EXTRACTION OF REPLICATED PUNJABI MULTIWORD EXPRESSIONS

[1]Kapil Dev Goyal, Research Scholar, Department of Computer Science, Punjabi University, Patiala. Email: kapildevgoyal@gmail.com

[2]Vishal Goyal, Professor, Department of Computer Science, Punjabi University, Patiala. Email: vishal.pup@gmail.com

ABSTRACT

Multiword Expressions (MWEs) play a vital role in Natural Language Processing. Multiword Expression is a combination of two or more words but treated as a single word. In Punjabi Language, there are varieties of MWEs and many of these are of the types that are not found in English. In this paper, we discuss different types of MWEs encountered in Punjabi. For example, replicated words, word combination with antonym, synonym, hyponym, gender, number and 'waala' morpheme have not been discovered as MWEs in English. Rule based approachs, statistical methods, and linguists' approaches were used for MWE identification and extraction. In this paper, we present a methodology for identification and extraction of Punjabi MWEs using statistical methods, rule base methods and linguists' approach.

INTRODUCTION

The identification and extraction of multi-word expressions (MWEs) is an important and hard task of all NLP tasks such as machine translation, information retrieval, question-answering etc. MWEs are combination of two or more words but consider as a single word. The individual meaning of each word will be different from collective meaning of MWEs.

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

These MWEs is treated as a single word with a single part of speech (POS) tagging, translation system and mostly all NLP applications. Our main task is to identify and extract replicated MWEs in Punjabi, which are not detected in English. No much work has been reported on Punjabi MWE.

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

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

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:** *"Pāṇī hī pāṇī"*

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 pāṭha*"*

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) 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 bālaṭī"*

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

RELATED WORK

(Brundage et al., 1992) discussed the characteristics of MWEs such as non-compositionality, non-substitutability and non-modifiability.

(Baldwin & Kim, 2010; Minia, 2012) presented an excellent review on Multiword Expression. They almost covered all important features of MWEs such as characteristics of MWEs, types of MWEs, extraction techniques, etc. Authors also explored some techniques using statistical approaches, linguistic approaches and rule based approaches.

(Poddar, 2013) also discovered a brilliant review on Multiword Expression. Author 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.

(Sinha, 2009) discoursed different types of MWEs detected in Indian Languages such as Hindi such as Replicating words, Samaas and Sandhi, Hindi acronyms and abbreviations, vaala morpheme contruct, etc.

(Church & Hanks, 1989; Pecina, 2009; Smadja, n.d.) coined an automatic extractor of MWEs using statistical approaches such as Point-wise Mutual Information (PMI) and other statistical hypothesis tests by measuring association between them.

(Agarwal et al., n.d.) coined a method to automatic extraction of Multi-word expression in Bengali mainly focusing on Noun-Verb MWEs.

(Fatima et al., n.d.) extracted trigram MWEs of Hindi language using rule-based approach by defining the set of rules based of grammatically relations. To distinguished grammatical relations and set of rules a Shallow parser is used.

(Singh et al., 2011) proposed a method to select the features of MWEs and CRF approach to automatically identify MWEs and named entities of Manipuri language using genetic algorithm. A huge set of data is required to train the system to learn new instances of MWEs of different domains and also discussed a method using a rule based approach for identifying and extracting of reduplicated MWEs in Manipuri and reviewed all types of reduplicated MWEs found in Manipuri corpus.

Above different researchers identified and extracted MWEs for different languages, but there is a no work done reported on Punjabi language in this field, so we are using our own hybrid approach for extraction of reduplicated MWEs in Punjabi languages.

IDENTIFICATION, EXTRACTION AND INTERPRETATION OF MWES IN PUNJABI

In this paper, following types of replicated MWEs are identified and extracted

1) Pure replicated words
2) Hyphen Separated Replicated Words
3) Particle containing replicated words
4) Word combination with Antonym
5) Word combination with hyponym
6) Word combination with near to synonym
7) Word combination with Gender/Number
8) 'Waala' Morpheme Construct

In all above-mentioned types, Rules based and statistical approaches are used to identify and extract replicated Punjabi Multiword Expressions.

1) Pure replicated words: To identify and extract pure replicated words, rule-based approach is used. In this approach two consecutives words are compared for equality. If these

consecutives words are equal, then they can be candidate of MWEs. The process is explained as below:

1) In first step, we generate list of all bi-grams of words.
2) Then iteratively take one bi-gram at a time and repeat steps 3 to 4.
3) Split each bigram and save to array BIGRAM
4) If BIGRAM[0] is equal to BIGRAM[1], then
   Save to output list
5) Print output list

2) Hyphen Separated Replicated Words: In case Hyphen separated MWEs, only those words are considered which contain hyphen. The extraction process is explained below:
   1) In first step, we generate list of all unigrams.
   2) Then iteratively take one unigram at a time and repeat steps 3 to 5.
   3) If unigram does not contain hyphen, then continue with next iteration. Otherwise goto step 4.
   4) Split each unigram with hyphen character and save to array UNIGRAM
   5) If UNIGRAM[0] is equal to UNIGRAM[1], then
      Save to output list
   6) Print output list

