Stepwise Mining of Multi-Word Expressions in Hindi

R. Mahesh K. Sinha
Indian Institute of Technology, Kanpur, India
sinharmk@gmail.com

Abstract

Multi-word expressions (MWEs) play an important role in all tasks that involve natural language processing. MWEs in Hindi are quite varied and many of these are of the types that are not encountered in English. In this paper, we examine different types of MWEs encountered in Hindi. Many of these have not received adequate attention of investigators. For example, 'vaalaa' constructs, doublets (word-pairs), replication, and a variety of verb group forms have not been explored *as MWEs*. We examine these MWEs from machine translation viewpoint. Many of these are frequently used in day-to-day conversations and informal communication but are not that frequently encountered in a formal textual corpus. Most of the conventional statistical methods for MWE identification use corpus with limited linguistic cues. These are found to be inadequate for detecting all types of MWEs that exist in real life. In this paper, we present a stepwise methodology for mining Hindi MWEs using linguistic knowledge. Interpretation and representation for some of these from machine translation perspective have also been explored.

1 Introduction

The identification and interpretation of multi-word expressions (MWEs) find application in almost all NLP tasks such as machine translation, information retrieval, question-answering etc. These are particularly helpful in parsing where the sequence of words forming the MWE is treated as a single word with a single part of speech (POS) tag. MWE information has been used for word alignment task (Venkatapathy et al., 2006). This is useful to lexicographers for deciding entry into the dictionary.

MWEs in Hindi are quite varied and many of these are of the types that are not encountered in English. No comprehensive work has been reported on Hindi MWE. In the following section a brief survey of related work is given. This is followed by a section on types of Hindi MWEs. Aspects of MWE identification, extraction and interpretation for Hindi are presented in section 4. Section 5 presents details of experimentation with results and section 6 concludes our investigation.

2 Related work

Baldwin et al. (2010) is an excellent review covering almost all aspects of MWEs. MWEs are characterized by non-compositionality, non-substitutability and non-modifiability (Brundage et al. 1992). Another definition of MWE is that it is 'any phrase that is not *entirely* predictable on the basis of standard grammar rules and lexical entries' (http://mwe.stanford.edu/reading-group.html). The design of a general purpose automated MWE extractor is dominated by using association measures such as point-wise mutual information and other statistical hypothesis tests (Church et al. 1990; Smadja 1993; Pecina 2008). Superior results have been reported when a supervised classifier is used with multiple association measures (Pecina 2008). The association measure is extended to include substitution to test semantic and statistical idiomaticity (Lin 1999). Moiron et al. (2006) use translation ambiguity to determine non-compositionality of MWEs.

For Hindi, there have been limited investigations on MWE extraction. Venkatapathy et al. (2005) considered N-V collocation extraction problem using MaxEnt classifier with certain syntactic and semantic features. Mukerjee et al. (2006) used POS projection from English to Hindi with corpus alignment for extracting complex predicates. Chakrabarti et al. (2008) present a method for extracting Hindi V+V compound verbs using linguistic features. Kunchukuttan et al. (2008) present a method for extracting compound nouns in Hindi using

Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE 2011), pages 110–115,
Portland, Oregon, USA, 23 June 2011.

statistical co-occurrence. Sinha (2009b) use linguistic property of light verbs in extracting complex predicates using Hindi-English parallel corpus. All of these works have considered only limited aspects of Hindi MWE. In this paper, we have considered almost all types of MWEs in Hindi and present method for their identification using linguistic features.

3 Types of MWEs in Hindi

Multi-word expressions appear in a variety of forms in Hindi. The primary criterion used in defining a MWE in this work is non-compositionality i.e. the meaning of MWE is not composed purely on the meanings of the constituent words (Baldin et al. 2002). From machine translation perspective, non-compositionality is of primary concern. In the following subsections, we enumerate different types of MWEs in Hindi.

