RESEARCH REVIEW JOURNALS
Volume-03
Issue-12
December -2018

ISSN: 2455-3085 (Online)
RESEARCH REVIEW International Journal of Multidisciplinary
www.rrjournals.com [UGC Listed Journal]

Identification and Extraction of Replicated Punbjabi Multi Word Expressions

Kapil Dev Goyal

Assistant Professor, SBAS Khalsa College,Sandaur, Sangrur (India)

ARTICLE DETAILS

Article History
Published Online: 10 December 2018

Keywords
Multiword Expressions (MWEs), Natural Language Processing

ABSTRACT

Multiword Expressions (MWEs) play an important role in Natural Language Processing. Multiword Expression is a combination of two or more words but treated as a single word. MWEs in Punjabi 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 Punjbai. 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 methodology for identification and extraction of Punjabi 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 Punjabi are quite varied and many of these are of the types that are not encountered in English. No comprehensive work has been reported on Punjabi MWE.

2. Related Work

Baldwin et al. (2010) [1]; Lahari Poddar [2]; MunishMinia [4] presented an excellent review on Multiword Expression. They reviewed almost all aspects of MWEs such as characteristics of MWEs, types of MWEs, extraction techniques, etc. Baldwin also introduced some analytic techniques for MWEs to analyze fixed expression, semi-fixed expression, and syntactical flexible expression using the constraint-based Head-driven Structure Grammar (HPSG), whereas Lahari Poddar reviewed all MWEs extraction approaches such as Rule base approaches, Statistical Methods, Word Association Measures, retrieving collocation using XTRACT and conceptual similarity and also discussed extraction of MWEs from small parallel corpora.

R. Mahesh K. Sinha (2011) [5] examined different types of MWEs encountered in Hindi such as Replicating words, Samaas and Sandhi, Hindi acronyms and abbreviations, vaala morpheme contruct, etc.

Brundage et al. (1992) [6] characterized MWEs by non-compositionality, nonsubstitutability and non-modifiability.

Church and Hanks (1990) [7]; Smadja (1993) [8]; Pecina (2008) [9] designed an automatic extractor of MWEs by measuring association using statistical methods such as Point-wise Mutual Information (PMI) and other statistical hypothesis tests. (Pecina 2008) reported superior results by using a supervised classifier used with multiple association measures and compared 55 statistical association measures to validate and rank German MWEs.

Agarwal et al. (2004) [10] proposed a method to automatic extraction of Multi-word expression in Bengali mainly focusing on Noun-Verb MWEs.

Fatima and Chaudhary (2010) [11] developed a method for extraction of trigram MWEs of Hindi using rule based approach by defining the set of rules based of grammatically relations. Shallow parser is used to distinguished grammatical relations and set of rules are applied to parsed output to extract trigram MWEs of specifics types such as noun compound and adjective-noun constructions.

Kishorjit and Bandyopadhyay (2011) [12] presented a method using genetic algorithm to choose the features of MWEs and CRF approach to automatically identify MWEs and named entities of morphological rich language, Manipuri. This method requires a large set of data to train the system to learn new instances of MWEs of different domains.

Kishorjit and Bandyopadhyay (2011) [13] presented a method for identifying of reduplicated MWEs in Manipuri using a rule based approach and reviewed all types of reduplicated MWEs found in Manipuri corpus.

3. Different types of MWEs in Indian Languages:

But there are some other types of MWEs which are not presented in English. These different types of MWEs in Indian Languages are given below:

(1) Replicated word: Most Indian Languages have replicated (repeated) words that have non-compositionality property. Mostly replicated words can be treated as MWEs. For example in Punjabi Language

ਰੋਜ਼ਰੋਜ਼ (Punjabi) Transliteration: *"Rōzrōz"*

Gloss: *Daily daily* Translation: *Every day*

ਹੋਲੀਹੋਲੀ (Punjabi) Transliteration: *"Hōlīhōlī"*

Gloss: *Slow Slow* Translation: *quite slowly*

Replicated words may contain a particle in between, For example

ਪਾਣੀਹੀਪਾਣੀ(Punjabi) Transliteration: *"Pāṇīhīpāṇī"*

Gloss: *water only water* Translation: *water all over*

Replicated words can be separated by hyphen sign '-' or without space as a singular word.

