Impact of Multiword Expression in English-Hindi Language

Vivek Dubey[1], **Pankaj Raghuwanshi**[2], **Sapna Vyas**[3]

[1]Alpine Institute of Technology,
Dewas Road, Ujjain

Abstract

The use of multi-word expressions (MWE) in the English and Hindi language has been long recognized. This paper presents a study and analysis of kinds and structure of multiword expressions. For the automatic identification of multiword expressions, number of words and number of sentences from parallel English-Hindi corpus available on net has been investigated. Also in this paper, methodologies to extract MWEs and associated constraints have been highlighted.

Keywords: collocation, n-chunk, association of word, words mapping, corpus.

1. INTRODUCTION

Multiword expressions have considerably attracted researchers both in terms of theory and practical. Initially, linguists described multiword expressions theoretically [1-4]. Then, researchers started experimenting with this knowledge practically [4]; however, identifying and treating multiword expressions properly has proven to be a pain in the neck for Natural Language Processing (NLP), due to lack of adequate resources such as manually annotated corpora in many languages. In recent years, there has been a growing awareness in the NLP community about problems related to multiword expressions [3].

Multi-word Expressions (MWEs) are lexical items that consist of multiple orthographic words. Multiword expressions appear frequently in human language, in any kind of phrases. Traditionally phrase is defined as a group of words that does not contain a verb and its subject and is used as a single part of speech. For example in English-Hindi pair the phrase in pair

great_name=बहुत_अच्छा_नाम is known as MWE.

Eng1: it is a great name- Hin1: यह बहुत अच्छा नाम है

Other examples of MWEs are such as

Eng2: make up mind

Hin2: खुद को तेयार करना,

Eng3: afraid anymore

Hin3: कोई डर की बात नहीं,

Eng4: good morning

Hin4: सुप्रभात

In above example, MWEs cross word space due to its characteristics. Also it is not possible to replace one word in the expression by its synonym. Even though, in example, it is awkward to say low sound as small sound, short sound, little sound, etc.

Eng5:1ow sound

Hin5:मंद शब्द

Many times during translation of sentence into another language, due to syntactically idiomatic of MWEs, source sentence is unable to parse and always translate meaningful and sometimes nonsense result.

Many common names are also called as MWE such as high school, petrol pump, smart city, etc. Similarly, many verbal MWE have used frequently now a day's such as call off, make up mind to lose weight, take the benefit of the doubt.

Many applications like Optical Character Recognition (OCR), Computer Aided Lexicography (CAL), Word Sense Disambiguation (WSD), Part of Speech Tagging and Parsing, Foreign Language Learning, Machine Translation (MT) and Information Retrieval (IR) are based on Natural Language Processing (NLP) and naturally handling multiword expressions.

Expression of more than one word with blank space i.e. multiword expression is also known as collocations. It is meant as an attempt to emphasize the frequent co-occurrence of their components. The large variation that multiword expressions exhibit is a main challenge and motivation - why there is no unified strict definition [5]. Continuous efforts have been made to bind up words in the form of definitions for Multiword expression. Some of them are as follows:

- MWEs are habitual recurrent word combinations of everyday language. [6]
- Multiword expressions are expressions consisting of two or more words that correspond to some conventional way of saying things. [4]
- A pair of words is considered to be a collocation if one of the words significantly prefers a particular lexical realization of the concept the other represents. [7]
- Chunk as Idiosyncratic interpretations that cross word boundaries (or spaces) [3], i.e., there is a mismatch between the interpretation of the expression as a whole and the standard meanings of the individual words that make it up.

In this paper, strong influence of MWEs in English-Hindi language has been analyzed and discussed. Section-2 explains kind of MWEs, section-3 briefs the structure of MWEs, section-4 describes availability of MWEs, section-5 briefs computability of MWEs and section-6 highlights constraints with MWEs.

2. KIND OF MWES

Multiword Expressions can be classified into two classes as Morphosyntactic Classes: study of grammatical categories or linguistic units that have both morphological and syntactic properties and Traditional Classes: study of orthogonal view of complex and compound words.