3.1 Replicating words

All South Asian languages have replicating word feature (Abbi 1975, Abbi 1992) that exhibit non-compositionality property of MWE. This is found for all parts of speech. Some examples from Hindi (Sinha et. al. 2005) are: *ghar ghar* {house house} 'every house' ; *ruk ruk* {stop stop}'after stopping'; *baRii baRii* {big big} 'quite big'; *ek ek* {one one}; 'every one' or 'one by one'; *dhiire dhiire* {slow slow} '(quite) slowly' or 'gradually'; Replicating words may also have a particle in between and the meaning changes. Example: *paani hi paani* (water only water) 'water all over'. Another class of MWE is where the replicating word is in singular form of the preceding word. An example is: *dinon-din* (days-day) 'day by day' or 'gradually'.

It should be noted here that not all replications make an MWE (see section 4).

3.2 Doublets / pair of words, Samaas and Sandhi

A pair of words that are antonym of each other may form an MWE. Example: *din-raat* (day night) 'all the time'. Yet another class is where the meaning of the doublet is usually a hyponym or a near synonym of the pair of the words. Example: *roji-roti* (job bread) 'employment'. When there is a change of gender in the pair of words, it may represent a group. Example: *betaa-betii* (son daughter) 'issues'. When the second word in the pair of words is a non-sensical word providing rhythm to the group, the meaning is hyponym of the preceding word. Examples: *chaay-vaaya* {tea vaaya} 'snacks'; *taix-viax* {tax viax} 'tax etc'.

Samaas (N+N, A+N) and Sandhi (means joining or fusion of words) are Hindi grammatical constructs at the morphological level and are borrowed concepts from Sanskrit. In Samaas, while combining the two words, the intervening postposition markers are deleted. Samaas are of different kinds depending upon the semantics of the constituent words involved and their importance (head word) in the resulting combined word. Examples: *rasoi* (cooking) +*ghar*(house) = *rasoighar* (house <u>for</u> cooking = kitchen); *ganga* (Ganges)+*jal*(water) = *gangajal* (water <u>from</u> Ganges). Sandhi is a process by which two words in Hindi get co-joined to yield a single word. This process could be recursively applied and quite complex compositions with multiple words are possible. The words formed by the process of Sandhi and some of the Samaas, result in a single word and as such cannot be called an MWE. However, they are very large in number in Hindi with innumerable combination of words. It is not practical to store all of them in a dictionary. Hence algorithms are designed to decompose the word into constituent words for interpretation. Thus, in a sense, it is the reverse process of MWE.

3.3 Vaalaa morpheme constructs

The 'vaalaa' Hindi morpheme may appear in different morphological forms as 'vaalaa', 'vaalii', 'vaale' or 'vaalo.M'. All the constructs involving 'vaalaa' are candidates for MWE. The multi-word may involve just the preceding word or both preceding and following words. The morpheme 'vaalaa' as such has no meaning. Examples (Sinha 2009a): *jaane vaalaa* (go vaalaa) 'about to go'; *doodh vaalii balti* (milk vaalii bucket) 'bucket filled with milk'; *lohe vaalii balti* (iron vaalii bucket) 'bucket made of iron'; *dilli vaalii gaadii* (Delhi vaalii train) 'train to/from Delhi'; *nahaane vaalaa sabun* (bathe vaalaa soap) 'soap used for bathing'; *sabzii vaalaa* (vegetable vaalaa) 'vegetable seller'.

3.4 Complex and Compound Verbs

The complex predicates and compound verb forms as MWEs have been widely studied (Hook, 1974; Abbi, 1992; Mohanan, 1994; Butt, 1995; Venkatapathy et.al., 2005; Mukerjee et. al., 2006; Chakrabarti et. al., 2008; Sinha 2009b). A complex predicate is a multi-word expression (MWE) where a noun, a verb or an adjective is followed by a light verb (LV) and the MWE behaves as a single verb unit. LV (Sinha 2009b) can also be a main verb. A compound verb form has the main verb in its root/stem form followed by conjugated light verbs. In Hindi compound verbs, the primary meaning of the light/helping verbs are often completely lost and may lead to a different semantic interpretation or result in affecting tense, aspect and modality of the compound verb. A few illustrative examples (light verbs are shown underlined): *daan denaa* (donation give) ‘to donate’; *mukka maaranaa* (fist kill/beat) ‘to punch’; *mukka de maaranaa* (fist give kill/beat) ‘to blow punch’; *mukka maaraa gaya* (fist kill/beat went) ‘was punched’; *mukka maaraa gaya thaa* (fist kill/beat went was) ‘had been punched’; *mukka maaraa jaa rahaa thaa* (fist kill/beat go continue was) ‘was being punched’; *mukka paRaa* (fist lie)‘got punched ’; *ruka jaao* (stop go) ‘stop’; *aa jaao* (come go) ‘come’; *galati kara baiThanaa* (mistake do sit); ‘commit mistake (unintentional)’.