3) Particle containing replicated words: If some combination of words contains some special particles, then they can be candidate of MWEs. In this case we considered three consecutives words, in which middle word is particle and compared left and right words. List of Punjabi particle that can separate replicated words is given below:

   ਤੇ, ਨਾ, ਦੇ, ਹੀ, ਤੋਂ, ਦਰ. The identification and extraction process is given below:

   1) In first step, we generate list of all trigrams and stores all particles in PARTICLE list.
   2) Then iteratively take one trigram at a time and repeat steps 3 to 4.
   3) Split each trigram with space and save to array TRIGRAM
   4) If UNIGRAM[0] = UNIGRAM[2] and PARTICLE list contains UNIGRAM[1] then

Save to output list

5) Print output list

4) Word combination with Antonym: If two consecutives words are antonym to each other, then they can be candidate of MWEs. There is no any algorithm exists to find out antonym. Therefore, manually a database of 500 Punjabi antonyms was created and used in this approach. The complete process is given below:
   1) In first step, we generate list of all bigrams and stores all Punjabi antonym in ANTONYM dictionary as a key-value pair.
   2) Then iteratively take one bigram at a time and repeat steps 3 to 4.
   3) Split each bigram with space and save to array BIGRAM
   4) If BIGRAM[0] and BIGRAM[1] are in Key-Value pair in ANTONYM dictionary, then

      Save to output list
   5) Print output list
5) Word combination with hyponym: In this category, Minimum edit distance (MED) and Pointwise Mutual information (PMI) statistical approaches are used to identify and extract these types of MWEs. In these types of expressions, mostly places of Gurmukhi matras were fixed, only few Gurmukhi letters were changed. For examples, in case of ਪਾਣੀ ਵਾਨੀ (Punjabi) and ਟੈਕ੍ਸ ਵੈਕ੍ਸ (Punjabi), places of Gurmukhi matras were unchanged in each consecutive word, but some of the Gurmukhi letters were changed. To compare the pattern of Gurmukhi matras, regular expressions were used and to compare the Gurmukhi letters, MED approach was used. The complete process of step by step is given below
   1) In first step, we generate list of all bigrams.
   2) Then iteratively take one bigram at a time and repeat steps 3 to 6.
   3) Split each bigram with space and save to array BIGRAM
   4) If Minimum edit distance of BIGRAM[0] and BIGRAM[1] is less than 20, then Go to step 5, Otherwise continue with next iteration.

5) Extract the Gurmukhi Matras pattern of BIGRAM[0] and BIGRAM[1] using regular expressions. If both patterns are matched, then go to step 6, otherwise continue with next Bigram.
6) Now measures the association between BIGRAM[0] and BIGRAM[1] using pointwise mutual information (PMI) statistical tool. If PMI score is greater than 0, then there is association between BIGRAM[0] and BIGRAM[1] and save to output list. Otherwise continue with next iteration.
7) Print output list

6) Word combination with near to synonym: In this category, Minimum edit distance (MED) and Pointwise Mutual information (PMI) statistical approaches are used to identify and extract these types of MWEs. Unlike hyponym, there is no similar patterns of Gurmukhi Matras, but most of the letters are similar. For examples, in case of ਰਾਤੋ ਰਾਤ, ਇਧਰ ਉਧਰ, there is no similarity of Gurmukhi matras, but most of the Gurmukhi letters are matched. The process of step by step is given below:
   1) In first step, we generate list of all bigrams.
   2) Then iteratively take one bigram at a time and repeat steps 3 to 5.
   3) Split each bigram with space and save to array BIGRAM
   4) If Minimum edit distance of BIGRAM[0] and BIGRAM[1] is less than 10, then Go to step 5, Otherwise continue with next iteration.
   5) Now measures the association between BIGRAM[0] and BIGRAM[1] using pointwise mutual information (PMI) statistical tool. If PMI score is greater than 0, then there is association between BIGRAM[0] and BIGRAM[1] and save to output list. Otherwise continue with next iteration.
   6) Print output list
7) Word combination with Gender/Number: In these types of MWEs second word can be Gender or Number. For example ਮਾਂ ਬਾਪ, ਦਿਨੇ ਦਿਨ. Like antonym, there is no any

algorithm exists to find out Gender/Number. Therefore, manually a database of 800 Punjabi Genders and Numbers were created and used in this approach. process is given below:

1) In first step, we generate list of all bigrams and stores all Punjabi Gender in GENDER dictionary and Numbers in NUMBER dictionary as a key-value pair.
2) Then iteratively take one bigram at a time and repeat steps 3 to 4.
3) Split each bigram with space and save to array BIGRAM
4) If BIGRAM[0] and BIGRAM[1] are in Key-Value pair in GENDER or NUMBER dictionaries, then