2.1 Morphosyntactic Classes

In these types of multiword expression classes, there is lead word i.e. starting of chunk and it is named based on lead word such as multiword noun chunk, multiword verb chunk, multiword adverbial chunk and multiword adjectival chunk.

2.1.a Multiword Noun Chunk

In such type of noun chunk, there is head noun and next are other elements appended to it such as street light = सड़कों पर प्रकाश, high heels = ऊँची एड़ी के जूते. It also includes proper names such as name of city and state (New Delhi = नई दिल्ली, Madhya Pradesh = मध्य प्रदेश) institution and Department (i.e. Shri Govindram Seksaria Institute of Technology and Science = श्री गोविन्दराम सेकसरिया प्रौद्योगिकी एवं विज्ञान संस्थान, Union Public Service Commission = संघ लोक सेवा आयोग) or person (Vikram Singh Rathod-विक्रम सिंह राठोड), multi form (i.e. Content Management System = सामग्री प्रबन्धन प्रणाली).

2.1.b *Multiword Verbal Chunk*

In such verbal chunk, there is a head verb and other elements such as adverbs, object, compliments, are next to head verb. They are broadly classified as phrasal verb chunk and light verb chunk.

a) Phrasal Verb Chunk

It consists of first word as a base verb and then preposition or adverb. This is again sub-classified as a transitive prepositional verb chunk dealt as compositional but use a specific preposition introducing the object like think about=इसके बारे में and other as verb particle chunks like Turn on=चालू करना.

b) Light Verb Chunk

In the construction of Light Verbal Chunk (LVC), verb acts as light and a noun that conveys most of the meaning of the chunk like such as Do the review = समीक्षा करना.

2.1.c Multiword Adverbial and Adjectival Chunk

In such type of adverbial chunk and adjectival chunk, separate class is considered in sentence for example: Upside down = अस्तव्यस्त, Second hand = पुराना.

2.2 Traditional Classes

In these types of MWEs, there are three orthogonal types of chunks such as fixed chunk, idiomatic chunk and True collocation chunk.

2.2.a Multiword Fixed Chunk

In such type of chunk, there is word with space in words. And also it is a collection of words with individual meaning that really have nothing to do with another. Such as Just in case = अगर.

2.2.b Multiword Idiomatic Chunk

It is completely non-compositional semantics that means the literal interpretation does not depend on meaning of the chunk. Without idiomatic chunk verbs are not identified easily in text. For example, the meaning of verb chunk of Put in place in Hindi is औक़ात बता देना.

2.2.c Multiword True Collocation

Collocation refers to a relationship between words that frequently occurs with each other. The words together can mean more than the sum of their parts (the hindu epaper, chilli paneer). And these collocations are completely compositional (both syntactically and semantically) but are statistically idiosyncratic. These are just fixed term which do not have any alternate representations. For example:

Eng6: he filled his lungs with the fresh_air, and straightened his shoulders.

Hin6: उसने ताज़ी_हवा को अपने फेफड़ों में भर लिया और कन्धे सीधे किए ।

3. STRUCTURE OF MWE

The structure of multiword expression is studied according to MWEs categories. However, it is essential to ensure their internal structure for managing appropriate representation of MWEs in corpus.

3.1 Multi word Noun Chunk

The structure of Noun chunk is a combination of Determiner (Det.), Adjective (Adj.), Head Noun (N) and Preposition Chunk (PC) in order (Det.) ((Adj.) N (PC), where () indicates optionality. Examples are shown in table-1.