There are innumerable numbers of such MWEs in Hindi. However not all verb forms are MWEs.

3.5 Acronyms and Abbreviations

The acronyms and abbreviations in Hindi differ from their English counterparts. For example, the name ‘Mohandas Karamchand Gandhi’ may be abbreviated as ‘ma. ka. gaandhii’ (taking the first letter) or ‘mo. ka. gaandhii’ (taking the first letter with associated vowel modifier) or ‘ema. ke. gaandhii’ (taking the English alphabet letter). Similarly, the Hindi acronym for ‘Bharatiya Janata Party’ could be ‘bee. je. pii.’ (first English characters with dots) or ‘beejepii.’ (first English characters with no dots) or ‘bhaa. ja. paa.’ (first Hindi character with associated vowel modifier with dots) or ‘bhaajapaa’ (first Hindi character with associated vowel modifier with no dots). Although acronyms without dots are single words but they represent MWEs.

3.6 MWEs with foreign words and terms

It is often a common practice to mix foreign words and terms in day-to-today conversation in Hindi (Sinha et al. 2005b). Sometimes there are morphological variations to these as per Hindi grammar. These may appear as MWEs with arbitrary combinations. Some of these are institutionalized MWEs. Examples: *skilda* (skilled) *mainegaron* (managers); *spektram* (spectrum) *laaiisenson* (licenses). Here, the words *mainegaron* and *laaiisenson* are plural forms of the transliterated English words ‘manager’ and ‘license’ respectively, but the morphological changes are as per the Hindi pluralization rule. Since the foreign root word may undergo morphological variation as per Hindi grammar or may retain its English form, a cross morphological analysis is required to be done to extract the root word. Further, the transliteration of foreign word has a number of phonetic variations which needs to be considered before a look up into the English dictionary is performed. This class of MWE is not focused in this study.

4 Identification, extraction and interpretation of MWEs in Hindi

In this paper, we have considered only those MWEs that are particularly applicable to Hindi. The general characteristics of these MWEs have been outlined in the preceding section. We use these very characteristics in extracting the MWEs from the corpus. The extraction of MWEs that are more generally based on collocation and cooccurrence, require exhaustive and representative corpus to succeed which is not available for Hindi.

For identifying MWEs, we use multiple strategies and resources depending upon the class of the MWEs. The process of identification is semiautomatic. The automatic process generates the probable MWEs and then filtered manually. In future, the process can be fully automated using this tagged data through machine learning. A monolingual corpus and a lexical database (dictionary) are used in all the cases. In addition, a bilingual English-Hindi corpus and a Hindi wordnet are used for identifying some. We attempt to provide limited interpretation for some of these. Our method is mostly based on linguistic knowledge. We also show how these interpretations are engineered for a machine translation task by making appropriate substitutions in the source text.

For identification, there is a preferred order in which we mine them as it helps in further processing. At a broad level, the processes are: sentence boundary identification; POS tagging;

morphological analysis; identification of acronym and abbreviation with dots; Hindi chunker and verb-phrase form separation; identification of replicating class; identification of doublet class; identification of *vaalaa* morpheme construct class; complex predicates and compound verb identification; identification of acronym (with no dots); and identification of named-entities.

