Multiword Expressions in Indian Languages

Kapil Dev Goyal

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

ARTICLE DETAILS

Article History
Received: 03 March 2017
Accepted: 18 March 2017
Published Online: 23 March 2017

Keywords:
Multiple Expression, Collocation, Machine Translation, Natural Language Processing, Corpus, Parallel Corpus

ABSTRACT

This paper presents a study of Multiword Expression for Punjabi Language and how Multiword Expression is a challenging task of Natural Language Processing? This paper also discusses features, methods for identification and extraction of MWE, multiple types and challenges of Multiword Expressions. Any system cannot ignore the importance of MWEs, otherwise results of that system may be different from actual results.

1. INTRODUCTION

Multi-word Expressions (MWEs) are a combination of two or more orthographic words and complete meaning of MWEs cannot be predicted from meaning of its individual words. Multiword Expression is combination of two or more words but it treated as a single word. Multiwordexpressions appear frequently in human language, in anykind of phrases.

Many times during translation of sentence into another language, due to MWEs, source sentence is parsed in multiple words instead of single word and always translate meaningful and sometimes nonsense result. Many common names are also called as MWE such as Middle school, Elementary School, Petrol pump, Smart Bank, etc. Similarly, many verbal MWE such as call off, make up mind to lose weight, take the benefit of the doubt.

Identification and extraction of Multiple Expressions is very hard task for 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.Identification and extraction of MWEs is most crucial, but hard to identify these Multiword Expressions (Calzolari et al. 2002)[3].

1.1 MWEs are not rare:

MWEs are frequently occurs in all natural languages. It’s assumed that MWEs are at least as many as numbers of single words (Jackendoff1997:156)[15]. About half of the entries in the semantic online lexicon WORDNET are MWEs (Fellbaum1998)[16].

1.2. 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 or Collocation. 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.

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 contract, etc.


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 Pointwise 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. FEATURES OF MWEs

(Manning and Schutze 1999:184) [14] presented the basic properties of MWEs are non-compositional, non-modifiable and non-substitutable.

3.1. Non-compositional: It means that MWE cannot be completely translated from the individual meaning of its parts. Individual meaning its parts is different from whole meaning of MWEs.

E.g. 

粛男Indicator

Transliteration: “Akhāndātārā”
Gloss: Star of Eyes
Translation: Lovely

E.g. लोहेकेचनेचचबाना (Hindi)
Transliteration: “Lōhēkēchanēchabānā”
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.

3.2. Non-modifiable: These types of Multiword Expressions are partially and completely fixed (frozen) and they can never change in any way. These types of expressions cannot be altered by grammatical transformations (like by changing Number/ Gender/ Tense, addition of adjective etc).

E.g. In रोजरोती (Rōjrōti) cannot be changed in number as रोजरोती (Rōjrōti) cannot be written as रोजरोती (Rōjrōti) or रोजरोती (Rōjrōti)

3.3. Non-Substitutable: In these types of MWEs are any word cannot be substituted by one of its synonym or related word/words without affecting the meaning of an expression.

E.g. रोजरोती (Rōjrōti) cannot be written as रोजरोती (Rōjrōti) or रोजरोती (Rōjrōti)

4. IDIOSYNCRASIES OBSERVED IN MWE

Most important types of idiosyncrasies are explained below:

4.1. Semantic Idiosyncrasy: One of the important classes of MWEs is Semantic idiosyncrasy in which meanings of whole MWEs are not predicted from the composition of the translation of their constituent words.

E.g. ਲੋਹੇ ਕੇਚਨ ਚੱਬਾਨਾ (Punjabi)
Transliteration: “Akhāndātārā”
Gloss: Star of Eyes
Translation: Lovely

4.2. Statistical Idiosyncrasy: Statistical idiosyncrasies are basically collocations which are combination of words that co-occur more frequently than expected by chance. They are fully compositional (can be predicted from the composition of the translation of their parts) and syntactically (cannot change their words order). E.g. “Traffic Light” and “Traffic Lamp” having same meaning but “Traffic Light” is more frequent than “Traffic Lamp”.

