# ANALYSIS OF MWES IN HINDI TEXT USING NLTK

Rakhi Joon and Archana Singhal

Department of Computer Science, University of Delhi, New Delhi, India

#### ABSTRACT

Natural Language Toolkit (NLTK) is a generic platform to process the data of various natural (human) languages and it provides various resources for Indian languages also like Hindi, Bangla, Marathi and so on. In the proposed work, the repositories provided by NLTK are used to carry out the processing of Hindi text and then further for analysis of Multi word Expressions (MWEs). MWEs are lexical items that can be decomposed into multiple lexemes and display lexical, syntactic, semantic, pragmatic and statistical idiomaticity. The main focus of this paper is on processing and analysis of MWEs for Hindi text. The corpus used for Hindi text processing is taken from the famous Hindi novel "KaramaBhumi by Munshi PremChand". The result analysis is done using the Hindi corpus provided by Resource Centre for Indian Language Technology Solutions (CFILT). Results are analysed to justify the accuracy of the proposed work.

#### **KEYWORDS**

NLTK, Hindi, Multi Words, MWEs, Analysis of MWEs, Hindi Text.

## **1. INTRODUCTION**

Hindi is most widely used language spoken as well as used for official work in India and other countries. It is originated from Sanskrit language and thus has a rich set of grammar and literature. The use of Hindi language has gained much popularity in NLP research. There are many resources developed for Hindi and some are under consideration. The NLTK framework is used for processing natural languages and providing comprehensive support for various NLP related tasks [8]. It is also used as a tool for carrying out the proposed work in which, the Hindi text is processed to generate the frequency graphs for various words and that can be further used for MWEs frequency measures.

MWEs comprises of two or more words (2-grams, 3-grams,..., n-grams) that convey some meaning as a whole, but independently the words convey some other meaning. The meaning of MWEs cannot be directly given by its components words. Some examples of MWE in Hindi are पश्चिमी सभ्यता(paschami sabhyata), जाल विवाह (baal Vivah), and so on. There are various views from different authors regarding MWEs :

"A unit whose exact meaning cannot be derived directly from the meaning of its parts" [12] "Arbitrary and recurrent word combination" [11] "Idiosyncratic concepts that cross word boundaries" [10]

As per the linguistic properties, "MWE are lexical items that can be decomposed into multiple lexemes and display lexical, syntactic, semantic, pragmatic and statistical idiomaticity" [5]. All the above definitions follow the property of "Idiomaticity" for MWEs. Idiomaticity refers to deviation from the basic properties of the words and applied at lexical, syntactic, semantic,

pragmatic and statistical level [2]. MWEs are the phrases that cannot be entirely predicted on the basis of standard grammar rules and lexical entries (http://mwe.stanford.edu/reading-group.html). The processing of text in any language depends on the grammatical constructs as well as the linguistic and syntactic properties of that particular language. The representation of a particular construct in any language is affected by the POS taggers. In Hindi, the extraction process of MWEs begins with the processing of text, finding n-grams, POS tagging and then applying the algorithms and procedures to carry out the required task. In the proposed task, the list of n-grams is extracted which can be further used for MWEs extraction. MWEs have not gained much popularity in Hindi language. In the proposed work an attempt has been made to process the Hindi text and to analyse the extracted types of MWEs present in Hindi text [2,3] on the basis of their properties and usage. The further analysis is done for accuracy. The corpus used for the processing is taken from one of the famous Hindi novel "KaramaBhumi by Munshi PremChand" and for result analysis, the Hindi corpus is obtained from CFILT.

The organization of the paper is done in the following manner: The next section gives a brief review of related work. In third section, Dataset used is explained with brief introduction about the corpus. In the next section, the description of Hindi Text processing using NLTK is given, how the various operations are performed on text and so on. In section 5, various types of Hindi MWEs and how to process the text for MWEs extraction in Hindi are proposed by the authors with detailed analysis of previous results. It is followed by the Conclusion and Reference sections.