Table 1: Example of Noun Chunk

Structure	English	Hindi
Noun + N	Water Tank, Vijay Verma, Madhya Pradesh, decision making	पानी की टंकी, विजय वर्मा,मध्य प्रदेश , निर्णय लेना
Det + N	The girl, some rice	लड़की,कुछ चावल
(det)+ Adj + N	Strong coffee, a rich person	कड़ा पेय, एक अमीर व्यक्ति
(det) + N + PC	A rise in inflation.	मुद्रास्फीति में वृद्धि

3.2 Multi Word Verbal Chunk

Structure of Verbal chunk is a combination of Verb (V), Noun Chunk (NC), Preposition Chunk (PC) and Adverb

(Adv) in the order of V (NP) (PP) (Adv), where () indicates optionality. Examples are shown in table-2.

Table 2: Example of Verbal Chunk

Structure	English	Hindi
V+NP	kick the bucket	मरना / निधन होना
V+ particles	Cut down, give up ,pass on	छोटा कर देना, त्यागना, मरना
V + particles + preposition	Come up with, put up with	सूझना/जुटाना, बर्दाश्त करना

3.3 Multi Word Ajectival or Adverbial Chunk

A compound adjective is an adjective that contains two or more adjective words either with hyphen or without hyphen. These are constructed in class three ways: open class, hyphenated class and closed class. However, the compound adjective can also be formed with an -ly adverb and a participle or other adjective. Table 3 highlights some examples:

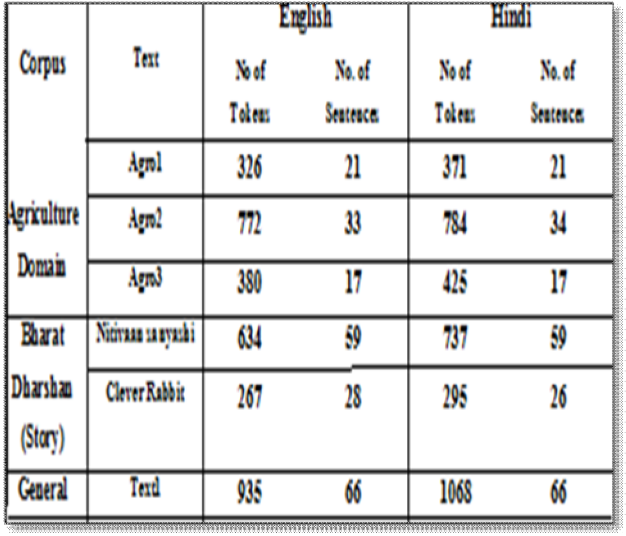
Table 3: Example of Adverbial & Adjectival Chunks

Structure	English	Hindi
Open Class	Cold blooded	निर्दयी
Hyphenated Class	well known / well-known	नामी गरामी
Close Class	housekeeping	गृह व्यवस्था
Adv + Past Participle	brightly colored	चमकीले रंग
Adv + Present Participle	regularly reviewing	नियमित निरीक्षण करना
Adv + Adj.	newly created post	नव सृजित पद

4. PARALLEL ENGLISH-HINDI CORPUS

A corpus is electronically stored and processed a large and structured set of texts used for statistical analysis and hypothesis testing. It is also useful to check linguistic rules. To analysis MWEs in English-Hindi Language, three corpus are used in this study. First is of agriculture domain, second is of bharat dharshan-hindi sahityik patrika and third is of general domain. In agriculture domain [12], there parallel English-Hindi text files are used and in bharat dharshan English-Hindi Panchtantra story corpus [13] two parallel stories are used. As general domain, HindEnCorp [15] is a parallel Hindi-English corpus freely available for non-commercial research purposes. However it contains 273 thousand sentences. As a sample, first 66 sentences are used in paper. The details of each English-Hindi file are analyzed in term of number of tokens and number of sentences in table-4. There is availability of MWEs in English-Hindi.

Table 4: Analysis of Eng-Hin Corpus



Corpus	Text	English		Hindi	
		No of Tokens	No. of Sentences	No of Tokens	No. of Sentences
Agriculture Domain	Agro1	326	21	371	21
	Agro2	772	33	784	34
	Agro3	380	17	425	17
Bharat Dharshan (Story)	Nitivaan sanyashi	634	59	737	59
	Clever Rabbit	267	28	295	26
General	Text	935	66	1068	66

5. COMPUTATIONAL METHODS

The raw text (i.e. html file, pdf file) has to be preprocessed. It consists of tools like tokenization, lemmatization and POS tagging. There are also flexible computation to handle MWEs [10] like Statistical Method, Linguistic Method, Hybrid Method and Machine Learning.

