A Comparative Analysis of Hindi Multi Word Expressions using Relevance Measure-RMMWE

Rakhi Joon, Archana Singhal

Abstract*: Text processing is a very complex and tedious task because of various types of ambiguities present in the text. There are various methods suggested by the researchers for text processing, which mainly include generic procedures of word extraction and phrase extraction. Word Extraction mainly deals with extraction of meaningful words from the text, while phrase extraction is the process of extracting relevant phrases. Multiword Expressions (MWEs) extraction is the phrase extraction procedure where key phrases are used as an alias to Multiword lexemes. MWEs are made up from two or more words conveying different meaning as compared to the meaning of the individual words. The measures used for extraction and analysis of MWEs, mainly include the baseline and statistical measures. In baseline measures, Precision, Recall and F-Measure are considered while, in statistical measures, Point wise Mutual Information (PMI), Dice Coefficient (DC), and Modified Dice Coefficient (MDC) measures are considered. In the proposed work one additional measure in statistical category i.e. Relevance Measure (RM), is proposed along with the existing ones. Relevance Measure is evaluated based on the frequency of occurrence of MWEs in Hindi Text. The dataset used in this paper for experimental purpose is Hindi Dataset taken from the famous Hindi novel 'Godaan'. An algorithm has also been designed for evaluating the relevance measure. Evaluation of these measures have been done for 2-grams MWEs and n-grams MWEs. The values calculated for each measure for different categories of Hindi MWEs are shown in tabular form and results are analyzed and discussed with the help of different histograms of RM and other measures. The statistical consideration of RM has not been done till now due to which it become difficult to find out which Hindi MWEs type is more relevant. To solve the above issue, RM for Hindi MWEs is explored in the proposed work, and the inclusion is justified by comparing the results with other existing measures.

***Index Terms*: Relevance Measure, Multiword Expressions, Hindi, Keyphrase, NLP, statistical measures.**

I. INTRODUCTION

Multi Word Expression (MWE) is one of the key terms used in NLP and Computational Linguistics (CL) over the years. Since there is no generic definition given for MWE yet, but there are different views from many authors on this important term. The term MWE refers to the combination of two or more independent words which individually convey some meaning, but when used together convey a different meaning. So, based on this concept there are various types of MWEs, which include Compound nouns compound verbs, compound adverbs, noun+verb, named entities, idioms, phrases, collocations and phrasal verbs. Earlier, before 90s, MWEs was used and studied under the phraseological unit only [28], but later in NLP, MWEs have gained a lot of attention. In every language, MWEs are used frequently in many practical applications of NLP and CL. In 2001, the project "Centre for the Study of Language and Information" (CSLI) at Stanford University was a milestone in the area of NLP as well as MWEs. The main focus of the project was to encode different MWEs into their precision grammar [28]. Later various dictionary resources and annotations tools were built for MWEs. And thus the journey of MWEs begins with such a historical background in the field of NLP. For a particular language, MWEs plays an important role as in formation of a sentence. The formation and categorization vary from language to language as the grammar rules are also considered for the formation of Multiwords. MWEs have gained popularity in many languages including English, Chinese, Hindi and many other Indian Languages. In this paper, the language used for MWEs is Hindi, which is morphological rich language like most of other Indian languages and thus provides a good platform for MWEs.

The text processing in a particular language depends on the grammatical constructs along with the linguistic and syntactic properties of the language. The formation of MWEs mainly deals with the combination of two or more meanings of the words rather than the words itself. These words can be any of the grammar constructs like Noun, Pronoun, Verb, Adverb, Adjectives, and so on. POS taggers mainly affect the representation of a particular construct in any language. In this paper since the work is based on Hindi MWEs, so the combinations of Hindi POS tags (https://bitbucket.org/sivareddyg/hindi-part-of-speech-tagger/src/434f9c388404?at=master) are considered for the experiments and results. The MWEs which were formed from the combinations of POS tags mainly includes 2-grams and n-grams. In 2-grams, Adj+Noun (साक्षात देवी - Saakshat Devi), Adj+Prep (बराबर वाला - Brabar vaala), Adv+Adj (बिलकुल गुड़िया-सी - *bilkul gudia si*), Adv+Adv (शायद फिर - *shyad fir*), Noun+Adj (पांचों पोसाक - *pancho poshak*), Noun+Prep (उत्साही मेंबर - *utsaahi member*, बीच-बीच में - *bich bich me*), Noun+Verb (ताल ठोंक - *taal thok*, दिवाला निकालना - *diwala nikalna*, तेल निकालना - *tail nikalna*) Noun+Noun (नीति पसंद - *niti pasand*, दो-चार पैसे - *do chaar paise*), Verb+Adv (चलो फिर - *chalo phir)*,

Revised Manuscript Received on June 15, 2019.

Rakhi Joon, Department of Computer Science, University of Delhi, Delhi, India,

Archana Singhal, Department of Computer Science, IP College for Women, University of Delhi, Delhi, India.