## **2. RELATED WORK**

The basic understanding of NLTK and its specifications are given by many authors, in [7], the basics of python programming language for NLP are given which are used for carrying out the task of text processing in NLTK. Python is a particularly convenient language to use for writing scripts to perform natural language processing tasks [9]. The simplified version of NLTK toolkit and its efficient usage is provided in [8]. Various processing tasks are discussed in the paper. The more simplified version of toolkit, NLTK-Lite is discussed in [9], which provides ready access to standard corpora, along with representations for common linguistic data structures, reference implementations for many NLP tasks, and extensive documentation including tutorials and library reference. More detailed knowledge and syntax related information is given in [13-15]. In [1] an empirical study of integrating n-grams and multi-word terms into topic models is presented by the authors while maintaining similarities between them and words based on their component structure with the help of LDA-ITER algorithm by which the most suitable n-grams and multiword terms are incorporated. In [4], the author thoroughly examined various types of MWE encountered in Hindi from machine translation viewpoint. These specific types were not given proper attention in research for e.g. 'vaalaa' construct, doublets, replication etc. Many of the types are frequently used in daily life but are not given proper place in formal textual corpus. The author presented a stepwise mining of various MWEs in Hindi and their machine translation perspectives. The lexical and syntactic configuration of MWEs is discussed by the authors in [5], i.e. sometimes the component words preserve original semantics but sometimes MWEs encodes extra semantics. Further base types of MWEs and the linguistic properties of MWEs are explained. The basic methodology for MWEs is presented in [10] by the author. In this paper the difficulty of extracting Multi Words from a particular document along with types and properties of MWEs are discussed. Collocations are the combination of random and repeated words, as described in [11]. A number of techniques based on statistical methods for extracting collocations from large textual corpora are presented by the author. These techniques were based on some original filtering methods that give higher-precision output. A lexicographic tool, Xtract is used for the implementation of proposed techniques. In [6], the authors tried to find out whether MWE

identification methods can be efficiently applied to different types of MWEs and various languages. The authors further investigated the hypothesis that MWEs can be detected independently by the unique statistical properties of their component words, regardless of their type. The performance is measured and compared by using three statistical measures: Mutual Information,  $\chi^2$  and Permutation Entropy. In the proposed work the main focus is on Hindi text processing and MEWs extraction analysis. Earlier, the identification and Extraction of Hindi MWEs are done by the authors [2]. A new classification of Hindi MWEs is given which include MWEs types: Acronyms and Abbreviations, Compound Adverb, Complex predicates, foreign words and terms, Idioms, morphemes, Named entities, proverbs, replicating words and Standard phrases and Expressions. Some existing types are also included along with new types and ontology concept to have a better classification representation. A brief introduction of Hindi text. Further the classification of the compound adverbs is proposed on the basis of types of adverbs.

#### **3. DATASET**

A corpus is the collection of text for a particular subject. In any text processing task corpus plays a vital role as all the operations are performed on the text and the training and test data categorization is done on the data. In the proposed work, the experiments are performed on the Hindi Dataset taken from one of the famous novel by Hindi writer, "Munshi Premchand": "कर्म भूमी (*Karama Bhumi*)". This novel is a political novel in which political problems are presented through some families and thus the collection of text entitles a particular subject as well as the identity. The dataset contains 128842 Hindi words with 14081 unique words. As the total data in the book is divided into five sections. Last four sections together (containing 84637 words and 11009 unique words) are taken for training purpose and the first section (containing 44205 words and 7301 unique words) is taken for testing purpose. The Text processing as well as MWEs extraction tasks are applied on the uniformly distributed data, like here in a particular ratio of text is used as training set and rest as text set.

The data used for result analysis is obtained from the Hindi corpus provided by CFILT, IIT Bombay (<u>http://www.cfilt.iitb.ac.in/hin\_corp\_unicode.tar</u>). The file can't be directly used for the required process so it is processed further to obtain a refined form of text. After obtaining the refined text file from the repository, the Hindi POS tagger is used to get the text file containing the POS tagged data [3]. That POS tagged file is further processed to get the final list of Hindi MWEs.