After the sentence boundary identification, POS tagging and the morphological analysis, the identification of acronyms and abbreviations that have dots associated with them, is carried out using a rule base. Next, chunking is performed. Chunking is a process of performing shallow parsing of the sentence where the words having affinity with each other at a syntactic level are grouped together. An example (chunks are shown within curly parentheses and English equivalent is enclosed within parentheses):{*bhagawaan raam ke haathon*}(by Lord Ram) {*mahaabalii raavana*}(mighty Ravan) {*yuddha bhoomi men*}(in battlefield) {*maara daalaa gayaa thaa*}(had been killed). In chunking, firstly the verb group is identified. Since Hindi is a verb ending language, a finite state machine (FSM) is designed which starts scanning the words from the rear end (right to left) for possible inclusion in the verb group based on the POS tag and the morphemes (Gune et al. 2010) of the words. A Hindi complex verb group may consist of auxiliaries, light verbs, predicate verbs and intensifiers besides the main verb. Such verb groups make an MWE because of its non-compositionality. In the above example, the last chunk which is the verb group chunk, is reproduced with meanings:{*maara* (kill) *daalaa* (put) *gayaa* (went) *thaa* (was)} (had been killed). Here main verb is *maara* (kill), *daalaa* (put) is a light verb making *maara daalaa* a predicate verb, *gayaa* (went) is an intensifier and *thaa* (was) is an auxiliary verb. The sequence of words that constitute the verb group could be quite long and is usually delimited by a postposition, a punctuation mark or a noun that does not form part of a predicate verb.

Identification of replicating words with a space, hyphen or a particle in between, and with plural-singular combination are searched within a chunk as identified in the earlier stage. The chunker creates a surface linear parse structure for the sentence and so is useful in eliminating false groupings of the replicating words. Replicating words (exact match) with a hyphen in between are definite MWEs while those without hyphen may not be so. In general, their identification and interpretation depends upon the associated POS and semantic role. Given below is an example rule (Sinha et al. 2005a) :

```
If the replicative verb has a suffix –te and the
   main verb is of the 'resultive:psych' type
then <verb_x-te><verb_x-te> =>
   due to|of <verb_x>+ing
```

This rule when applied to the Hindi sentence, *vah daurate daurate thak gayaa* (he run run tire went), yields the interpretation as 'He got tired of running'. For machine translation, the replicating words 'daurate daurate' is substituted by a dummy variable (say 'dv1') with POS as an adverb and its value will be stored as 'of running'. The Hindi sentence is modified to '*vah dv1 thak gayaa*' for machine translation. This kind of strategy is applied for all interpretations. The ambiguity resolution, if any, is left to the translation engine to tackle.

Hindi wordnet (Narayan et al., 2002) is used for checking antonym, hyponym and near synonym relationships in the pair of words. The doublets with hyphens are sure candidates of MWE but the doublets without hyphen are considered MWEs if they belong to the same chunk. In a semi-onomatopoeia combination, the second word is usually an unknown word and its suffix provides a rhythmic companionship. This is what is used in their identification. For example, in "*chaaya vaaya*". '*vaaya*' is an unknown word and the suffix '*aaya*' is common to the two words. The interpretation of the semi-onomatopoeia combination is usually the hyponym of the first word. Thus "*chaaya* (tea) *vaaya*" is interpreted as 'snacks'.

Since all 'vaalaa' constructs are MWEs, the mere presence of 'vaalaa' morpheme facilitates their identification. The major issue is that of determining the adjoining words that form the MWE. For this a number of rules are devised based on the semantic interpretation of the MWE. Given below is an illustration (Sinha 2009a):

> "If 'vaalaa' is preceded by a verb in infinitive form and followed by an auxiliary verb, then it represents a future event (about to action representing the verb). The verb+vaalaa is a MWE."

A number of such rules are devised using semantic relationships obtained through wordnet or a lexical database.

For identification of compound verb, we use a list of 30 light verbs (Sinha 2009b). When a verb in its stem form, is followed by a light verb, it is identified as a compound verb (strategy used is similar to Chakrabarti et al. 2008). This rule is applied recursively to make a larger group.

For the identification of complex predicates, we use a parallel aligned Hindi-English corpus. A simple heuristic of the absence of the light verb translated into English in the parallel corpus is taken as the complex predicate (Sinha 2009b).

