 as given below:

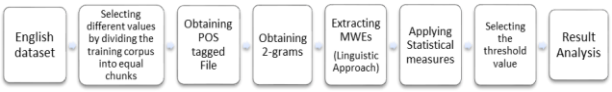


**Fig. 1. Process flow of the proposed approach**

The procedure starts with the analysis of English dataset. The dataset considered is the movie review dataset containing the reviews of different movies, so it is rich in MWEs. The first 100 text files of the dataset are considered, out of which 70 text documents are selected for training purpose. The POS tagging is done for each file and thus tagged files are created using Stanford POS tagger for English dataset. From the POS tagged files 2-grams are extracted and linguistic patterns are applied to extract the MWEs. Then the statistical measures are applied on the obtained list of extracted MWEs. It has been observed that after 40 text files, the values of the existing as well as proposed measures almost stays constant, so 40 is selected as the threshold value for the dataset. The results are compared with the results of existing approaches for accuracy of the basis of best statistics. The proposed methodology is explained in the following subsections:

### A. Selecting Linguistic patterns

The functional categories of the English MWEs are considered for the evaluation purpose and these categories are selected on the basis of existing literature. The categories selected are Adverb+Adjective, Noun+Preposition, Adjective+Preposition, Noun+Verb, Preposition+Noun, Verb+Noun, Noun+Noun, Verb+Preposition, Verb+Adverb, Adjective+Adverb, Adjective+Noun, Noun+Adjective, Compound Noun, Verb+Verb and Verb+Particle. The experiments are done for these 2-grams categories of English MWEs. The description of these categories are shown below in table I.

**Table-I: Description of 2-grams patterns**

| 2-grams pattern | First POS | Second POS | Example |
|---|---|---|---|
| Adj+Adv | Adjective | Adverb | Simply better, alcoholic critically |
| Adj+Noun | Adjective | Noun | Real deal, Japanese girl |
| Adj+Prep | Adjective | Preposition | Aware of, more than |
| Adv+Adj | Adverb | Adjective | How sweet, very strange |
| Noun+Adj | Noun | Adjective | Bad result, fox musical |
| Noun+Noun | Noun | Preposition | Play around |
| Noun+Prep | Noun | Verb | earthquake grimaced, comedy making |
| Noun+Verb | Preposition | Noun | Around Venice, through images |
| Prep+Noun | Verb | Adverb | Fail miserably, badly dubbed |
| Verb+Adv | Verb | Particle | Fill in, cool down |
| Verb+Noun | Verb | Preposition | Move around |
| Verb+Particle | Verb | Verb | Being hunted |

The next section discusses about the existing and the proposed statistical measures.

### B. Statistical measures

Various statistical measures used in proposed work are discussed below in brief.

**Point wise Mutual Information (PMI)**

PMI is the logarithmic ratio of the n-gram probability and the constituent words probabilities. For example for bigram ($w_1$ $w_2$), the PMI score is calculated as:

$$\text{PMI}(w_1, w_2) = \log_2 \frac{P(w_1,w_2)}{P(w_1)*P(w_2)}$$

$P(w_1, w_2)$ is the probability of the bigram MWE which comprised of two words $w_1$ and $w_2$, whereas $P(w_1)$ indicate the probability of an individual word.

**Dice Coefficient (DC)**

The DC is based on the frequency of occurring rather that probability. Like PMI, for bigram the DC is calculated as:

$$DC(w_1, w_2) = \frac{2f(w_1w_2)}{f(w_1) + f(w_2)}$$

**Modified Dice Coefficient**

Similar to DC, the MDC is also based on the frequency of occurrence and thus calculated as:

$$MDC(w_1, w_2) = \frac{2f(w_1w_2)}{f(w_1) * f(w_2)}$$

In the above two measures, the observed frequency of the bigram MWEs $f(w_1w_2)$, comprised of the words $w_1$ and $w_2$. The frequencies, $f(w_1)$ and $f(w_2)$ are the observed frequencies of $w_1$ and $w_2$.