# 4. HINDI TEXT PROCESSING WITH NLTK

The Hindi text is processed in Unicode format as NLTK supports this format. There are several repositories provided by NLTK for Indian Languages via "Indian Language POS-Tagged Corpus Reader". The following sections discusses about the various text processing tasks.

#### **4.1. Corpus Exploration**

Several corpora are provided by NLTK in many languages, for Indian languages also Indian Language POS-Tagged Corpus is there on which experiments can be performed. On the other hand, a new corpus can be added into Python and can be processed using NLTK as done in the proposed work. Use the NLTK corpus module to read the corpus "*kb\_complete.txt*", included in the Python directory, and various operations are performed on this text file to get the complete

understanding of how to process the Hindi text using NLTK. The sorted set as well as unique words are extracted out from the corpus.



Figure 1. Frequency Distribution of three common words in complete dataset



Lexical Dispersion Plot

Figure 2. Frequency Distribution of three common words in the Training dataset



Figure 3. Frequency Distribution of three common words in the Test dataset

The most commonly used words are obtained from the corpus, here most common 50 words are discovered and from those 3 words (" $\ddot{H}$ ", " $\exists \epsilon$ "," $\exists \epsilon$ ","," $\exists \epsilon$ ",",",";

#### 4.2. Operations done on text

There are a number of ways in which one can use the inbuilt NLTK supported python functions to process the text in a particular language. Since for Multiword, the individual words are combined together to form a new one, so Bi-grams, Tri-grams and n-grams are needed to be extracted from the text and thus can be further filtered out to extract multiword. Like here from the text the bi-grams and tri-grams are obtained:

>>> list(bigrams(words))

[('हमारे', 'स्कूलों'), ('स्कूलों', 'और'), ('और', 'कॉलेजों'), ('कॉलेजों', 'में'), ('में', 'जिस'), ('जिस', 'तत्परता'), ('तत्परता', 'से'), ('से', 'फीस'), ('फीस', 'वस्ल'), ('वसूल', 'की'), ('की', 'जाती'), ('जाती', 'है,'), ('है,', 'शायद'), ('शायद', 'मालगुजारी'), ...]

>>> list(trigrams(words))

[('हमारे', 'स्कूलों', 'और'), ('स्कूलों', 'और', 'कॉलेजों'), ('और', 'कॉलेजों', 'में'), ('कॉलेजों', 'में', 'जिस'), ('में', 'जिस', 'तत्परता'), ('जिस', 'तत्परता', 'से'), ('तत्परता', 'से', 'फीस'), ('से', 'फीस', 'वसूल'), ('फीस', 'वसूल', 'की'), ('वसूल', 'की', 'जाती'), ('की', 'जाती', 'है,')...]

Similarly n-grams can be obtained from the text. In MWEs extraction the initial step is to group the words and that is initiated using the n-grams operation on the text. Then further steps are followed to get the final MWEs list.

### 4.3. N-Grams as MWEs

A Multi word can be 2-grams, 3-grams or extended up to any number of words depending on the type as well as the property. With the help of NLTK, one can easily extract n-grams from the text as explained in the above section. The identification of MWEs from the n-grams is an important task because the first step is to extract the n-grams list that can be further processed to identify and extract MWEs. In figure 4 some n-grams extracted from the Dataset ("कर्म भूमी (*Karama Bhumi*)") are shown which can act as MWEs.