Verb+Particle (देखते ही - *dekhte hi*), Verb+Prep (बोलने वाले - *bolne vale),* Verb+Verb (हार मानना - *haar manana)* are included. In n-grams, Compound Noun (eg. अंगारे की-सी आंखें - *Angare ki si aankeh*), Verb-Noun (eg. काटते देख कर घूंघट - *katthe dekhkar ghunghat),* Noun-Verb (eg. हाथ-पांव ठंडे होना - *hath pav thande hona),* Noun-Noun (हवा की तरह - *hawa ki tarah)* are included. In the proposed work, process of extraction begins with processing text followed by POS tagging and applying the extraction process to extract MWEs in Hindi and then measuring Relevance score. So, the main focus is to find the relevance score of the Multiwords extracted from the dataset. The idea is taken from [10], in which highly ranked different sentences were selected and relevance score of each sentence with respect to the complete document was calculated. In proposed work, highly ranked multiword expressions are selected and the relevance score of each multiword is measured with the complete set of multiword as well as with complete set of words in the corpus. The procedure followed is explained using the proposed algorithm and flowchart. The main significance of the proposed work is based on the relevance measure itself, because earlier no work has been done considering the relevance measure for statistical category in MWEs research. Further the comparison of RM is done with other existing measures and result analysis is done using various histograms.

The paper is organized in the following manner. Section 2 describes the related work in this area and different experiments done by other researchers on MWEs. Section 3 describes the dataset used for the experimental purpose. In section 4, the proposed methodology is explained along with the statistical measures. The experimental results are shown in section 5 and the results are analyzed for accuracy. In the last section conclusion and future scope are given.

II. RELATED WORK

There are many papers concerned with the introduction to MWEs and the extraction process, some of the important ones are considered here. NLTK was suggested and used for Hindi MWEs Extraction and processing [13]. The new strategies to normalize multiword into two different versions of multi word features are suggested by authors in [21]. Decomposition and combination strategies were followed and the effectiveness of strategies was measured using text classification. Support Vector machine was used in both linear and nonlinear kernel and results were compared for both types of kernel as well as for both strategies proposed. The lexical and syntactic configuration of the MWEs were discussed which mainly considered that the original semantics was preserved by component words but sometimes extra semantics was encoded by MWEs [25]. Further base types of MWEs and the linguistic properties of MWEs were explained.

The frequency information of MWEs, mainly 4-grams was collected and used as a measurement for the language processing. In this study, larger units were used for parsing and models of processing where maximum work has been done for smaller units only before this [3]. Taking reference from this, the concept of n-grams is implemented in the proposed work which also cover 4-grams. The frequency effects were considered by the authors which are mainly related to the relevancy of the document. The author in [17] explained the detailed overview of various types of MWE encountered in Hindi and a stepwise mining of MWEs in Hindi with their machine translation perspectives.

There are many different techniques for MWEs extraction and statistical measures suggested by various researchers, but Relevance Measure for MWEs has not been considered so far. Some of the authors suggested some tasks related to relevancy are as follows:

Experiments were carried out by the researchers to show the effect of earlier acquired phrases and importance of multiword phrases in developing a language system [2]. Multi-Word Relevant Expressions (REs) and the extraction process of REs were based on the ParLocalMax algorithm, which used large number of machines than the LocalMax algorithm [9]. In [10], two different methods of text summarization were defined in which the very first method i.e. relevance measure which is based on standard Information Retrieval methods to rank the relevance score of the sentences is used as an idea in the proposed work. In the procedure for relevance measure proposed by the authors, highly ranked sentences which are different from each other were selected while in proposed work highly ranked multiword expressions are selected and the relevance score of each multiword is measured with the complete set of multiword as well as with complete set of words in the document/corpus. Non compositionality [14], association measures [15] which are further combined by standard statistical classification methods and further modified to provide scores for ranking are also considered. Relevancy and its measure were described [22] and for the measurements and calculations, correlations and mutual information are used. Correlation is the measurement useful for continuous data rather than nominal one whereas mutual information results into a discrete measure of the mutually shared information between the two dimensions in a particular question. The relevance measure should be maximized for better performance.

There are many types of Hindi MWEs associated with the research. Many categories of MWEs in English and other languages are considered, which itself made a common categorization. In our previous research many categories of Hindi MWEs [11] are considered and some new categories were also added to it [12]. Other associated categories were discussed by many authors: In [23], the authors presented a system for compound noun MWEs extraction and the method adopted for extraction use many statistical co-occurrence measures to satisfy statistical idiosyncrasy of MWEs. The Noun+Verb expressions in Hindi MWEs were considered in [24] and their relative compositionality was measured using Maximum Entropy Model (MaXEnt). N+V expressions in Hindi were mapped to V-N expressions in English to compute some features. In [26], the authors presented an empirical study of integrating n-grams and multi-word terms into topic models in which the most suitable n-grams and multi-word terms were incorporated. From machine translation viewpoint, there are various types of MWEs in Hindi and these types were not given proper attention in research for e.g. 'vaalaa' construct, replication, doublets and so on.

Many of the types are frequently used in daily life but are not given proper place in formal textual corpus [4]. The statistical measures are also considered for the extraction process and are discussed by many authors. In [27], the hypothesis explained was that MWEs can be detected independently by the unique statistical properties of their constituent words, regardless of their types and the performance measure and comparison of three statistical measures (Mutual Information, $\chi 2$ and Permutation Entropy).

In the proposed work the focus is on Hindi text processing and MWEs extraction analysis using the baseline and statistical measures. The authors have proposed the hybrid methods for extracting multiword expressions based on linguistic and statistical information [1]. Various n-grams are extracted and the statistical measures are applied to classify these as multiword expressions. In [6], the authors analysed the important properties relating to the semantic idiosyncrasy of MWEs. The relation between properties and four types of verb+noun combinations were also elaborated from less to more semantically idiosyncratic. A statistical corpus based measure was also developed for quantification of the properties. In [7], an algorithm (LocalMaxs) for extracting MWEs, by using statistical measures was explained and in [8], the authors proposed a domain independent method for extraction of MWEs from a machine readable corpus. The text representations including indexing and weighing tasks were also discussed [20]. The importance of MWEs in various languages as well as the experiments carried out for calculating the performance of designed systems or algorithms were discussed which proved to be useful for carrying out the proposed task.