We use an in-house named-entity recognizer. All the forms of the names as outlined in section 2.11 are detected and interpreted accordingly. All the unknown word sequences are considered probable candidates for MWEs. A name gazetteer is used to identify the named entities and the rest are checked for being acronyms. A majority of acronyms without dots in Hindi are mappings of English acronyms. Therefore, the individual Roman alphabet character mapping to Hindi is utilized to detect these. The names that are also valid dictionary words do not get identified.

5 Experimentation and Results

As a general corpus is very sparse in terms of occurrences of each type of MWE, we created corpus consisting of instances of different types sampled from various sources such as news articles, grammar books and corpora available at http://www.cfilt.iitb.ac.in/hin_corp_unicode.tar, www.cdacnoida.in/snlp/digital_library/gyan_nidhi.asp. The sampling was mostly done through an automatic process where templates of patterns were supplied with randomly picking up words from a list of frequent words created by an analysis of a Hindi corpus. These were further clubbed into six different classes of MWEs where each class consisted of similar MWE type. This helped us in taking care of sparseness to some extent to make our study more meaningful. Our sample space for each class consisted of approximately 5000 words.

Table 1 shows the results of our experimentation. The f-score varied from 27% to 97%. The identification of named entities is poor as it is based on a gazetteer and unknown words. The performance of the MWEs identification in the doublet class is affected due to inadequacy of the Hindi wordnet that has been used for some of its subclasses. The Hindi wordnet is not complete and many of the antonyms, hypernyms/hyponyms and ontological classification are not present.

Table 1: Experimental results

MWE Type	F-score
acronym and abbreviation with dots	92.2%
replicating class	97.4%
doublet class	73.6%
‘vaala’ construct class	90.7%
Complex predicates and compound verbs	77.2%
acronym (with no dots) and named entity	27.5%

6 Conclusions and Discussions

In this paper, we have provided comprehensive details and characteristics of the MWEs that are specific to Hindi. Many of these characteristics are generic in nature in the sense that it is not based on any statistical inference but it is the linguistic property that helps in MWE extraction. For example, all replicating words irrespective of their POS, all doublets with plural-singular form combinations, ‘vaala’ forms, complex verb forms etc are all strong candidates for MWEs in Hindi irrespective of whether these have earlier been encountered in the corpus or not. This means that even the low frequency MWEs can be captured. All the statistical approaches require the corpus to be representative and exhaustive in order to be able to yield reliable results (limitations: Kunchukuttan et al., 2008). Moreover, most of the idiosyncrasies of the language surface in informal conversations and are rarely available in regular textual corpora (Baldwin et al., 2010). The statistical approach will anyway be needed to mine other types of MWEs and discover new and institutionalized MWEs (mostly domain specific) that keep getting added (Baldwin et al., 2010). However, our stepwise methodology of filtering MWEs in stages provides a reduced sample space for searching the MWEs. Thus the size of the bag of the context words (Katz, 2006) needed for their identification and interpretation gets reduced. One of the primary aims of this study is to collect MWEs of different types in a semi-automatic way for use by the lexicographers for possible entry in the dictionary and stepwise mining is helpful.

Our contribution lies in presenting a comprehensive study of all types of MWEs encountered in Hindi and devise methods for their mining. We have not been able to present a detailed description of our method due to space constraints. In future work, we would like to hybridize rule based and statistical methods with bootstrapping of the data obtained for different classes.

References

Amitabh Mukerjee, A. Soni, and A. Raina. 2006. Detecting Complex Predicates in Hindi using POS Projection across Parallel corpora. ACL Workshop on Multiword Expressions

Anoop Kunchukuttan and Om P. Damani. 2008. A System for Compound Noun Multiword Expression Extraction for Hindi. Proceedings of International Conference on Natural Language Processing (ICON2008)

Anvita Abbi. 1992. Reduplication in South Asian Languages: An Areal, Typological and Historical Study. Allied Publishers, New Delhi.

Anvita Abbi. 1975. Reduplication in Hindi: A Generative Semantic Study. Dissertation Abstracts Internacional, Vol. 36, University of NY (1975).

Anvita Abbi. 1992. The explicator compound verb:some definitional issues and criteria for identification. Indian Linguistics, 53, 27-46.

