International Journal of New Technology and Research (IJNTR)
ISSN:2454-4116, Volume-2, Issue-4, April 2016 Pages 18-21

System for Identification and Analysis of Reduplication Words in Hindi Corpus

Dr. (Mrs.) Kamlesh Dutta, Anshul Jindal

***Abstract*— Reduplication words is a class of MWE which is rapidly expanding due to the continuous need for coinage of new terms for describing new concepts, such as multi word expression, gold standard, and web page. Identification of reduplication words can particularly help parsing, and dictionary based applications like machine translation, and cross lingual information retrieval, since such word sequences should be treated as a single unit. The purpose of our work is to come up with a list of potential reduplication MWEs which a lexicographer can look at and decide whether a given word sequence should be added to the lexicon. This will aid the construction of a quality lexicon which incorporates MWE entries.**

A system is to be developed which is focused on first extracting reduplication words from the given text and then identify them into different categories based upon their semantic and syntactic analysis. The system should store different categories words into different files based upon their classification. The approach used in identification of reduplication words is that the two hyphen separated words are first translated to English language and then they are compared from the backside. Depending upon the degree of similarity they are classified into different categories of reduplication

I. Introduction

Reduplication in linguistics is a morphological process in which the root or stem of a word (or part of it) or even the whole word is repeated exactly or with a slight change. Reduplication is used in inflections to convey a grammatical function, such as plurality, intensification, etc., and in lexical derivation to create new words. It is often used when a speaker adopts a tone more "expressive" or figurative than ordinary speech and is also often, but not exclusively, iconic in meaning.

Reduplication is found in a wide range of languages and language groups, though its level of linguistic productivity varies. Reduplication is often described phonologically in one of two different ways

(a) Reduplicated *segments* (sequences of consonants/vowels),

(b) Reduplicated *prosodic units* (syllables or moras).

In addition to phonological description, reduplication often needs to be described morphologically as a reduplication of linguistic constituents (i.e. words, stems, and roots). As a result, reduplication is interesting theoretically as it involves the interface between phonology and morphology.

The *base* is the word (or part of the word) that is to be copied. The reduplicated element is called the *reduplicant*, in reduplication, the reduplicant is most often repeated only once. However, in some languages, reduplication can occur more than once, resulting in a tripled form, and not a *duple* as in most reduplication. Triplication is the term for this phenomenon of copying two times.

Dr. (Mrs) Kamlesh Dutta, Computer Science and Engineering Department, National Institute of Technology Hamirpur , Hamirpur, Himachal Pradesh, India,

Anshul Jindal, Computer Science and Engineering Department, National Institute of Technology Hamirpur, Hamirpur, Himachal Pradesh, India,

II. Related Work

Most MWE extraction methods are based on exploiting the various idiosyncrasies exhibited by MWEs. The variation in statistical distributional characteristics has been widely employed to test for evidence of a collocation being an institutionalized MWE. Point wise Mutual Information is one of the earliest measures of association used for collocations [3]. Word association has also been measured using measures like Jaccard, Odds Ratio, etc. [5]. Tanmoy, Dipankar and Sivaji use clustering technique to group all nouns that are related to the meaning of the individual component of an expression[9]. Anoop and OM uses statistical co-occurrence measures to exploit the statistical idiosyncrasy of MWEs [1]. In addition to the constituent words, the context in which the collocation is found can give clues about whether the collocation is a non-compositional MWE. Katz [6] and Baldwin [7] use the context as a bag of words and build context vectors for representing collocations and their constituents. If an annotated training set is available, the MWE extraction problem can be set up as a classification problem.

III. Reduplication of Words in Hindi

This describes the various categories in which Hindi reduplication words can be divided:

A. *Onomatopoeic Expressions:* The constituent words imitate a sound, and the unit as a whole refers to that sound. E.g. छन-छन (Chan Chan, sound of water falling on a hot surface), खट-खट (khat khat, knock knock).

B. *Complete Reduplication*: The individual words are meaningful, and they are repeated. E.g. कदम-कदम (kadam kadam, at every step), धीरे-धीरे (Dheere Dheere, slowly).

C. *Partial Reduplication*: Only one of the words is meaningful, while the other word is constructed by partially reduplicating the first word. There are various ways of constructing such reduplications, but the most common type in Hindi is one where the first syllable alone is

changed. E.g. अलग-थलग *(alag thalag, separated)*, रंग-िबरंगा *(rang birangaa, colorful)*.

D.Semantic Reduplication: The two paired members are semantically related. The most common forms of relation between the words are synonymy (बाग़-बगीचा, *baag bagichaa, garden*), antonyms (लेन-देन, *len den, dealing*), class representative (चाय-पानी, *chaay paanee, snacks*)).