III. DATASET

Due to large Corpus, number of text files from the novel are collected and used for the experiments. This dataset is very useful as it is based on real life scenarios. It contains number of Multiwords in form of compound nouns, named entities, compound verbs, idioms, phrases, and so on. The dataset contains 36 text files, which are processed to get the final dataset. The dataset contains 173691 total words and out of these 15448 are unique words. The data used for training contains 130159 words and 13050 unique words, while the data used for testing contains 43533 words and 6802 unique words. The experiments are done for 2-grams and n-grams. The n-grams obtained from the raw text are then divided into different combinations of POS tags and then filtered statistically to obtain the MWEs. The selection of a particular POS tag combination is decided by the linguistic rules for that particular grammar. Table I shows the different patterns chosen for the Hindi MWEs and the total number of 2-grams existing in the dataset with example, while Table II shows the same information for n-grams.

Two types of measures are used to calculate the performance of designed algorithm, i.e. baseline measures and the statistical measures. The baseline measures mainly cover precision, recall and F-measure, while the statistical measures covers PMI, DC, MDC. One new measure i.e. Relevance Measure has been added to the statistical list of measures for accuracy. Different categories of MWEs are extracted from the corpus and Relevance Score is calculated for each category. In the corpus, total number of 2-grams are 35361, out of which 3381 are Multiwords and 2117 are unique Multiwords, while total number of n-grams are 60024, out of which 9498 are Multiwords and 3814 are unique Multiwords.

For using statistical techniques on any corpus the size of corpus in terms of number of words, unique words, categories etc., should be ensured prior to the experiments. In this paper Hindi MWEs are evaluated over various statistical techniques, in which all are existing techniques except RM, which is introduced as a new statistical measure proposed for the Hindi MWEs.

Table I: Detailed description of 2-grams with example

MWE Pattern	POS first tag	POS second tag	Total 2-grams	Example
JJ+NN	Adjective	Noun	2876	*चिरस्थायी जीर्णावस्था (chirnasthai jirnavstha), अच्छी तरह (Aachi tarah)*
JJ+PSP	Adjective	Preposition	61	*बराबर वाला (barabar vaala), मुश्किल से (mushkil se)*
RB+JJ	Adverb	Adjective	54	*बिलकुल अलग (bilkul alag), सदैव बंद (saidav bandh)*
RB+RB	Adverb	Adverb	8	*फिर कभी (phir kabhi), फिर एकाएक (phir ekaaek)*
NN+JJ	Noun	Adjective	1043	*विधवा बहू (vidhwa bahu), चार-पांच सेर (chaar panch ser)*
NN+PSP	Noun	Preposition	15119	*उत्साही मेंबर (uthsahi member), बीच-बीच में (bich bich me)*

NN+VM	Noun	Verb	11975	*ताल ठोंक (taal thok), दिवाला निकालना (diwala nikalna)*
NN+NN	Noun	Noun	2108	*नीति पसंद (niti pasand), दो-चार पैसे (do chaar paise)*
VM+RB	Verb	Adverb	12	*भागा जरूर (bhaga jarur), चलो फिर (chalo phir)*
VM+RP	Verb	Particle	470	*देखते ही (dekhte hi), सूंघा तक (sungha tak)*
VM+PSP	Verb	Preposition	1162	*देने वाला (dene vala), मिल के (mil ke)*
VM+VM	Verb	Verb	473	*मिलते-जुलते रहना (miltr julte rehna), मार डालती (maar dalti)*

Table II: Detailed description of n-grams with example

MWE Pattern	POS Description	Total n-grams	Example
Compound NN	Includes series of Nouns existing in the sentence	54	*बाधा गले पड़ी (baadha gale padi), अंगारे की-सी आंखें (aangare ki si aankeh)*
VM-NN	Includes sequence of Verb as first word and Noun as last verb in the sentence	17309	*भरे हुए माथे (bhare hue mathe), आए दिन संग्राम (aae din sangraam), धूल मत झोंको (dhool mat jhoko), पिचके हुए चेहरे (pichke hue chehre), पूछो तो कोई जवाब (pucho toh koi jwab nhi)*
NN-VM	Includes sequence of Noun as first word and Verb as last word in the sentence	19979	*हाथ-पांव ठंडे होना (hath pav thnde hona), पांवों-तले अपनी गर्दन दबी (paavo tale apni garden dabi), यथार्थ का ज्ञान होता (yathrath ka gyan hota)*
NN-NN	Includes sequence of Noun as first word and Noun as last word in the sentence	22682	*तप और व्रत (tap or varat), हवा की तरह(hawa ki tarah)*

IV. PROPOSED METHODOLOGY

As discussed earlier the corpus used is Hindi corpus and the raw text from one of the famous Hindi Novel 'Godaan' is collected and filtered for carrying out the required phenomenon. There are mainly two views of Natural Language Processing (NLP) approaches, Classical view and Statistical/Machine Learning view.

The classical view deals with various rules of grammar made by humans and the basic phenomenon related to NLP like Phonology, Morphology, Lexical Analysis, Syntactic Analysis, Semantic Analysis, Pragmatics and Discourse. While, the Statistical/Machine Learning view helps in resolving the problem of ambiguity which is an inherent part of NLP.