Figure 4. Pictorial representation of n-grams (हवाई जहाज, हृदयहीन दफ्तरी शासन, लफ्जों की झंकार का नाम गजल) as MWEs in Hindi

# 5. HINDI MWES CLASSIFICATION AND ANALYSIS

## 5.1. MWEs in Hindi and Significance

In any language, a sentence itself exhibits various properties and predicates like noun, pronoun, verb, adjective, adverb and so on. For the formation and extraction of MWEs from a corpus, first it is detected whether the combination of words which is to be categorized as MWEs exhibits the necessary and sufficient conditions of MWE which mainly include that words have to be separated by space or delimiter, non-compositionality of meaning and idiomaticity and many more [2]. The existing and new types of Hindi MWEs which were defined by the authors are listed in the below mentioned table.

S.No.	Туре	Example		
1.	Acronyms and Abbreviations	'ए पी जे अब्दुल कलाम '(A. P. J. Abdul Kalam)		
2.	Complex Predicates	वापस लेना (vaapas lena, return back)		
3.	Foreign Words and Terms	सीनियर डॉक्टर (Senior Doctor)		
4.	Idioms	ऊँगली उठाना: आलोचना करना (to lift finger, to blame)		
5.	Morphemes	पीने वाला पानी (pine vaalaa pani, drinking water)		
6.	Replicating words	अभी अभी (now now, recently)		
7.	Proverbs	दूर के ढोल सुहावने लगते हैं (The drums sound better at a distance, The grass is always greener on the other side)		
8.	Named Entities	नई दिल्ली (New Delhi)		
9.	Adverbs: Adverb of Place(स्थानवाचक) Adverb of Quantity (परिमाणवाचक) Adverb of Time (कालवाचक) Adverb of Manner (रीतिवाचक)	आगे पीछे (aage piche, front and back) बारी बारी (baari baari, one by one) आज कल (aaj kal, now-a-days) भली भांति (bhali bhanti, clearly)		

Table 1. Types of Hindi MWEs with examples [2].

Here, the last three types were proposed by the authors as MWEs in Hindi and rest six types were defined in [4]. The examples are taken from the paper itself. The result analysis of these types will be done in section 5.2 based on most of these types.



Figure 5. Ontological Classification of MWEs in Hindi

In [2], the classification tools used for the Hindi MWEs is Protégé, which provides ontological classification for the various class types. The representation of ontological classification of Hindi MWEs is shown below. The highlighted types [3] are the additions to the previous classification [4] of Hindi MWEs. In this way the classification is done in the previous work [2-4] and the results are analyzed in the next section.

## 5.2. MWEs Extraction: Result Analysis of existing work

The experiments in [3] and [4] are performed on the Hindi corpus taken from CFILT, IIT Bombay (<u>http://www.cfilt.iitb.ac.in/hin\_corp\_unicode.tar</u>) and results are collected from these papers. In this paper the results are used for analysis purpose only. Since the F-Measure score of many types are >90% as they frequently occur in the text and the properties exhibited by the words fulfil the necessary and sufficient condition to be a multiword.

S.No.	Туре	F-Measure Score
1.	Acronyms and Abbreviations	92.2%
	with dots	
2.	Replicating class	97.4%
3.	Doublet class	73.6%
4.	'vaala' construct class	90.7%
5.	Complex predicates and	77.2%
	compound words	
6.	Acronym and named entities	27.5%
7.	Compound Adverbs	90.66%
		(Average of three sample
		experiments)

Table 2	F-Measure sc	ore of various	types of	Hindi MWEs
1 uoie 2.	1 micubule be	ore or various	cypes or .	inter tor to Lo.

The identification of Acronyms and abbreviations and 'vaala' construct classes can be done easily and they occur frequently in text so there is a large collection of these types which results in higher f-score. Replicating words also occur frequently in any text and thus added to highest score. The Doublet class mainly includes pair of words that are antonyms, hyponyms or near synonyms of each other [4]. There is no such frequent occurrence of these types exist in any document but an average number exists, so the score is also an average. Since the acronyms and named entities occur in many verities in any text and are very difficult to find out as well as identified, this is the reason behind the least score of these types. The Compound adverbs also occur frequently in any text and these are of various types also as shown in Table 1, so gained a good F-Measure score. The pictorial representation of analysis is shown below.



