Multiword Expressions (MWEs): A Challenging Task of Natural Language Processing

Kapil Dev Goyal *(Assistant Professor)*

SBAS Khalsa College, Sandaur (Sangrur).

Email: kapildevgoyal@gmail.com

Abstract

Identification and extraction of Multiword Expressions (MWEs) is very hard task faced by various Natural Language processing applications like Information Retrieval (IR), Information Extraction (IE), Question-Answering systems, Speech Recognition and Synthesis, Text Summarization and Machine Translation (MT) are highly affected by MWEs. In this paper we discuss various features and types of Multiword Expressions (MWEs) of Punjabi language and how Punjabi MWEs are different from English MWEs.

Keywords: *Multiword Expression, Collocation, Corpus.*

Introduction

When two or more constituent words together represent an expression and the proper meaning of that expression cannot be extracted from individual meaning of each word, then it is mandatory to identify these constituent words to improve the quality of many natural language processing applications. Otherwise if any system considers these expressions as separate words, then system will skip the basic meaning of the combined words and the result will be different from the actual meaning.

E.g. *Kick the bucket*

Gloss: ਬਾਲਟੀ ਨੂੰ ਲੱਤ ਮਾਰਨਾ

Translation: ਮਰਨਾ

E.g. ਅੱਖਾਂ ਦਾ ਤਾਰਾ (Punjabi)

Transliteration: *"Akhān dā tārā"*

Gloss: *Star of Eyes*

Translation: *Lovely*

In these examples each constituent word has a different meaning from the collective meaning of all separated words. These types of words are called Multiword Expression and it is mandatory to identify them in all Natural Language applications.

The major problem in NLP is to identify and extract Multiword Expressions (MWEs). Proper extraction and interpretation of MWEs is very important for most of the NLP tasks. MWEs may exist in all types of natural language and poses major problems in all kinds of NLP application (Sag et. al 2002)[1]. However, Identification and extraction of MWEs is most crucial, but hard to identify these Multiword Expressions (Calzolari et al. 2002)[3].

Related Work

Baldwin et al. (2010) [1]; Lahari Poddar [2]; Munish Minia [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.

Features of MWEs: (Manning and Schutze 1999:184)[14] described that non-compositional, non-modifiable and non-substitutable are basic features of MWE.

Non-compositional: It means that MWE cannot be completely translated from the meaning of its parts.

E.g. ਅੱਖਾਂ ਦਾ ਤਾਰਾ (Punjabi) **Transliteration:** *"Akhān dā tārā"*

Gloss: *Star of Eyes* **Translation:** *Lovely*

E.g. लोहे के चने चबाना (Hindi) **Transliteration:** *"Lōhē kē chanē chabānā"*

Gloss: *To chew iron gram* **Translation:** *Difficult task*

In above examples, actual translations cannot be predicted from their parts, which are completely different from its basic meaning.

Non-modifiable: Many Multiword Expressions are frozen and they cannot be changed in any way. These types of expressions cannot be modified by grammatical transformations (like by changing Number/ Gender/ Tense, addition of adjective etc).

Eg. In ਰੋਜੀ ਰੋਟੀ *(Rōjī rōtī)* cannot be changed in number as ਰੋਜੀ ਰੋਟੀਆਂ *(Rōjī rōtī'ān)*

Non-Substitutable:- Any word of Multiword Expression cannot be substituted by one of its synonym without affecting the meaning of an expression.

E.g. ਰੋਜੀ ਰੋਟੀ *(Rōjī rōtī)* cannot be written as ਰੋਜੀ ਖਾਣਾ *(Rōjī khānā)* or ਰੋਜ ਰੋਟੀ *(Rōj rōtī)*

MWES are not rare: MWEs are not rare, they are frequently occur in all natural languages. It's assumed that MWEs are at least as many as numbers of single words (Jackendoff 1997:156)[15]. About half of the entries in the semantic online lexicon WORDNET are MWEs (Fellbaum 1998)[16].

How Punjabi Language is different from English Language: In fixed order language like English Language, two or more words that co-occur more often than chance can be MWE. But in partially free-ordered languages like Punjabi, if two or more words co-occur frequently, they need not be MWE, it may depend upon the surrounding context. Consequently, the identification and extraction of MWEs in Punjabi text is also very challenging task.