Generally, the Indian languages are free from order of words, because the sentences contain some particles of language in between. For example

छात्र को महाप्रबंधक जी ने सम्मानित किया|

(*chhatr ko mahaparbhandhak ji ne sammanit kiya*)

अथवा (aathva)

महाप्रबंधक जी ने छात्र को सम्मानित किया|

(*mahaparbhandhak ji ne chhatr ko sammanit kiya*)

This type of problem exists many times with Indian Languages. There are various solutions provided in NLP in the form of statistical approaches for the above types of problems. For Multiwords also, this type of problem exists when some language particles exist in an n-gram sentence which can be a Multiword Expression, but avoided to become a MWE. For example

धूप में ऐसे ही बाल सफेद नहीं हुए है|

(*dhoop me aise hi baal safed nhi hue hai*)

Here, the words underlined are the language particles which are avoiding the idiom to become Hindi MWE. So it is the need of the hour to implement the baseline and the statistical techniques for any Information Retrieval or any other NLP tasks. The baseline measures mainly cover precision, recall and f-measure. The following section discusses various statistical measures used in this paper along with these baseline measures.

A. Statistical Measures for Hindi MWEs

Various statistical measures used in proposed work have been discussed in brief in the next section.

The Point wise Mutual Information (PMI)

The Point wise Mutual Information is the logarithmic ratio of the probabilities of an n-gram and its constituent words. 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. Likewise for n-gram (w_1 w_2 w_3 … w_n), the PMI is calculated as:

$$\text{PMI}(w_1, w_2, w_3 \dots w_n) = \log_2 \frac{P(w_1 w_2 w_3 \dots w_n)}{P(w_1)*P(w_2)*P(w_3)*\dots*P(w_n)}$$

The Dice Coefficient (DC) and Modified Dice Coefficient

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

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

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

Here, $f(w_1w_2)$ is the frequency observation of the bigram MWE which comprised of two words w_1 and w_2, whereas $f(w_1)$ and $f(w_2)$ are the frequency observations of the individual words. Likewise for n-gram (w_1 w_2 w_3 … w_n), DC and MDC are calculated as:

$$DC(w_1, w_2, w_3 \dots w_n) = \frac{n*f(w_1w_2w_3\dots w_n)}{f(w_1)+f(w_2)+f(w_3)+\dots+f(w_n)}$$

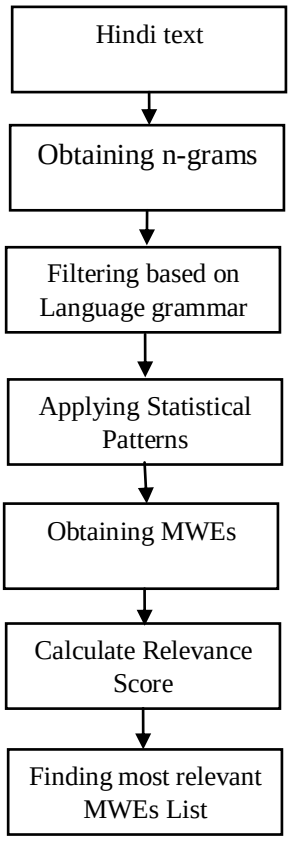
$$MDC(w_1, w_2, w_3 \dots w_n) = \frac{n*f(w_1w_2w_3\dots w_n)}{f(w_1)*f(w_2)*f(w_3)*\dots*f(w_n)}$$

In the proposed approach, PMI, DC and MDC formulas are used for both the bigrams and n-grams. Results are shown using tables and analysis is done using histograms.

Relevance Measure

Relevance is the measure of how relevant a given query is to the classification [22], 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. The procedure for finding the relevance measure is explained in the next section.

Proposed Methodology for Relevance Measure of Hindi MWEs*:*

As explained in literature survey author in [10] selected highly ranked sentences, which were different from each other. In the proposed work, highly ranked multiword expressions are selected and the relevance score of each multiword is measured with the complete set of multiword as well as with complete set of words in the corpus. The algorithm to calculate Relevance Measure for MWEs is given below:

Algorithm: RMMWE (Relevance Measure for MWEs)
1. Let W be the set of all the words in the Hindi document and M be the set of words participating in formation of Multiword Expressions, where M is a subset of W. 2. Let W_m be the weighted term frequency vector of words participating in formation of multiword expression, for all m ∈ W. 3. For each m ∈ W, find the relevance score of W_m and W, i.e. the inner product between the two (eg. W_{m1}*W_1+W_{m2}* W_2 + W_{m3} *W_3+…). 4. Select the highest relevance scored words participating in formation of Multiword Expressions, i.e. 'e'(say), and add it to summary. 5. Delete 'e' from W and recomputed M. 6. If number of sentences reaches a predefined value, terminate, else go to step 3.

B. Process Model

The process followed in this work is shown as below:

1. The text is collected from one of the famous Hindi novel "Godaan" and the Hindi corpus is created which is taken as input to the system.
2. Processing is done using POS taggers to obtained tagged file
3. Various categories of 2-grams and n-grams are considered which are most suitable for formation of MWEs in Hindi.
4. The n-grams are filtered using the Language grammar and proper procedure is used to obtain MWEs out of various n-grams.
5. The baseline measures and statistical measures are applied on the obtained Hindi MWEs to check the performance of developed system.
   5.1. In baseline measure, Precision, Recall and F-Measure are calculated and in statistical measures, PMI, DC and MDC are calculated with one more proposed measure, i.e. Relevance measure.
   5.2. Relevance measure is calculated and results are obtained for both types of measures.
6. Compare the results obtained from the baseline measures and the statistical measures. The relevance score is compared with all other statistical measure using histograms.