**Lexical Fixedness (LF)**

The LF is calculated on the basis of Association measure and lexical variants.

$$fixedness_{lex}(v, n) = \frac{PMI(v,n) - \overline{PMI}}{std}$$

The mean is $\overline{PMI}$ and the standard deviation is *std*, calculated over the PMI.

**Syntactic Fixedness (SF)**

The SF is calculated by considering target text and the Verb-Object pair.

$$Fixedness_{syn}(v, n) = \text{D}(\text{P}(pt|\text{v,n})\|\text{P}(pt))$$

$$= \sum_{pt_k \epsilon P} P(pt_k|v, n)\ log \frac{P(pt_k|v,n)}{P(pt_k)}$$

The pattern set P, comprised of different POS tags. The syntactic properties of the target terms are expressed by $P(pt_k|v, n)$.

**Relevance Measure**

Relevance is the measure of how relevant a given MWEs is to the classification, the formula proposed was as follows:

$$\text{Relevancy} = \frac{\sum_{i=1}^{n} c_i x_i}{\sum_{i=1}^{n} x_i}$$

Here $c_i$ is the correlation coefficient and $x_i$ is the input vector. In proposed approach relevance is measured in terms of frequency, in which the frequency of a particular Hindi MWEs is measured with its corresponding individual words. Next section discusses about the experimental results.

## V. EXPERIMENTAL RESULTS

The calculation of results are based on the measures explained in the earlier section for the Multiple Threshold method and are shown in table II. The values shown in the bold face are the results obtained for the proposed measures, while the rest are the results for the existing measures as discussed by the authors in [1].

The best statistics is also updated on the basis of the new measures evaluated. It can be analyzed from the results that LF, SF and RM are more significant than the existing measures. The graphical representation for the proposed measures are also shown using the graphs in the figures 2-4. The lexical fixedness and relevance measure are more relevant for the noun-noun pair, while the syntactic fixedness is more relevant for verb-verb pair among the top three most significant types of MWEs selected from the table II. It can be observed from the results that the proposed measures are significant among all other categories of statistical measures used for MWEs extraction.