(2) Samaas and Sandhi: *Samaas* is a process to develop a new word by combination of two or more words by removing some particles. But *sandhi* is just joining two or more words to obtain a new word. In these pairs of words, second word may be antonym, hyponym, near to synonym, change in gender, change in number, etc. In these pairs, words may be separated by blank space, hyphen sign or without any space as a singular word.

Word combination with Antonym: In these pairs, the second words are antonym having opposite meaning of previous words. For example

ਦਿਨਰਾਤ (Punjabi) Transliteration: *"Din rāt"*

Gloss: *Day Night* Translation: *Day and Night*

ਹਾਰਜਿਤ (Punjabi) Transliteration: *"Hārjit"*

Gloss: *Loss Win* Translation: *Loss and Win*

Word combination with near to synonym: Second words in these pairs are synonym or near to synonym having same or related meaning of previous word. For example

ਦਾਲਰੋਟੀ(Punjabi) Transliteration: *"Dālrōṭī"*

Gloss: *Pulses Chapati* Translation: *Food*

ਪੂਜਾਪਾਠ(Punjabi) Transliteration: *"Pūjāpāṭha"*

Gloss: *Worship Lesson* Translation: *Worship*

Word combination with hyponym: In these second words are hyponym having same sound as previous words, but second words have no sense and these may or may not be presented in lexicons. For example

ਪਾਣੀਵਾਨੀ(Punjabi) Transliteration: *"Pāṇīvānī"*

Gloss: *Water Speech* Translation: *Water*

ਟੈਕ੍ਸਵੈਕ੍ਸ(Punjabi) Transliteration: *"TaixVaix"*

Gloss: *Tax Vaix* Translation: *Tax*

In these examples *vaani/speech* and *vaix* has no any sense.

Word combination with Gender/Number: In these pairs, the second words are change in gender or number of previous words. For example

ਮਾਂਬਾਪ(Punjabi) Transliteration: *"Māṁbāp"*

Gloss: *Mother Father* Translation: *Mother and Father*

ਦਿਨੇਦਿਨ(Punjabi) Transliteration: *"Dinō din"*

Gloss: *Days Day* Translation: *Day by day*

(3) Acronyms and Abbreviations: Deriving of acronyms and abbreviations in Punjabi is different from English. In English abbreviations may be derived by taking just first letter of each word, but in Punjabi by taking first letter along with vowel modifier. For example the Hindi acronym for *"Bhartiya JantaParti"* may be written as in English as B.J.P. or BJP (by taking first letter) and in Hindi it may be *Bha.Ja.Paa* or *Bhajapaa* (by taking first letter with vowel). All acronyms or abbreviations without dots are single words represent MWEs.

(4) Waala Morpheme Construct: *'waala'* has many morphological forms such as *'waalaa', 'waalii', 'waale'* or *'waalean'*. Any word combination with these *waala* morpheme construct can be candidates of MWEs. *Waala* morpheme can be last word or in between word of the construct. For example

ਕੰਮਵਾਲੀ(Punjabi) Transliteration: *"Kam wāli"*

Gloss: *Work waali* Translation: *Maid*

ਦੁੱਧਵਾਲਾ(Punjabi) Transliteration: *"Dudhwālā"*

Gloss: *Milk waala* Translation: *Milkman*

ਦੁੱਧਵਾਲੀਬਾਲਟੀ(Punjabi) Transliteration: *"Dudhawālībālaṭī"*

Gloss: *Milk waali bucket* Translation: *Milk bucket*

4. Identification, extraction and interpretation of MWEs in Punjabi

In this paper, we have considered only those MWEs that are particularly applicable to Punjabi. 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 Punjabi.