Fig. 1 Process Model for Hindi MWEs Extraction and Relevance measure

Considering the above mentioned procedure, the model works in the following manner as shown in figure 1.

V. EXPERIMENTAL RESULTS

Various experiments were carried out using base line methods as well as Statistical measures. The next section evaluates and discusses the result obtained from these experiments.

A. Baseline measures

The baseline methods, Precision, Recall and F-Score which were used in [12] for Hindi MWEs, are applied on the current dataset, and following results are obtained for 2-grams and n-grams based on baseline measures as shown in table III and table IV.

The evaluated results are analysed using histograms. The figure 2 and figure 3 clearly show how the baselines measures vary for different type of Hindi MWEs. Observations show that in case of 2-grams, compound adverb scores the highest score, while in case of n-grams; compound noun achieved the highest score.

B. Statistical Measures

The statistical measures, PMI, DC, MDC, and Relevance Measures for Hindi MWEs are applied on the current dataset, and results obtained are shown in table V. Here only 2-grams are considered for the statistical measures, which are extended for n-grams in table VI. The relevance measure is mainly considered for the evaluation and analysis of results. For 2-grams, Compound verb Hindi MWEs have the highest relevance score with value 10.6359, after that verb and adverb combination has scored 9.3781 and the compound adverb Hindi MWEs have the third highest relevance score of 8.8139. In baseline measures also, the above mentioned 2-grams have top scores.

Table III: Evaluation of baselines measures for n-grams

MWE Pattern	Total MWEs	Total Unique MWEs	Precision	Recall	F-Measure
Compound NN	10	7	0.5392	0.4118	0.4668
VM-NN	2243	1003	0.4502	0.3090	0.3665
NN-VM	1570	706	0.3318	0.3102	0.3206
NN-NN	5675	2098	0.6126	0.2700	0.3747

Table IV: Evaluation of baselines measures for 2-grams

2-grams	Total MWEs	Unique MWEs	Precision	Recall	F-Measure
JJ+NN	725	419	0.7250	0.3663	0.4867
JJ+PSP	8	5	0.5784	0.3846	0.4620
RB+JJ	30	11	0.8532	0.2683	0.4082
RB+RB	4	3	0.8395	0.4286	0.5675
NN+JJ	302	110	0.7517	0.2670	0.3940
NN+PSP	130	76	0.0825	0.3689	0.1349
NN+VM	1022	814	0.4716	0.4434	0.4570
NN+NN	806	529	0.800	0.3963	0.5300
VM+RB	6	4	0.8395	0.4000	0.5418
VM+RP	63	23	0.5837	0.2674	0.3668
VM+PSP	32	14	0.2236	0.3044	0.2578
VM+VM	253	109	0.8483	0.3011	0.4445
All 2-grams	3381	2117	0.500	0.3851	0.4350

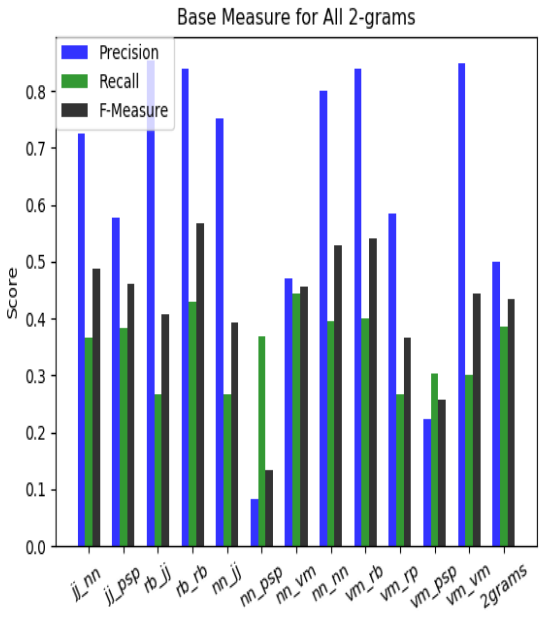


Fig. 2 Comparison of Baseline Measures for 2-grams

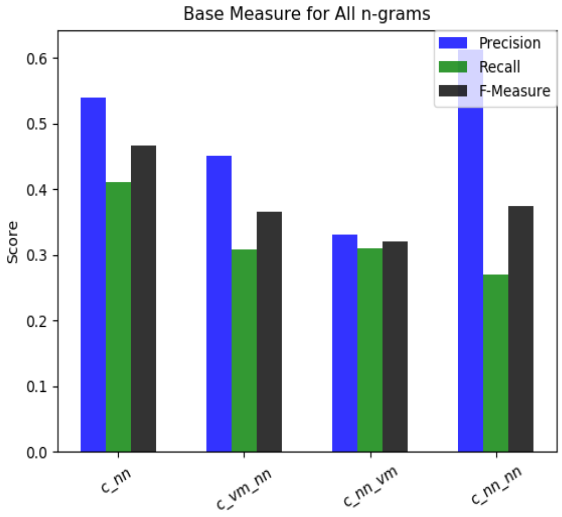


Fig. 3 Comparison of Baseline Measures for n-grams

Table V: Evaluation of statistical measures for 2-grams Hindi MWEs

2-grams	Total MWEs	Unique MWEs	PMI	DC	MDC	Relevance
JJ+NN	725	419	3.9516	0.3284	0.00106	3.6816
JJ+PSP	8	5	3.9172	0.0116	0.00005	1.3398
RB+JJ	30	11	4.8174	0.1446	0.00222	7.7287
RB+RB	4	3	4.1645	0.0494	0.00122	8.8139
NN+JJ	302	110	4.0901	0.1368	0.00044	4.2287
NN+PSP	130	76	0.5693	0.0507	0.00006	0.1256
NN+VM	1022	814	2.2005	0.2901	0.00017	1.6599
NN+NN	806	529	3.7449	0.1975	0.00010	7.2699
VM+RB	6	4	4.0441	0.0039	0.00005	9.3781
VM+RP	63	23	2.6548	0.0238	0.00002	2.6034
VM+PSP	32	14	1.6885	0.0160	0.00002	0.4148
VM+VM	253	109	3.9910	0.0854	0.00006	10.6359