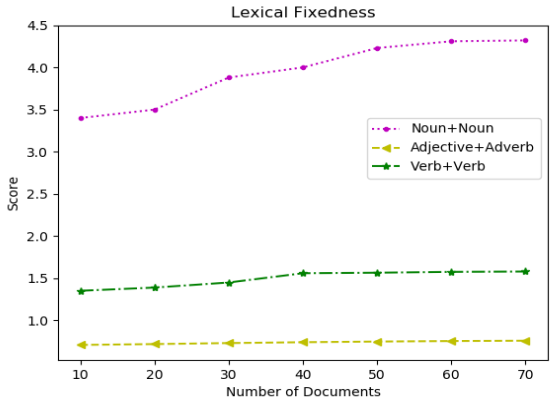


**Fig. 2. Lexical Fixedness measure for most significant MWEs**

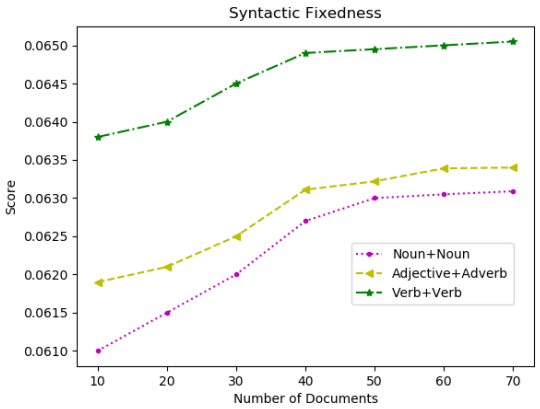


**Fig. 3. Syntactic Fixedness measure for most significant MWEs**

**Table - II: Result evaluation and analysis**

| 2-grams | Recall | Precision | F-Measure | PMI | DC | MDC | Previous Best Score | **Lexical Fixedness** | **Syntactic Fixedness** | **Relevance Measure (RM)** | **Updated Best statistical Measure** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| All Multiwords | 0.601 | 0.337 | 0.432 | 15.1558 | 0.2692 | 0.1119 | DC | **-0.568** | **0.033** | **5.576** | **DC** |
| Adv+Adj | 0.5 | 0.017 | 0.033 | 19.6411 | 0.2857 | 0.3333 | MDC | **0.531** | **0.0317** | **6.519** | **RM** |
| Noun+Prep | 0.667 | 0.043 | 0.081 | 13.3792 | 0.0299 | 0.0177 | PMI | **-1.673** | **0.02217** | **2.25** | **PMI** |
| Adj+Prep | 1 | 0.07 | 0.131 | 11.4524 | 0.0458 | 0.0007 | PMI | **-1.5** | **0.0645** | **4.236** | **SF** |
| Noun+Verb | 0.5 | 0.1 | 0.167 | 19.4372 | 1 | 0.4 | PMI | **0.551** | **0.054** | **4.258** | **LF** |
| Prep+Noun | 0.429 | 0.107 | 0.171 | 15.3095 | 0.0952 | 0.037 | DC | **-0.57** | **0.0629** | **5.405** | **SF** |
| Verb+Noun | 0.321 | 0.173 | 0.225 | 14.327 | 0.3333 | 0.1538 | DC | **-0.75** | **0.05534** | **6.504** | **RM** |
| Noun+Noun | 0.231 | 0.429 | 0.3 | 33.5329 | 0.4286 | 0.375 | MDC | **4** | **0.0627** | **10.513** | **RM, LF** |
| Verb+Prep | 0.308 | 0.381 | 0.341 | 15.1558 | 0.0435 | 0.025 | PMI | **-0.568** | **0.0624** | **10.468** | **RM** |
| Verb+Adv | 0.875 | 0.241 | 0.378 | 14.4428 | 0.0385 | 0.0035 | DC | **-0.761** | **0.0424** | **4.482** | **DC** |
| Adj+Adv | 1 | 0.25 | 0.4 | 20.3558 | 0.0769 | 0.0357 | MDC, DC | **0.742** | **0.06311** | **8.189** | **RM,LF,SF** |
| Adj+Noun | 0.543 | 0.452 | 0.493 | 18.4474 | 0.4848 | 0.6667 | PMI | **0.25** | **0.00527** | **5.742** | **PMI** |
| Noun+Adj | 1 | 0.357 | 0.526 | 19.9991 | 0.3333 | 0.0952 | DC | **0.542** | **0.05472** | **5.459** | **LF** |
| Compound Noun | 0.9489 | 0.65 | 0.769 | 14.727 | 0.1111 | 0.0294 | DC | **-0.772** | **0.00824** | **2.536** | **DC** |
| Verb+Verb | 1 | 0.667 | 0.8 | 23.2539 | 0.4 | 0.5 | PMI | **1.56** | **0.0649** | **9.625** | **SF, RM** |
| Verb+Particle | 1 | 0.889 | 0.941 | 8.9896 | 0.0023 | 0.0001 | All | **-2.25** | **0.00423** | **2.085** | **PMI , DC, MDC** |

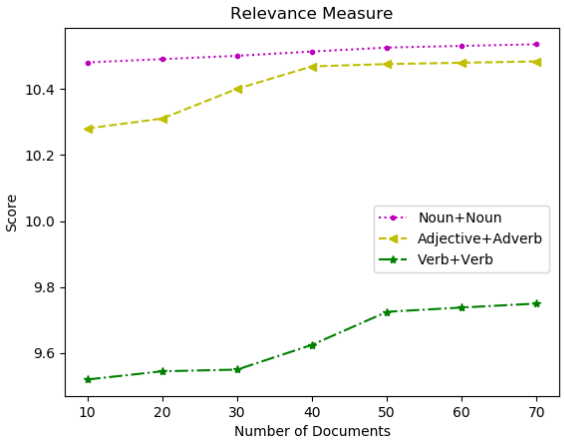


**Fig. 4. Relevance Measure for most significant MWEs**

## VI. CONCLUSION AND DISCUSSION

The identification and evaluation of MWEs is a tedious task in text processing. However, these are very important for many NLP applications so need to be evaluated properly. The proposed methodology mainly focuses on the usage of the linguistic and statistical information of text for extraction and analysis of MWEs. The linguistic patterns are selected for different categories of 2-grams MWEs and each category is evaluated for different statistical measures. The results shows that Noun+Noun, Adjective+Adverb and Verb+Verb are the top three most significant categories of MWEs for English dataset. For other categories also the approach is satisfactory as depicted from the results. The main focus of the proposed approach is to improve the existing classification of MWEs by adding the new measures. It has been observed that for many categories of MWEs, the proposed measures are best measures.