Unlike in English Language, Punjabi has several difficulties to identify correct MWEs because of morphologically rich language and partially free ordered language. Therefore the word can occur at any place of sentence without affecting the actual meaning of sentence. Another problem with morphological rich language is that they do not have capitalization information. Capitalization information makes easier to identify Named Entities (NE), which may be Multiword Expression.

Beside MWEs, there is one more related term known as Collocation. However, Collocations do not always represent the same range of MWE characteristics.

Collocation: A collocation is combination of two or more words whose semantic and/or syntactic characteristics cannot be fully derived from its individual words (Evert 2004)[17]. Collocations are frequently co-occurring words independent of any their semantics (Moon 1998)[18]. Collocation is conventional way to say something. (Manning and Schutze 1999:151)[14].

Difference between Collocations and MWEs: A collocation is group of two or more words that co-occur (recurrent word combination) more often than expected by chance and they are often different from idioms or non compositional phrases. But non-compositional is one of basic feature of MWEs.

Types of MWE/Collocation: MWEs are classified in **Lexical phrases** and **Institutionalized Phrases** according to their lexical and semantic properties which are explained as follows:

(1) Lexical Phrases: Lexical phrases are MWEs that have syntactically idiosyncratic or semantic in their parts or in their combination. These types of phrases are idiosyncratic and added meaning to their structures. Lexical phrases are further divided into 3 parts

- Fixed Expressions
- Semi-Fixed Expressions
- Syntactical-Flexible Expressions

(a) Fixed Expression: Fixed Expressions are syntactically fixed that are completely frozen and neither undergo morph-syntactic variation nor internal modification. Eg. *by and large, ad hoc, of course*, etc.

(b) Semi-Fixed Expression: Semi-Fixed Expressions can undergo lexical variation but they hard restriction on word order and composition. They do not allow any syntactical variation. For example *"Kick the bucket"* can be written as *"Kicks the bucket"*, *"Kicking the bucket"*, *"Kicked the bucket",* etc. Semi-Fixed Expressions are further divided into following subparts.

- Non-Decomposable Idioms
- Compound Nominals
- Name Entities

i. Non-Decomposable Idioms: On the basis of semantic composition, Idioms can be divided into two types: **Decomposable Idioms** and **Non-Decomposable Idioms.**

In decomposable idioms, semantic of overall idioms can be derived from their parts. For example in *"Spill the beans"* idiom, meaning of *"spill"* is *"reveal"* and beans means *"Secret"*. The combined meaning of its parts is *"reveal the secret"*, which is same as meaning of overall idioms. But in the case of non-decomposable idioms such analysis is not possible. In non-decomposable idioms, the semantic of whole idiom cannot be completely derived from its parts. For example in *"Kick the bucket"* idiom which means is *"to die"* and cannot derive from its parts.

Non-decomposable idioms are semi-fixed expression. Due to their restrict word order they cannot undergo any syntactic variation but some lexical variation might allow. For example *"Kick the bucket"* can be written as "*Kicking the bucket"*

ii. Compound Nominals: Compound Nominals are similar to the non-decomposable idioms in that there are restricted syntactically terms. But in case of compound nominals expression,

they allow lexical inflection for number. So they can be written as singular or plural forms. For example "*Railway station"* can be written as "*Railway Stations"*

iii. **Named Entities:** Named Entities are highly syntactically idiosyncratic. Names of persons/objects or places are basically named entities. For example *"Boota Singh", "New Delhi", "Knight Riders"*, etc.

Named Entities in English basically represent by capital letters, but in Punjabi it is very hard task to identify them due to lack of capitalization.

(c) Syntactically-Flexible Expression: Unlike semi-fixed expression, syntactically flexible expressions permit to make syntactical variation. Syntactically variation can be possible in following 3 types:

- Verb-Particle Construction
- Decomposable Idioms
- Light-Verb Construction

i. **Verb-Particle Construction:** A combination of main verb and one or more particles/prepositions is called verb particle constructions or phrasal verb. For example come down, look up, etc. A combination of VPC with noun object is called transitive VPC. Noun particle may be placed either between or following the verb and particle(s). For example *"Call Ajay up"* or *"Call up Ajay"*. VPC combination with adverbs is called intransitive VPC, adverbs can often be placed in between of verb and particle(s). (For example fight bravely on). It is very hard task to identify complete range of transitive VPC as words-with-space.