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-Punjabi corpus and a Punjabi 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; Punjabichunker 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):{bhagawaanraamkehaathon}(by Lord Ram) {mahaabaliiraavana}(mighty Ravan) {yuddhabhoomi men}(in battlefield) {maaradaalaagayaathaa}(had been killed). In chunking, firstly the verb group is identified. Since Punjabi 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 Punjabi 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 maaradaalaa 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 pluralsingular 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 aredefinite 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 => due to|of +ing

This rule when applied to the Punjabi sentence, vahdauratedauratethakgayaa (he run run tire went), yields the interpretation as 'He got tired of running'. For machine translation, the replicating words 'dauratedaurate' is substituted by a dummy variable (say 'dv1') with POS as an adverb and its value will be stored as 'of running'. The Punjabi sentence is modified to 'vah dv1 thakgayaa' 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 semionomatopoeia 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 "chaayavaaya". '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 verbin 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 Punjabi-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 Punjabi are mappings of English acronyms. Therefore, the individual Roman alphabet character mapping to Punjabi 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 Punjabi 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 Punjabi wordnet that has been used for some of its subclasses. The Punjabi 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	92.20%
replicating class	97.40%
doublet class	73.60%
‘vaala’ construct class	90.70%
Complex predicates and compound verbs	77.20%
acronym (with no dots) and named entity	27.50%

6. Conclusions and Discussions

In this paper, we have provided comprehensive details and characteristics of the MWEs that are specific to Punjabi. 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 Punjabi 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 Punjabi 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

[1] Sag, I. A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In *Computational Linguistics and Intelligent Text Processing*. Springer Berlin Heidelberg. pp. 1-15

[2] Poddar, L., Bhattacharyya, P. (2013, June). Multilingual Multiword Expression. Literature Survey Report. Department of Computer Science and Engineering. Indian Institute of Technology, Bombay

[3] Calzolari, N., Fillmore, C. J., Grishman, R., Ide, N., Lenci, A., MacLeod, C., &Zampolli, A. (2002, May). Towards Best Practice for Multiword Expressions in Computational Lexicons. In *LREC.* pp. 1934-1940.

[4] MunishMinia, Pushpak Bhattacharyya. (2012, June), Litrature Survey on Multi-Lingual Multiword Expressions. Literature Survey Report. Department of Computer Science and Engineering. Indian Institute of Technology, Bombay

[5] Sinha, R. M. K. (2011, June). Stepwise mining of multi-word expressions in Hindi. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*. Association for Computational Linguistics. pp. 110-115

[6] Brundage, J., Kresse, M., Schwall, U., &Storrer, A. (1992). *Multiword lexemes: A monolingual and contrastive typology for natural language processing and machine translation*. Technical Report 232, InstitutfuerWissensbasierteSysteme, IBM Deutschland GmbH, Heidelberg.

[7] Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational linguistics*, *16*(1). pp. 22-29.

[8] Smadja, F. (1993). Retrieving collocations from text: Xtract. *Computational linguistics*, *19*(1). pp. 143-177.

[9] Pavel Pecina. (2008). Lexical Association Measures: Collocation Extraction. PhD thesis,*Faculty of Mathematics and Physics, Charles University*, Prague, Czech Republic.

[10] Agarwal, A., Ray, B., Choudhury, M., Sarkar, S., &Basu, A. (2004, December). Automatic Extraction of Multiword Expressions in Bengali: An Approach for Miserly Resource Scenario. In *Proceedings of International Conference on Natural Language Processing (ICON)*. pp. 165-174.

[11] Fatima, Z., & Chaudhary, N. (2010, October). Extracting Hindi Multiword Expressions Using a Rule Based Tool. In *Proceedings of the 2010 International Conference on*

Advances in Communication, Network, and Computing. IEEE Computer Society. pp. 434-438.

[12] Nongmeikapam, K., & Bandyopadhyay, S. (2011). Genetic algorithm (GA) in feature selection for CRF based manipuri multiword expression (MWE) identification. International Journal of Computer Science & Information Technology (IJCSIT) 3(5). pp.53-66.

[13] Nongmeikapam, K., & Bandyopadhyay, S. (2010). Identification of Reduplicated MWEs in Manipuri: A Rule Based Approach. In Proceedings of ICCPOL 2010. Redwood City, San Francisco, USA. pp. 49-54

[14] Manning, C. D., &Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT press.

[15] Jackendoff, Ray (1997). *The Architecture of the Language Faculty*. Cambridge, MA: MIT Press.

[16] Fellbaum, C. (1998): WordNet: An Electronic Lexical Database. Cambridge, Mass.: MIT Press

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

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

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

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

[21] Goyal, V., &Lehal, G. S. (2011, June). Hindi to Punjabi machine translation system. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Systems Demonstrations. Association for Computational Linguistics. pp. 1-6.