4.3. Lexical Idiosyncrasy: Lexical idiosyncrasy is most straight forward properties of MWEs in which one or more components are not part of Lexicon. For example “ad hoc” is Lexical idiosyncrasy where “ad” and “hoc” both are not part of English lexicon.

4.4. Syntactic Idiosyncrasy: These types of MWEs do not follow the rules of the conventional grammar. E.g. “By and large”, here ‘by’, ‘and’ and ‘large’ are preposition, conjunction and adjective respectively (ByPREPandCONJlargeADJ). It is very difficult to identify these types of expressions and what are the rules to identify them.

5. COLLOCATION

There is one correlated term same as Multiple Expression. Combination of more than one word with blank space i.e. multiword expression is also known as collocations. Collocations are frequent co-occurrence of their components. 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. [17]
- Multiword expressions are expressions consisting of two or more words that correspond to some conventional way of saying things. [14]
- 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.[18]
- Chunk as Idiosyncratic interpretations that cross word boundaries (or spaces) [1], 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.

5.1. Multiword True Collocation

Collocation refers to a relationship between words that frequently occurs with each other. The words whose are frequently co-occurs together canmean more than the sum of their parts (Traffic Light, Railway Stations). 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.

6. ENGLISH-PUNJABI PARALLEL CORPUS

A corpus is huge collection of text and it is electronically stored and processed a large and structured set of texts. Corpus can be used for statistical analysis and hypothesis testing and It is also useful to check linguistic rules.

There is no availability of Parallel English-Punjabi corpus and development of English Punjabi parallel corpus is itself is very challenging task, because of limited source of English Punjabi electronically translated text. In this paper, an alternative method has used to create English-Punjabi parallel corpus by using Hindi to Punjabi machine translation machine [21]. The motivating factor behind this idea is that there is large collection of English-Hindi parallel text available from various resources and good result of Hindi-Punjabi Machine translation system.

Mainly three English-Hindi parallel corpuses are used in this study. HindEnCorp[20], bharatdharshan English-Hindi Panchtantra story corpus[19] and English-Hindi Bible.

7. 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.

7.1 Statistical Method: There are various statistical methods are used to measure word association (chunk) and then MI (mutual information) as an objective are measured for estimating word association norms. These 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.

7.2 Linguistic Method: There are various linguistic methods are used such as POS tagger, Chunker, Morphological analyser, etc are methods for MWE extraction. Both grammatical and syntactical requirements are there for a MWE.

7.3 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.

7.4 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.

8. CHALLENGES

There is no single method or algorithm to identify, extraction, presentation in text because of its multiple types. There are number of challenges in NLP tasks and applications [14]like translation, word-sentence alignment, inpreprocessing, incompleteness, etc.

8.1. Poor quality of the parallel English-Punjabi corpus
It has been earlier discussed that there is no availability of Parallel English-Punjabi corpus and development of English Punjabi parallel corpus is itself is very challenging task, because of limited source of English Punjabi electronically translated text. In this paper, an alternative method has used to create English-Punjabi parallel corpus by using Hindi to Punjabi machine translation machine [21] and quality of parallel English-Punjabi corpus is poor.

8.2. Sentence alignment errors: There are various automatic sentence aligner are available to automatically align English Punjabi translated text. These all are purely statistical sentence aligner 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 oneto-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.

8.3. 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. Oftentimes 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 incases where the word aligner proposes N:1 alignment; usually these N words contain the correct sequence or apart of the sequence and the correct analysis of the bidirectional alignments may help filter out the incorrect parts (i.e., the analysis of the intersection of N and M sequences, where M:1 is Punjabi-to-English and N:1 is English-to-Punjabi alignments detected by the word alignment tool).

8.4 Noise introduced by preprocessing: Errors caused by morphological analysis and disambiguation tools may lead to wrong tokenization and 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 wordpairs cannot be found in the dictionary. An additional source of errors stems from languagespecific differences in word order between the languages.

8.5. Incomplete dictionary: There is no limit of words as language evolves new words. Dictionary cannot be completed. 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.

9. 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.

REFERENCES