B. V. Moiron and J. Tiedemann. 2006. Identifying idiomatic expressions using automatic word alignment. EACL 2006 Workshop on Multiword Expressions in a multilingual context.

Debasri Chakrabarti, Hemang Mandalia, Ritwik Priya, Vaijayanthi Sarma and Pushpak Bhattacharyya.2008. Hindi Compound Verbs and their Automatic Extraction, Computational Linguistics (COLING08), Manchester, UK.

D. Lin. 1999. Automatic identification of noncompositional phrases. ACL-1999.

D. Narayan, D. Chakrabarti, P. Pandey, and P.Bhattacharyya. 2002. An experience in building the IndoWordNet - a WordNet for Hindi. Global WordNet Conference.

G. Katz and E. Giesbrechts. 2006. Automatic identification of noncompositional multi-word expressions using Latent Semantic Analysis. ACL Workshop on Multiword Expressions.

Harshada Gune, Mugdha Bapat, Mitesh Khapra and Pushpak Bhattacharyya. 2010. Verbs are where all the Action Lies: Experiences of Shallow Parsing of a Morphologically Rich Language, Computational Linguistics Conference (COLING 2010), Beijing, China.

I. A. Sag,, T. Baldwin, F. Bond, A. C-opestake, and D. Flickinger. 2002. 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), Mexico City, Mexico. 1–15

Jennifer Brundage, M. Kresse, U. Schwall and A. Storrer. 1992. Multiword lexemes: A monolingual and contrastive typology for natural language processing and machine translation. Technical Report 232, Institut fuer Wissensbasierte Systeme, IBM Deutschland GmbH, Heidelberg.

K. Church and P. Hanks. 1990. Word association norms, mutual information, and lexicography. Computational Linguistics. 16(1).

Miriam Butt. 1995. The Structure of Complex Predicates in Urdu. CSLI Publications.

P. Pecina. 2008. Lexical Association Measures. Ph. D. thesis, Charles University.

Peter Edwin Hook. 1974. The Compound Verb in Hindi. Center for South and Southeast Asian Studies: The University of Michigan.

R. Mahesh K. Sinha. 2009a. Learning Disambiguation of Hindi Morpheme 'vaalaa' with a Sparse Corpus, The Eighth International Conference on Machine Learning and Applications (ICMLA 2009), Miami, Florida, USA

R. Mahesh K. Sinha. 2009b. Mining Complex Predicates In Hindi Using Parallel Hindi-English Corpus, ACL-IJCNLP 2009 Workshop on Multi Word Expression (MWE 2009), Singapore.

R. M. K. Sinha and Anil Thakur. 2005a. Dealing with Replicative Words in Hindi for Machine Translation to English, 10th Machine Translation summit (MT Summit X), Phuket, Thailand., 157-164.

R. M. K. Sinha and Anil Thakur. 2005b. Machine Translation of Bi-lingual Hindi-English (Hinglish) Text, 10th Machine Translation summit (MT Summit X), Phuket, Thailand.. 149-156.

Sriram Venkatapathy and A. Joshi. 2006. Using information about multi-word expressions for the word alignment task. In Proceedings of the COLING/ACL Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties, Sydney, Australia, 53–60.

Sriram Venkatapathy and Aravind K. Joshi, 2005. Relative compositionality of multi-word expressions: a study of verb-noun (V-N) collocations, In Proceedings of International Joint Conference on Natural Language Processing - 2005, Jeju Island, Korea, 553-564.

Tara Mohanan. 1994. Argument Structure in Hindi. CSLI Publications, Stanford, California.

Timothy Baldwin and F. Bond. 2002. Multiword expressions: Some problems for Japanese NLP. In Proceedings of the 8th Annual Meeting of the Association for Natural Language Processing (Japan), Keihanna, Japan, 379–382.

Timothy Baldwin and Su Nam Kim. 2010. Multiword Expressions, in Nitin Indurkhya and Fred J. Damerau (eds.) Handbook of Natural Language Processing, Second Edition, CRC Press, Boca Raton, USA. 267-292.