IV. PROPOSED SYSTEM DESIGN

A system is developed which is focused on first extracting reduplication words from the given text and then identify them into different categories based upon their semantic and syntactic analysis. The system stores different categories words into different files based upon their classification.

In our approach we have devised an efficient algorithm to identify reduplication words in which the two hyphen separated words are first translated to English language and then they are compared from the backside. Depending upon the degree of similarity they are classified into different categories of reduplication.

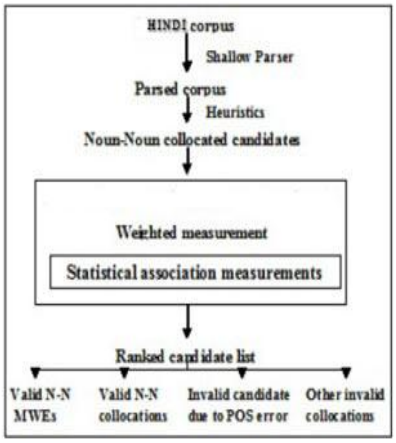

Figure 1: Basic Architecture

Basic system architecture is shown in Figure 1. The complete extraction procedure has been divided mainly into three phases. In the first phase, after initial pre-processing, candidate selection has been done using some heuristics to feed them into the main extraction phase. This process is divided into two phases.

A. Preparation of Corpus and POS tagging: The crawled corpus is so scattered and unformatted that a basic semi-automatic preprocessing has been needed. Some of them are like sentence boundary detection and make the corpus suitable for parsing. Parsing using Hindi shallow parser has been done for identifying the POS, chunk, root and inflection of ach token. Some of the tokens are misspelled due to typographic or phonetic error.

B. Identification of Reduplication: For identification we have followed a unique strategy in which the parsed English version is used for identification. In this the two words separated by hyphen are matched from the reverse side and then they are classified on the degree of their matching.

1. *Complete Reduplication:* In this the individual words arev meaningful, and they are repeated. e.g. कदम-कदम (kadam kadam, at every step), धीरे- धीरे (Dheere Dheere, slowly). Hence to identify these type of words the two words should be matched completely.
2. *Partial Reduplication:* In this three cases are possible-
    1. Change of the first vowel or the matra attached with first consonant,
    2. Change of consonant itself in first position or
    3. Change of both matra and consonant.

    Exception is reported where vowel in first position is changed to consonant and its corresponding matra is added. For e.g. अलग-थलग (alag thalag, separated), रंग-िबरंगा (rang birangaa, colourful). For identification of these type of words the degree is kept more than 1. If the matched degree is more than 1 then they come under this category.
3. *Onomatopoeic Expression:* In this, mainly words are repeated twice and may be with some matra In this case, after removing inflection, words are divided equally and then the comparison is done. For e.g. छन- छन (Chan Chan, sound of water falling on a hot surface), खट खट (khat khat, knock knock). So here also degree is kept complete matching.

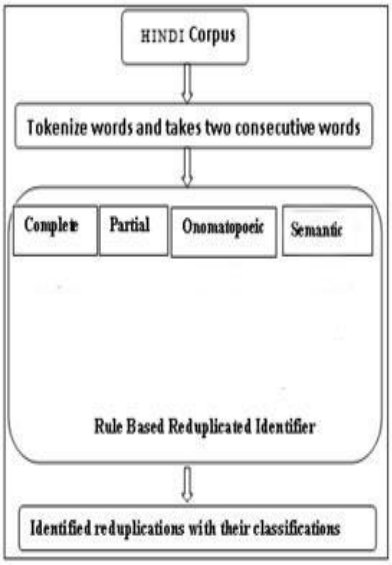

Figure 2. Identification Reduplication

4. *Semantic Reduplication*: In this the two paired members are semantically related. The most common forms of relation between the words are synonymy (बाग़- बगीचा, baag bagichaa, garden), antonymy (लेन देन, len den, dealing), class representative (चाय-पानी, chaay paanee, snacks)). For this, matching degree is kept null.