All 2-grams	3381	2117	3.6579	0.6232	0.00032	4.5855

The analysis of evaluated results of statistical measures is performed using the histograms as shown in the below mentioned figures.

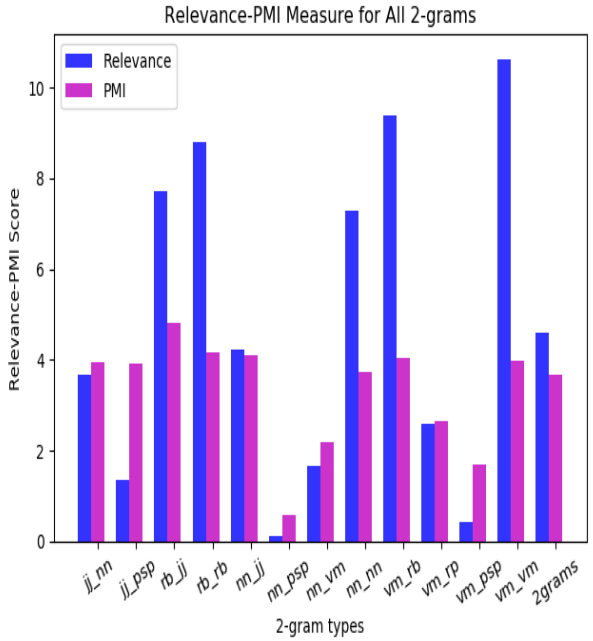


Fig. 4 Comparison of PMI and Relevance measure for 2-grams

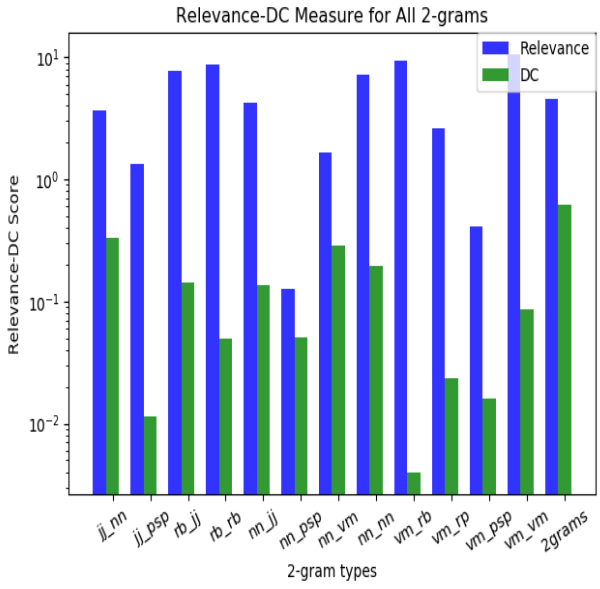


Fig. 5 Comparison of DC and Relevance measure for 2-grams

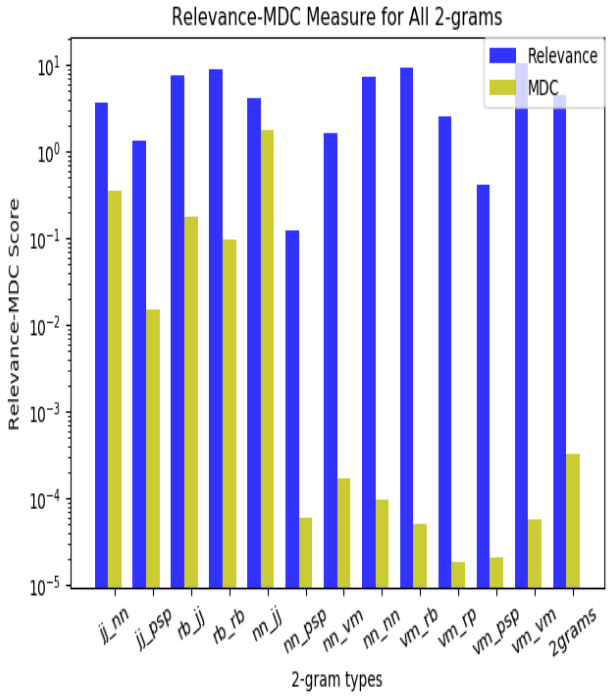


Fig. 6 Comparison of MDC and Relevance measure for 2-grams

In figure 4, PMI and Relevance score are compared. In figure 5, DC and relevance score are compared and in figure 6, MDC and Relevance score are compared. It is depicted from the figures that the PMI measure is most significant after the relevance measure while, DC and MDC are very less significant for all possible 2-grams as well as for all n-grams Hindi MWEs.

For n-grams Hindi MWEs, the statistics obtained is shown in table VI, where, Compound Noun Hindi MWEs pattern have the highest Relevance score with value 6.69. The other statistical measures, PMI, DC and MDC also have top scores in case of compound Noun Hindi MWEs pattern.