   Save to output list
5) Print output list

8) 'Waala' Morpheme Construct : Word combination with 'waala' morpheme (ਵਾਲਾ ਵਾਲੇ ਵਾਲੀ ਵਾਲਿਆਂ ਵਾਲੀਆਂ) can be candidate of MWEs. In this category, we generate a list of bigrams and trigrams. In bigram, first word will be Noun and second word will be 'waala' morpheme. Similarly, in trigram, first and last word will be Noun and middle word will be 'waala' morpheme. To check whether given word is Noun or not, Rule based Part of Speech (POS) was used. The step by step working of process is given below:

1) First of all, we generate list of all 5-grams, because Punjabi POS tagger will not produce good results for less than 5 words, and stores all Punjabi 'waala' morpheme words in WAALA list.
2) Then iteratively take one 5-gram at a time and repeat steps 3 to 7.
3) Pass each 5-gram to Punjabi POS tagger and store output of POS tagger to POSSTR string.
4) Split each POSSTR with space and save to array POSARRAY
5) If WAALA list contains POSARRAY[2] words and POSARRAY[1] and POSARRAY[3] are nouns, then concatenate POSARRAY[1], POSARRAY[2] and POSARRAY[3] with space.

result = POSARRAY[1] + " " + POSARRAY[2] + " ' + POSARRAY[3]

then result save to output list.

6) Otherwise, If WAALA list contains POSARRAY[2] words and only POSARRAY[1] is noun, then concatenate POSARRAY[1] and POSARRAY[2] space.

result = POSARRAY[1] + " " + POSARRAY[2]

then result save to output list.

7) Otherwise continue with next 5-grams.
8) Print output list

EXPERIMENTATION AND RESULTS

First of all, Punjabi corpus is required for experiment. We created corpus from various sources such as Punjabi newspaper http://epaper.ajitjalandhar.com, https://epaper.punjabitribuneonline.com, Emille Project http://www.emille.lancs.ac.uk and various e-books. We considered approximately 5000 words as a sample. Table 1 shows the f-score results. The f-score varied from 59% to 96%. The identification of "waala" morpheme is very poor than identification of Pure replicated words. The performance of the Waala Morpheme MWEs identification is affected due to poor performance of Rule base POS tagger and performance of the antonyms, gender and numbers MWEs identification is affected due to inadequacy of the Punjabi wordnet.

Table 1: Experimental results

MWE Type	F-score
Pure replicated words	96.06%
Hyphen Separated Replicated Words	95.99%
Particle containing replicated words	94.34%
Word combination with Antonym	60.50%
Word combination with hyponym	75.10%
Word combination with near to synonym	74.90%
Word combination with Gender/Number	59.68%
'Waala' Morpheme Construct	48.88%

CONCLUSIONS AND DISCUSSIONS

In this paper, we have identified and extracted different types of Punjabi MWEs which were not encountered in English language. In this paper, we discussed step by step processes to identify and extract MWEs of various types such as pure replicated words, hyphen separated replicated words, particle containing replicated words and many more. All these algorithms are implemented in PHP and achieved f-score value varied from 59% to 96%. In future work, we would like to extract parallel extraction of MWEs from Punjabi-English parallel corpus.

REFERENCES

Agarwal, A., Ray, B., Choudhury, M., Basu, A., & Sarkar, S. (n.d.). Automatic Extraction of Multiword Expressions in Bengali: An Approach for Miserly Resource Scenarios. In *academia.edu*. Retrieved August 31, 2020, from http://www.academia.edu/download/30405011/icon2004_mwe.pdf

Baldwin, T., & Kim, S. N. (2010). Multiword expressions. *Handbook of Natural Language Processing, Second Edition*, 267–292.

Brundage, J., Kresse, M., Schwall, U., & Storrer, A. (1992). *Multiword lexemes: A monolingual and contrastive typology for natural language processing and machine translation*.

Church, K. W., & Hanks, P. (1989). *Word association norms, mutual information, and lexicography*. *April*, 76–83. https://doi.org/10.3115/981623.981633

Fatima, Z., 2010, N. C.-P. of the, & 2010, undefined. (n.d.). Extracting Hindi Multiword Expressions Using a Rule Based Tool. *IEEE Computer Society*.

Minia, M. (2012). *Literature Survey on Multi-Lingual Multiword Expressions*.

Pecina, P. (2009). Collocation Extraction AND THEORETICAL LINGUISTICS. In *Studies in computational and theoretical linguistics*.

Poddar, L. (2013). Multilingual Multiword Expressions. *Detection of MultiWord Expression and Name*

Entity Recognition, *113050029*.

Singh, N. B., Bandyopadhyay, S., Nongmeikapam, K., Laishram, D., & Mayekleima Chanu, N. (2011). Identification of Reduplicated Multiword Expressions Using CRF. *LNCS*, *6608*(PART 1), 41–51. https://doi.org/10.1007/978-3-642-19400-9_4

Sinha, R. M. K. (2009). *Mining complex predicates in Hindi using a parallel Hindi-English corpus*. *August*, 40. https://doi.org/10.3115/1698239.1698247

Smadja, F. (n.d.). Retrieving Collocations from Text: Xtract. In *dl.acm.org*. Retrieved August 31, 2020, from https://dl.acm.org/doi/abs/10.5555/972450.972458