## REFERENCES

1. S. Agrawal, R. Sanyal, and S. Sanyal (2014). Statistics and linguistic rules in multiword extraction: a comparative analysis. *Int. J. Reason. Intell. Syst.*, vol. 6(1), 2014, pp. 59–70.
2. T. Baldwin, "A resource for evaluating the deep lexical acquisition of english verb-particle constructions", *in Proceedings of the LREC Workshop towards a Shared Task for Multiword Expressions (MWE 2008)*, Marrakech, Morocco, 2008, p.1–2,.
3. T. Baldwin, C. Bannard, T. Tanaka and D. Widdows, "An empirical model of multiword expressions decomposability", *in Proceedings of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, Sapporo, Japan, 2003, p.89–96.
4. S. Boulaknadel, B. Daille and D. Aboutajdine, "A multi-word term extraction program for arabic language", *in Proceedings of 6th International Conference on Language Resources and Evaluation (LREC 2008),*

Marrakech, Morocco, 2008, p.1485–1488.
5. L. Bauer, "Compounds and multi-word expressions in English." *Complex Lexical Units: Compounds and Multi-Word Expressions* vol. 9, 2019, pp. 45.
6. D. Chakrabarti, D.K. Narayan, P. Pandey and P. Bhattacharyya, "Experiences in building the indo WordNet – a WordNet for Hindi", *in Proceedings of International Conference on Global WordNet (GWC 02), Mysore, India*, 2002.
7. T. Chakraborty and S. Bandyopadhyay, "Identification of reduplication in bengali corpus and their semantics analysis: a rule based approach", *in Proceedings of the Multiword Expressions: From Theory to Applications (MWE2010)*, Beijing, China, 2010, p.73–76.
8. P. Cook, A. Fazly and S. Stevenson, "The vnc-tokens dataset", *in Proceedings of the LREC Workshop towards a Shared Task for Multiword Expressions (MWE 2008)*, Marrakech, Morocco, 2008, p.19–22.
9. T.V.d. Cruys and B.V. Moir´on, "Semantics-based multiword expression extraction", *in Proceedings of the ACL-2007 Workshop on A Broader Perspective on Multiword Expressions*, Prague, Czech Republic, 2007, p.25–32.
10. M. Diab and P. Bhutada, "Verb noun construction mwe token classification", *in Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation, Applications*, Suntec, Singapore, 2009, p.17–22.
11. J. Duan, M. Zhang, L. Tong, and F. Guo, "A hybrid approach to improve bilingual multiword expression extraction", *in Proceedings of 13th Pacific-Asia Conference on Knowledge Discovery and Data (PAKDD 2009)*, Bangkok, Thailand, 2009, p.541–547,.
12. S. Evert and B. Krenn, "Using small random samples for the manual evaluation of statistical association measures", *Computer Speech and Language*, vol. 19(4), 2005, pp.450–466.
13. A. and S. Fazly Suzanne, "Distinguishing Subtypes of Multiword Expressions Using Linguistically-motivated Statistical Measures," *Proc. Work. a Broader Perspect. Multiword Expressions*, June, 2007, pp. 9–16.
14. R. Girju, D. Moldovan, M. Tatu and D. Antohe, "On the semantics of noun compounds", *Computer Speech and Language*, vol. 19(4), 2005, pp.479–496.
15. J. S. Justeson and S.M. Katz, "Technical terminology: some linguistic properties and an algorithm for identification in text", *Natural Language Engineering*, vol. 1(1), 1995, pp.9–27.
16. G. Katz and E. Giesbrecht, "Automatic identification of non-compositional multi-word expressions using latent semantic analysis," *Proc. Work. Multiword Expressions Identifying Exploit. Underlying Prop.*, July, 2006, pp. 12–19.
17. M. King and P. Cook, "Leveraging distributed representations and lexico-syntactic fixedness for token-level prediction of the idiomaticity of English verb-noun combinations", 2009, pp. 345–350.