5.1 Statistical Method: In statistical methods for MWE extraction, firstly word association (chunk) is carried out and then proposed MI (mutual information) as an objective are measured for estimating word association norms. There are two developments of Statistical Methods for multi-word extraction: new association measures to rank candidates and new strategies to align the best candidate as MWE.

5.2 Linguistic Method: It is a traditional method for MWE extraction is words' POS tags. Both grammatical and syntactical requirements are there for a MWE.

5.3 Hybrid Method: A Hybrid method is generic method for MWE extraction by using both statistical and linguistic information of a word sequence to measure its possibility to be a multi-word expression.

5.4 Machine Learning Method: This is a dynamic method for MWE extraction employed AI (Artificial Intelligence) methods. It is used to discover initially new knowledge from either word information (frequency, association (chunk)) or Part-of-Speech information of words (order, POS sequence, etc) then the new knowledge is used to determine whether or not a new coming word sequence is a MWE.

6. LIMITATION

Multiword Expression is unique, intelligent, challenging for writer as well as reader. There is no simple solution to identify, extraction, presentation in text. There are a number of limitations in NLP tasks and applications [14] like translation, word-sentence alignment, in preprocessing, incompleteness, etc.

6.1 Translation quality of the parallel English-Hindi corpus

Whereas the sentences are indeed translations, the translations are largely and non-lexical, in the sense that context is used in order to extract the meaning and deliver it in different wording. As a result, it is sometimes hard or even impossible to align words based on the sentence alone.

6.2 Sentence alignment errors

A purely statistical sentence aligner is used to align sentences based on their length and token co-occurrence information. As a result, some sentences of similar length may incorrectly be marked as mutual translations. Of course, most of the word sequences in such sentences cannot be aligned and hence become MWE candidates.

The output of the sentence aligner contains only one-to-one sentence translations. As parallel corpora include non-lexical translations that sometimes can only be expressed in terms of one-to-many or many-to-one translated sentences, the sentence aligner may output one-to-one alignment, where one of the sentences is only a partial translation of another. The non-translated part of the sentence may contain false MWE candidates.

6.3 Word alignment errors

Sometimes a word sequence has a translation, but it is not aligned properly. Possible reasons for such errors are: Insufficient statistics of word co-occurrence due to the small size of the parallel corpus.

Errors caused by bidirectional translation merge. Often the alignment is correct only in one direction, but this information is lost after merging the alignment; this often happens in very long sentences. Another example of the problematic alignment caused by bi-directional merge in cases where the word aligner proposes N:1 alignment; usually these N words contain the correct sequence or a part of the sequence and the correct analysis of the bi-directional alignments may help filter out the incorrect parts (i.e., the analysis of the intersection of N and M sequences, where M:1 is Hindi-to-English and N:1 is English-to- Hindi alignments detected by the word alignment tool).

6.4 Noise introduced by preprocessing

Errors caused by morphological analysis and disambiguation tools may lead to wrong tokenization or to the extraction of an incorrect base form the surface form of the word. As a result, the extracted citation form cannot be aligned to its translation, and correctly aligned word-pairs cannot be found in the dictionary.

An additional source of errors stems from language specific differences in word order between the languages: Such problems can be handled with more sophisticated preprocessing that eliminates language specific differences, where not only morphology and function words are taken into account, but also language-specific word order.

6.5 Incomplete dictionary

If sentence and word alignment results are correct, and the correct word-to- word translation exists, but the translated pair is not in the dictionary, the word sequence may erroneously be considered an MWE candidate.