V. ANALYSIS

TABLE I
REDUPLICATION WORDS FROM HUMAN ANALYSIS

S No.	Reduplication Words	Type of Reduplication
1	सानी-पानी	Partial
2	स्त्री-पुरुष	Semantic
3	खली-भूसा	Semantic
4	चुरा-चुरा	Complete
5	स्त्री-पुरुष	Semantic
6	हाथ-पाँव	Semantic
7	भूल-चूक	Semantic
8	महीने-दो-महीने	Partial
9	कभी-कभी	Complete
10	लड़ाई-दंगा	Semantic
11	कुछ-कुछ	Complete
12	पीछे-पीछे	Complete
13	उम्र-भर	Semantic
14	बनी-ठनी	Partial

TABLE II
REDUPLICATION WORDS FROM SYSTEM ANALYSIS

S No.	Reduplication Words	English Parsed Word	Type of Reduplication
1	चुरा-चुरा	curA-curA	Complete
2	कभी-कभी	kaBI-kaBI	Complete
3	कुछ-कुछ	kuCa-kuCa	Complete
4	सानी-पानी	sAnI-pAnI	Partial
5	सानी-पानी	sAnI-pAnI	Partial
6	हाथ-पाँव	hAWa-pAzva	Partial
7	भूल-चूक	BUla-cUka	Partial
8	महीना-दो-महीने	mahIne-xo-mahIne	Partial
9	उम्र-भर	umra-Bara	Partial
10	बन-ठन	banI-TanI	Partial
11	स्त्री-पुरुष	swrI-puruRa	Semantic
12	खली-भूसा	KalI-BUsA	Semantic
13	स्त्री-पुरुष	swrI-puruRa	Semantic
14	लड़ाई-दंगा	ladZAI-xaMgA	Semantic
15	वह-चल	usake-calI	No Category

From the above results we found the following:
All the multiword expressions are correctly read by the system. There was no difference in output of complete reduplication and onomatopoeic expression. Complete Reduplication words were correctly interpreted. Partial reduplication words contain some words which should come to semantic reduplication Semantic reduplication also contain some words which should come to partial reduplication. System designed is not 100% correct but give a different point of view of classification of words.

VI. CONCLUSION

An attempt has been made in the project to model the syntax and semantics of Hindi Multi Word Expressions (MWEs) based on the rule based theory. Our approach of differentiating the words into different reduplication categories by matching the words from the back side is very efficient and different from all the other approaches. Although there are some errors in the results obtained but if more improvements can be applied on this system then we can get to 90% precision. The purpose of our work was to come up with a list of potential reduplication MWEs which a lexicographer can look at and decide whether a given word sequence should be added to the lexicon and we have fulfilled that. With the use of this system we get different reduplication words which we can used to design a lexicon. Complete automation of the MWE extraction is still a difficult task.

REFERENCES

[1] Anoop Kunchukuttan and Om P. Damani A System for Compound Noun Multiword Expression Extraction for Hindi 2006
[2] A. Wray. *Formulaic Language and the Lexicon.* Cambridge University Press. 2002.
[3] K. Church and P. Hanks. *Word association norms, mutual information, and lexicography*. Computational Linguistics. 16(1), 1990..
[4] T. Dunning. Accurate methods for the statistics of surprise and coincidence. Computational Linguistics. 19(1), 1993.
[5] P. Pecina. An extensive empirical study of collocation extraction methods. ACL Student Research Workshop. 2005..
[6] G. Katz and E. Giesbrechts. Automatic identification of noncompositional multi-word expressions using Latent Semantic Analysis. ACL-2006 Workshop on Multiword Expressions. 2006.
[7] T. Baldwin, C. Bannard, T. Tanaka, and D.Widdow. *An empirical model of multiword expressions decomposability*. ACL-2003 Workshop on Multiword Expressions. 2003.
[8] B.V. Moiron and J. Tiedemann. *Identifying idiomatic expressions using automatic word alignment*. EACL 2006 Workshop on Multiword Expressions in a multilingual context. 2006.
[9] Identifying Bengali Multiword Expressions using Semantic Clustering by Tanmoy ,Dipankar and Sivaji

Dr. (Mrs.) Kamlesh Dutta, Associate Professor, Computer Science & Engineering Department, National Institute of Technology, Hamirpur (HP), India. Coordinator videoconferencing and communication, Member SUGC
Member SUGC (curriculum), Member APEC, Member Committee for the revision of UG curriculum, Member, Community Service Cell, Member, Campus video networking committee, Member, Automation Committee
Convener DUGC, CSE Department, Coordinator Project, CSE Department
Faculty-In charge CSE , SMDP-II programme, NIT Hamirpur (HP), Member Committee - Sexual Harassment of Women at Work Place, NIT Hamirpur (HP). http://nith.ac.in/~kd/?Home

Anshul Jindal, B-Tech in Computer Science and Engineering from National Institute of technology Hamirpur. Secured 94% in Physics, 98% in Mathematics & 95% in Computer-Science in AISSCE-10, received a Shield in National Mathematics Olympiad Contest in High School for scoring 100% marks. Received Certificate of Merit for being among top 0.1% students in India by securing 100% Marks in Mathematics in AISSE-08