18. A. Kunchukuttan, O.P. Damani, "A System for Compound Noun Multiword Expression Extraction for Hindi", *In: Proceedings of ICON-2008: 6th International Conference on Natural Language Processing*, Pune, India, 2008, p.20-29.
19. P. Lambert and N. Castell, "Alignment of parallel corpora exploiting asymmetrically aligned phrases", *in Proceedings of the LREC 2004 Workshop on the Amazing Utility of Parallel and Comparable Corpora*, 2004, p.26–29.
20. B.V. Moir´on and J. Tiedemann, "Identifying idiomatic expressions using automatic word alignment", *in Proceedings of the EACL-2006 Workshop on Multiword Expressions in a Multilingual Context*, Trento, Italy, 2006, p.33–40.
21. R. Moon and J. R. Taylor, "Multi-word items", *The Oxford Handbook of the Word*, 2015, pp.120-140.
22. P. Nakov, "Paraphrasing verbs for noun compound interpretation", *in Proceedings of the LREC Workshop towards a Shared Task for Multiword Expressions (MWE 2008)*, Marrakech, Morocco, 2008, p.46–49.
23. C. Ramisch, "A generic framework for multiword expressions treatment: from acquisition to applications", in *Proceedings of ACL 2012 Student Research Workshop*, Jeju Island, Korea, 2012, p.61–66.
24. I.A. Sag, T. Baldwin, F. Bond, et al. "Multiword expressions: A pain in the neck for NLP". *In: Proceedings of Third International Conference on Computational Linguistics and Intelligent Text Processing: CICLing-2002*, Springer, Berlin, Heidelberg, 2002, p. 1-15.
25. V. Seretan, "A collocation-driven approach to text summarization", *in Proceedings of Traitement Automatique des Langues Naturelles (TALN 2011)*, Montpellier, France, 2011, p.9–14.
26. Y. Tsvetkov and S. Wintner, "Extraction of multiword expressions from small parallel corpora", *Natural Language Engineering*, vol. 18(4), 2012, pp.549–573.
27. V. Sriram, P. Agrawal, A.K. Joshi, "Relative Compositionality of Noun Verb Multi-word Expressions in Hindi", *In: Proceedings of 5th International Conference on Natural Language Processing (ICON)*, Kanpur, India, 2005.
28. R. Joon and A. Singhal, "A System for Compound Adverbs MWEs extraction in Hindi." *In Eighth International Conference on Contemporary Computing (IC3)*, Noida, India, 2015, p. 336-341.
29. R. Joon and A. Singhal, "Role of Lexical and Syntactic Fixedness in Acquisition of Hindi MWEs", *in proceedings of third International Conference on Advances in Computing and Data Sciences*, Ghaziabad, India, 2019, p. 155-163,
30. R. Joon and A. Singhal, "A comparative analysis of Hindi Multi Word Expressions using relevancy measure-RMMWE." *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 8(8), pp. 3436-3445.
31. S. Venkatapathy and A. K. Joshi, "Relative compositionality of multi-word expressions: A study of Verb-Noun (V-N) collocations", *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 3651 LNAI(Ijcnlp)*, 2005, pp. 553–564.

## AUTHORS PROFILE

**Rakhi Joon** is pursuing her Ph.D. in Computer Science from University of Delhi, New Delhi. She did her M.Tech. in Computer Science & Engineering from GJU S&T, Hisar, Haryana and B.Tech. in Information Technology from MDU, Rohtak, Haryana. Her research areas include Natural Language Processing, Wireless Networks.

**Dr. Archana Singhal** is working as an Associate Professor, Department of Computer Science, Indraprastha College for Women, University of Delhi, New Delhi. Her research areas include Natural Language Processing, Semantic Web, Multi-agent Systems, Information Retrieval and Ontologies, Secure Software Systems and Social Networks. She has many publications to her credit in reputed journals and International conferences.