Table VI: Evaluation of statistical measures for n-grams Hindi MWEs

MWE Pattern	Total MWEs	Total Unique MWEs	PMI	DC	MDC	Relevance
Compound NN	10	7	5.26	0.352	0.0082	6.69
VM-NN	2243	1003	3.02	0.026	0.0003	4.24
NN-VM	1570	706	3.61	0.132	0.0026	3.64
NN-NN	5675	2098	4.96	0.216	0.0031	5.36

The same trend is followed for n-grams Hindi MWEs. Due to very less value of DC and MDC, logarithmic scales are used for the scores of DC and MDC on y-axis for both 2-grams and n-grams.

PMI, DC and MDC measures are compared with Relevance score in figure 7, 8, and 9 respectively for n-grams Hindi MWEs. It can be observed from these figures that PMI is more significant than the other two measures when compared with the relevance measure.

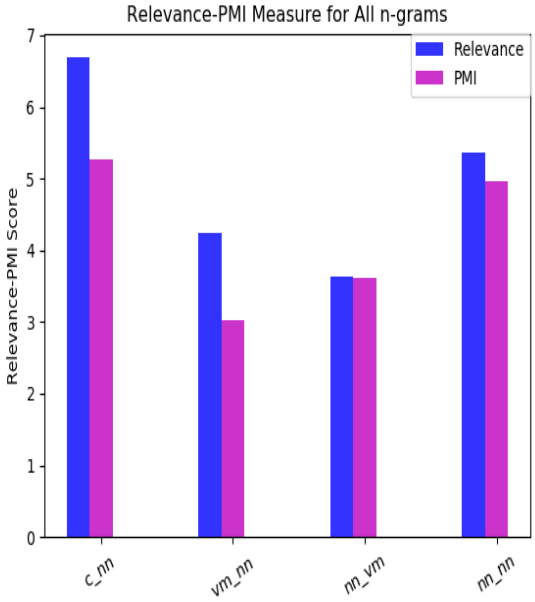


Fig. 7 Comparison of PMI and Relevance measure for n-grams

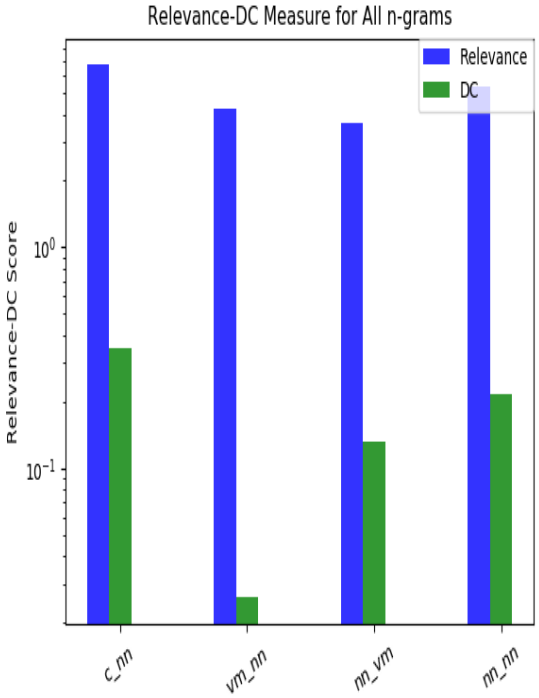


Fig. 8 Comparison of DC and Relevance measure for n-grams

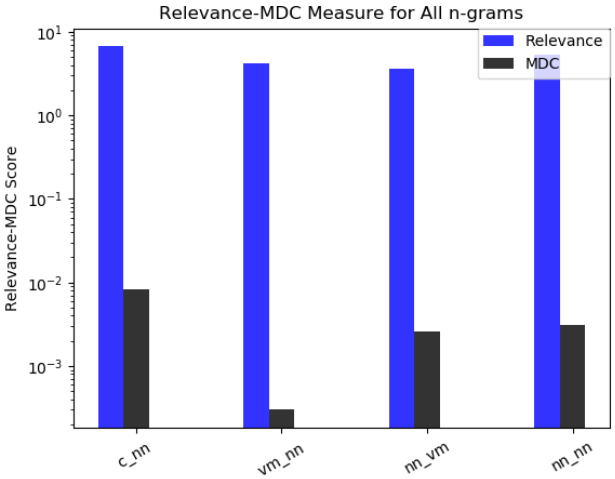


Fig. 9 Comparison of MDC and Relevance measure for n-grams

VI. CONCLUSION

Hindi MWEs extraction is considered as one of the key concepts in text processing to find the correct meaning of a sentence. In the proposed work, most relevant MWEs are considered by exploring relevance measure for Hindi MWEs which are not much researched like the English MWEs. It is necessary to find out how much relevant a particular type of Hindi MWEs is for 2-grams as well as for n-grams. In this paper the algorithm is designed and implemented for relevance measure of Hindi MWEs extraction. An analysis of Hindi MWEs is also done using the baseline and statistical measures. The baseline measures, Precision, Recall and F-Measure are evaluated for both 2-grams MWEs and n-grams MWEs. In the same way, statistical measures, PMI, DC, MDC are evaluated for 2-grams MWEs and n-grams MWEs. One additional statistical measure i.e. relevance measure is proposed in this paper for the accuracy. Relevance measure is evaluated on the basis of the frequency of occurrence of MWEs in Hindi Text. Earlier no such work has been done by using RM for statistical analysis of MWEs. The results are compared for both baseline measures and statistical measures. Later, histograms are used for analysis of results. It is predicted that the Relevance Measure and PMI measure are most significant and the DC and MDC are very less significant for all possible 2-grams and n-grams Hindi MWEs. Further the most relevant 2-grams and n-grams Hindi MWEs are evaluated and the significance of relevance measure is properly justified by analysing the results of experimental study. It is also suggested to consider relevance measure as one of the categories of statistical measures along with other measures like PMI, DC, MDC and so on.