Figure 6. Analysis of various Hindi MWEs types

#### 6. CONCLUSIONS AND FUTURE SCOPE

The MWEs are very important aspect in any language which is based on the property of "Idiomaticity" (deviation from the basic properties of the words) applied at lexical, syntactic, semantic, pragmatic and statistical level [3]. In this paper a detailed understanding of Hindi Text processing as well as MWEs is provided which is comprehensive enough for the processing of Hindi MWEs. NLTK is a toolkit available for processing various natural languages tasks. It supports Unicode format and provides separate packages for Indian languages. Various repositories of NLTK are being used for processing the tasks. Further the result analysis of previous work [2,3] has been done and shown using the graphical measures. The corpus used for the experiments has been collected from various sources for accuracy. For text processing the corpus is collected from the famous Hindi novel "कर्म भूमी (*Karama Bhumi*)" by Munshi Premchand and for analysis purpose the Hindi corpus provided by CFILT has been used.

The future scope lies in various directions, mainly in extension of ontologies for Hindi MWEs types and exploration for more new types.

## REFERENCES

- Nokel, M. & Loukachevitch, N., (2016), "Accounting ngrams and multi-word terms can improve topic models" Proceedings of the 12th Workshop on Multiword Expressions, Association for Computational Linguistics, pp44–49
- [2] Joon, R. & Singhal, A., (2015) "Classification of MWEs in Hindi using ontology," ITC 2015, in the proceedings of international conference on Recent Trends in Information, Telecommunication and Computing -ITC 2015, Vol. 6, pp84-92.
- [3] Joon, R. & Singhal, A., (2015), "A System for Compound Adverbs MWEs extraction in Hindi." in the proceedings of International Conference on Contemporary Computing (IC3), IEEE Computer Society, Vol. 8, pp336-341.
- [4] Sinha, RMK, (2011) "Stepwise Mining of Multi-Word Expressions in Hindi," In: Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE 2011), Portland Oregon, USA, pp110–115.
- [5] Baldwin, Tim & Kim, SN (2010) Multiword Expressions Handbook of Natural Language Processing, Edited by: Nitin Indurkhya and Fred J. Damerau (eds.) Second Edition, CRC Press, Boca Raton, USA, pp.267-292.
- [6] Ramisch, C., Schreiner, P., Idiart, M. & Villavicencio, A., (2008) "An Evaluation of Methods for the Extraction of Multiword Expressions", in the proceedings of International conference on Language Resources and Evaluation, Vol. 6, pp50-53.
- [7] Madnani, N., (2007), "Getting Started on Natural Language Processing with Python", Crossroads.
- [8] Bird, S., (2006), "NLTK: The Natural Language Toolkit" Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions, Association for Computational Linguistics, pp69–72.
- [9] Bird, S., (2005), "NLTK-Lite: Efficient Scripting for Natural Language Processing", Proceedings of the 4th International Conference on Natural Language Processing, 2005, pp. 11 – 18.
- [10] Sag, I.A., Baldwin, T., Bond, F., Copestake, A. & Flickinger, D., (2002) "Multiword expressions: A pain in the neck for NLP," In: Proceedings of Third International Conference on Computational Linguistics and Intelligent Text Processing: CICLing-2002, Lecture Notes in Computer, pp. 1-15.
- [11] Smadja, F. (1993) "Retrieving Collocations from Text: Xtract", Computational Linguistics 19, pp143-177
- [12] Choueka, Y. (1988), "Looking for needles in a haystack or locating interesting collocational expressions in large textual databases" In RIAO 88:(Recherche d'Information Assistée par Ordinateur). Conference, pp609-623.
- [13] Natural Language Toolkit. <u>http://nltk.sourceforge.net</u>
- [14] NLTK Tutorial. http://nltk.sourceforge.net/index.php/Book
- [15] <u>http://www.nltk.org/</u>

#### Authors

**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. She has many publications to her credit in reputed journals and International conferences.