6.6 Parameters of the algorithm

Setting the threshold too high causes bi-grams that are subsequences of the longer MWEs to be false positives. During error analysis the algorithm drawback may be exposed: false MWE candidates that occur several times in the parallel corpus are selected to be MWE candidates only in a minority of these occurrences.

7. CONCLUSION

As the manual extension of resources is very costly and time consuming, there is a need for building resources and developing techniques for the automatic acquisition of MWEs from corpora to unknown words, to words missing relevant syntactic/semantic categories, to missing grammatical constructions, since it lies in between the lexicon and syntax and also tools/resources for it limited.

8. ACKNOWLEDGEMENT

Authors are grateful to Prof. Vineet Chataniya, IIIT Hyderabad and Dr. Swati Dubey, Vikram University, Ujjain for their valuable suggestions and Madhya Pradesh Council of Science and Technology Bhopal for the sanction of research project no. A/RD/RP-2/2014-15/234.

References

[1] Nunberg, G., Wasow, T., & Sag, I. A. Idioms. Language, 70(3), 491–539, 1994.

[2] Jackendoff, R. S., The Architecture of the Language Faculty. MIT Press, 1997.

[3] Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger, Multiword expressions: A pain in the neck for NLP. In Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002), pages 1–15, Mexico City, Mexico.

[4] Manning, C. & Schutze, H., Foundations of Statistical Natural Language Processing, chapter 5: Collocations. MIT Press, 1995.

[5] Rayson, P., Piao, S., Sharoff, S., Evert, S., & Moir´on, B. n. V., Multiword expressions: hard going or plain sailing? Language Resources and Evaluation, 1997.

[6] FIRTH J. R. Papers in Linguistics 1934-1951. Oxford, UK: Oxford UP, 1957. 233p.

[7] Darren Pearce, "Using conceptual similarity for collocation extraction." in Proceedings of the Fourth annual CLUK colloquium, 2001.

[8] www1:http://www.grammar.cl/english/compoundadjectives.htm. (last seen on 31/05/2015)

[9] www2:http://www.englishexercises.org/makeagame/viewgame.asp?id=8591 (last seen on 31/05/2015)

[10] Wen Zhang, Taketoshi Yoshida, Xijin Tang, Tu-Bao Ho. 2009. Improving effectiveness of mutual information for substantival multiword expression extraction. Expert Systems with Applications 36, 10919-10930, 2009 (ELSEVIER).

[11] www3:http://www.characterscountonline.biz/p/hindi-character-count.html. (last seen on 31/05/2015)

[12] www4: http://www.cfilt.iitb.ac.in/Downloads.html. (seen on 31/05/2015)

[13] www5:http://www.bharatdarshan.co.nz/magazine/article/child/113/panchtantra-stories-nitivaan.html. (last seen on 31/05/2015)

[14] Yulia Tsvetkov and Shuly Winte, Extraction of multi-word expressions from small parallel corpora. Natural Language Engineering,18, pp 549-573, 2011.

[15] www6: http://ufal.mff.cuni.cz/hindencorp. (last seen on 31/05/2015)

AUTHORS

Vivek Dubey is the Principal of Alpine Institute of Technology Ujjain, MP, India. He is the incharge of NLP Laboratory at the Institute. He did BE (CSE), M.Tech. (CT) and Ph.D. in Computer Science & Engineering. He has 15 years of engineering teaching experience, 3 years industry experience and 7 years in other. He has published around 45 papers in various national and international journals/conferences. He is also Editor and Reviewer in various journals.

Pankaj Raghuwanshi is working in Alpine Institute of Technology as Project Assistant in the department of Computer Science Engineering. He received BE degree in Mahakal Institute of Technology Ujjain in 2009.

Sapna Vyas is a Ph. D. Scholar of Pacific University, Udaipur, Rajasthan. She completed her MCA in 2013 from RGPV, Bhopal, M.P. She has participated in college projects- HR Summit, Indore and CSI Votting. Her interest is in Artificial Intelligence, Data Mining, and Text Processing.