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. 6(1). pp. 59–70.
2. I. Arnon, S.M. McCauley and M.H. Christiansen (2017). Digging up the building blocks of language: Age-of-acquisition effects for multiword phrases. Journal of Memory and Language. 92. pp. 265–280.
3. I. Arnon and N. Snider (2010). More than words: Frequency effects for multi-word phrases. J. Mem. Lang. 62(1). pp. 67–82.

4. C. Bannard, "A measure of syntactic flexibility for automatically identifying multiword expressions in corpora." In: Proceedings of the Workshop on a Broader Perspective on Multiword Expressions, pp. 1-8, 2007.
5. D. Chankrabarti, M. Mandalia and R. Priya, "Hindi Compound Verbs and their Automatic Extraction". Coling-2008. Manchester, pp. 27-30, 2008.
6. A. and S. Fazly Suzanne, "Distinguishing Subtypes of Multiword Expressions Using Linguistically-motivated Statistical Measures," Proc. Work. a Broader Perspect. Multiword Expressions, pp. 9–16, June, 2007.
7. J.F. Silva, G.P. Lopes. "A Local Maxima method and a Fair Dispersion Normalization for extracting multi-word units from corpora". In: Proceedings of Sixth Meeting on Mathematics of Language. pp. 369–381, 1999.
8. K. T. Frantzi, K. Frantzi, S. Ananiadou, and H. Mima (2000). Automatic Recognition of Multi-word Terms : The C-value / NC-value Method. Natural language processing for digital libraries. pp. 115–130.
9. C. Gonçalves, J. F. Silva, and J. C. Cunha, "A parallel algorithm for statistical multiword term extraction from very large corpora," Proc. - 2015 IEEE 17th Int. Conf. High Perform. Comput. Commun. pp. 219–224, 2015.
10. Y. Gong and X. Liu, "Generic text summarization using relevance measure and latent semantic analysis," Proc. 24th Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retr. - SIGIR '01, pp. 19–25, 2001.
11. R. Joon and A. Singhal, "Classification of MWEs in Hindi using Ontology". In: Proceedings of Sixth International Conference on Recent Trends in Information, Telecommunication and Computing – ITC 2015, Chennai, India, pp. 84-92, 28 March 2015.
12. R. Joon and A. Singhal, "A system for compound adverbs MWEs extraction in Hindi," 2015 8th Int. Conf. Contemp. Comput. IC3 2015, pp. 336–341, 2015.
13. R. Joon and A. Singhal (2017). Analysis of MWES in Hindi Text Using NLTK. Int. J. Nat. Lang. Comput. 6(1). pp. 13–22.
14. 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., pp. 12–19, July, 2006.
15. P. Pecina, "A machine learning approach to multiword expression extraction," Proc. Lr. Work. Towar. a Shar. Task Multiword Expressions (MWE 2008), pp. 54–61, June, 2008.
16. M. Peters, "Determining Relevance: How Similarity Is Scored. Moz", 2013. Available: https://moz.com/blog/determining-relevance-how-similarity-is-scored.
17. R.M.K. Sinha, "Stepwise Mining of Multi-Word Expressions in Hindi", In: Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World, Portland, Oregon, USA, pp. 110–115, 2011.
18. J.R. Williams (2017). Boundary-based MWE segmentation with text partitioning. Computation and Language. 3(3).
19. J. Xu, J. Yu and H. Wang, "Automatic Extraction of MWEs combining statistical and similarity approaches". In: Proceedings of 4th International conference on Genetic and Evolutionary computing, Shenzhen, China, pp. 256-259, 2010.
20. W. Zhang, T. Yoshida, and X. Tang (2011). A comparative study of TF*IDF, LSI and multi-words for text classification. Expert Syst. Appl. 38(3). pp. 2758–2765.
21. W. Zhang, T. Yoshida, and X. Tang, "Text Classification using Multi-word Features," In: Proceedings of IEEE International Conference on Systems, Man and Cybernetics, Montreal, Que., Canada, pp. 3519–3524, 2007.
22. M. Kirk (2017). Thoughtful Machine Learning with Python: A Test-Driven Approach. O'Reilly.
23. 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, pp.20-29. 2008.
24. 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.
25. T. Baldwin and S.N. Kim (2010). Multiword Expressions Handbook of Natural Language Processing. Boca Raton, USA: CRC Press.
26. M. Nokel, N. Loukachevitch, "Accounting ngrams and multi-word terms can improve topic models". In: Proceedings of the 12th Workshop on Multiword Expressions, Association for Computational Linguistics, pp.44–49, 2016.
27. C. Ramisch, P. Schreiner and M. Idiart. "An Evaluation of Methods for the Extraction of Multiword Expressions". In: Proceedings of 6th International conference on Language Resources and Evaluation, pp.50-53, 2008.
28. P. Rayson, S. Piao and S. Sharoff (2010). Multiword expressions: hard going or plain sailing?. Language Resources and Evaluation. 44(1). pp.1-5.
29. 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, pp. 1-15, 2002.

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.