ii. **Decomposable Idioms:** These types of idioms are syntactically flexible to some degree and behave like semantically linked parts. For example *"Let the cat out of the bag"* is syntactically flexible idioms and can be written as *"The cat was let out of the bag"* and *"sweep under the rug"* can be written as *"The whole issue was swept under the rug"*, etc.

iii. **Light-Verb Construction:** It is a combination of verb with noun where the noun usually taken in literal sense but the verb usually loses its original meaning. For example to *"give a lecture", "make a decision", "take a picture"* etc. LVC are highly idiosyncratic and it is very difficult to identify which light verb combines with a given noun.

(2) Institutionalized Phrases: These types of phrase are fully syntactically and semantically compositional, but statistically idiosyncratic. Consider an example of *"traffic light". "Traffic light"* is statistical idiosyncratic rather than linguistic, due to its relatively high frequency than its any alternative compositions (*"Traffic descriptor", "traffic director"*).

Different types of MWEs in Punjabi 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

ਰੋਜ਼ ਰੋਜ਼ (Punjabi)	**Transliteration:** *"*Rōz rōz*"*
Gloss: *Daily daily*	**Translation:** *Every day*
ਹੋਲੀ ਹੋਲੀ (Punjabi)	**Transliteration:** *"*Hōlī hōlī*"*
Gloss: *Slow Slow*	**Translation:** *quite slowly*

Replicated words may contain a particle in between, For example

ਪਾਣੀ ਹੀ ਪਾਣੀ (Punjabi)	**Transliteration:** *"*Pānī hī pānī*"*
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.

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

ਦਿਨ ਰਾਤ (Punjabi) **Transliteration:** *"*Din rāt*"*

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

ਹਾਰ ਜਿਤ (Punjabi) **Transliteration:** *"*Hār jit*"*

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

(b) 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

ਦਾਲ ਰੋਟੀ (Punjabi) **Transliteration:** *"*Dāl rōṭī*"*

Gloss: *Pulses Chapati* **Translation:** *Food*

ਪੂਜਾ ਪਾਠ (Punjabi) **Transliteration:** *"*Pūjā pātha*"*

Gloss: *Worship Lesson* **Translation:** *Worship*

(c) 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

ਪਾਣੀ ਵਾਨੀ (Punjabi) **Transliteration:** *"*Pāṇī vānī*"*

Gloss: *Water Speech* **Translation:** *Water*

ਟੈਕ੍ਸ ਵੈਕ੍ਸ (Punjabi) **Transliteration:** *"Taix Vaix"*

Gloss: *Tax Vaix* **Translation:** *Tax*

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

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

ਮਾਂ ਬਾਪ (Punjabi) **Transliteration:** "Māṁ bāp"

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

ਦਿਨੇ ਦਿਨ (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 Janta Parti"* 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

ਕੰਮ ਵਾਲੀ (Punjabi) **Transliteration:** *"Kam wāli"*

Gloss: *Work waali* **Translation:** *Maid*

ਦੁੱਧ ਵਾਲਾ (Punjabi) **Transliteration:** *"Dudh wālā"*

Gloss: *Milk waala* **Translation:** *Milkman*

ਦੁੱਧ ਵਾਲੀ ਬਾਲਟੀ (Punjabi) **Transliteration:** *"Dudha wālī bālatī"*

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

Conclusion: Identification and extraction of MWEs is an importance task of many NLP application, if any system ignore the importance of MWEs, then overall performance and accuracy of that system will be reduced and results may be different. But identification and extraction of MWEs is very task. It has been discussed that there are multiple types of MWEs, Therefore no single algorithm or single approach cannot identify all types of MWEs. Punjabi language has another different types of MWEs which were not presented in English language such as Replicated words, Samaas and Sandhi, Waala Morpheme, Punjabi acronyms and abbreviations.

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] Munish Minia, 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, Institut fuer Wissensbasierte Systeme, 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] Evert, S., Heid, U., & Spranger, K. (2004). Identifying Morphosyntactic Preferences in Collocations. In *LREC*. pp. 907-910.

[18] Moon, Rosamund 1998. Fixed Expressions and Idioms in English: A Corpus-Based Approach. Clarendom Press, Oxford.