Identification of Reduplication in Bengali Corpus and their Semantic Analysis: A Rule-Based Approach

Tanmoy Chakraborty
Department of Computer Science and Engineering
Jadavpur University
its_tanmoy@yahoo.co.in

Sivaji Bandyopadhyay
Department of Computer Science and Engineering
Jadavpur University
sivaji_cse_ju@yahoo.co.in

Abstract

In linguistic studies, reduplication generally means the repetition of any linguistic unit such as a phoneme, morpheme, word, phrase, clause or the utterance as a whole. The identification of reduplication is a part of general task of identification of multiword expressions (MWE). In the present work, reduplications have been identified from the Bengali corpus of the articles of Rabindranath Tagore. The present rule-based approach is divided into two phases. In the first phase, identification of reduplications has been done mainly at general expression level and in the second phase, their structural and semantics classifications are analyzed. The system has been evaluated with average Precision, Recall and F-Score values of 92.82%, 91.50% and 92.15% respectively.

1 Introduction

In all languages, the repetition of noun, pronoun, adjective and verb are broadly classified under two coarse-grained categories: repetition at the (a) ***expression level***, and at the (b) ***contents or semantic level.*** The repetition at both the levels is mainly used for emphasis, generality, intensity or to show continuation of an act. The paper deals with the identification of reduplications at both levels in Bengali. Reduplication phenomenon is not an exotic feature of Indian Languages. For instance, Yiddish English has duplication of the form X schm-X, as in "duplication schmuplication". Semantic duplication is also rich in English and Onomatopoeic repetition is not uncommon either (e.g., ha-ha, blah-blah etc).

Reduplication carries various semantic meanings and sometime helps to identify the mental state of the speaker as well. Some correlative words are used in Bengali to express the possessiveness, relative or descriptiveness. They are called '***secondary descriptive compounds***'.

The related studies on MWEs are discussed in Section 2. Various types of reduplications in Bengali and their semantic interpretations are discussed in Section 3. The proposed system architecture and the procedures are discussed in Section 4. The evaluation metrics used for evaluating the system are discussed in Section 5. Experimental results are presented in Section 6 and conclusions are drawn in Section 7.

2 Related Work

The works on MWE identification and extraction have been continuing in English (Fillmore, 2003; Sag et. al, 2002). After tokenization, multiword expressions are important in understanding the meaning in applications like Machine Translation, Information Retrieval system etc. Some of the MWE extraction tasks in English can be seen in (Diab and Bhutada, 2009; Enivre and Nilson, 2004). Among Indian languages, Hindi compound noun MWE extraction has been studied in (Kunchukuttan and Damani, 2008). Manipuri reduplicated MWE identification is discussed in (Nongmeikapam and Bandyopadhyay, 2010). There are no published works on reduplicated MWE identification in Bengali.

3 Reduplication of Words in Bengali

Identification of MWEs is done during the tokenization phase and is absolutely necessary

Proceedings of the Multiword Expressions: From Theory to Applications (MWE 2010), pages 73–76,
Beijing, August 2010

during POS tagging as is outlined in (Thoudam and Bandyopadhyay, 2008). POS tagger identifies MWE as unknown word at token level. Bengali Shallow Parser[1] can only identify hyphened reduplication and gives them separate tags like RDP (reduplication) or ECH (echo).

Another objective for identifying reduplicated MWEs is to extract correct sense of reduplicated MWEs as discussed in Section 3.2. Sometime, reduplication is used for sentiment marking to identify whether the speaker uses it in positive or negative sense. For example,

(i) Eto **Bara Bara** Asha Kisher?*(Why are you thinking so high?*) (Positive Sense)

(ii) Ki **Bara Bara** Bari Ekhane! *(Here, the buildings are very large.)* (Negative Sense)

3.1 Expression Level Classification of Reduplication

Four classes of reduplications commonly occur in the Indian language (Bengali, Hindi, Tamil[2], Manipuri etc.). In Bengali, another type called ***correlated word*** is also classified as reduplication.

Onomatopoeic expressions: Such words represent an imitation of a particular sound or imitation of an action along with the sound, etc. For example, **khat khat**, *(knock knock).*

Complete Reduplication: The individual words carry certain meaning, and they are repeated. e.g. **bara-bara** *(big big)*, **dheere dheere**, *slowly)*. In some cases, both the speaker and the listener repeat certain clauses or phrases in long utterances or narrations. The repetition of such utterances breaks the monotony of the narration, allows a pause for the listener to comprehend the situation, and also provides an opportunity to the speaker to change the style of narration.

Partial Reduplication: Only one of the words is meaningful, while the second word is constructed by partially reduplicating the first word. Most common type in Bengali is one where the first letter or the associated matra or both is changed, e.g. **thakur-thukur** *(God)*, **boka-*soka*** *(Foolish*) etc.

Semantic Reduplication: The most common forms of semantic relations between paired words are *synonym* (**matha-mundu**, *head*), *antonym* (**din-rat**, *day and night*), *class representative* (**cha-paani**, *snacks*)).

Correlative Reduplication: To express a sense of exchange or barter or interchange, the style of corresponding correlative words is used just preceding the main root verb. For example, **maramari**(*fighting*).

3.2 Reduplication at the Sense Level

Different types of reduplication at the sense level are described below:

i. **Sense of repetition:**
Bachar Bachar Ek Kaj Kara .
(*Do the same job every year.)*

ii. **Sense of plurality:**
Ki **Bara Bara** Bari Ekhane.
(Here, the houses are very large.)

iii. **Sense of Emphatic or Modifying Meaning:**
Lala-Lala phul. (Deep *red rose*)

iv. **Sense of completion:**
Kheye Deye Ami Shute Jaba.
After eating, I shall go to sleep.

v. **Sense of hesitation or softness:**
Eta **Hasi Hasi** Mukh Kena?
Why does your face smiling?

vi. **Sense of incompleteness of the verbs:**
Katha **Bolte Bolte** Hatat Se Chup Kore Gelo.
Talking about something, suddenly he stopped.

vii. **Sense of corresponding correlative words:**
Nijera **Maramari** Kara Na.
Don't fight among yourselves.

viii. **Sense of Onomatopoeia:**
Shyamal Darja **Khata khata** Karchhe .
Shyamal is knocking at the door.

4 System Design

The system is designed in two phases. The first phase identifies mainly five cases of reduplication discussed in Section 3.1 and the second phase attempts to extract the associated sense or semantics discussed in Section 3.2.

4.1 Identifying Reduplications

Reduplication is considered as two consecutive words W1 and W2. For **complete reduplication**, after removing matra, comparison for complete equality of two words is checked.

[1] http://ltrc.iiit.ac.in/analyzer/bengali

[2] http://users.ox.ac.uk/~sjoh0535/thesis.html

In **partial reduplication**, three cases are possible- (i) change of the first vowel or the matra attached with first consonant, (ii) change of consonant itself in first position or (iii) 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 example, **আবল-তাবল (***abal-tabal, incoherent* or *irrelevant*)**.** Linguistic study (Chattopadhyay, 1992) reveals that the consonants that can be produced after changing are 'ট', 'ফ', 'ম', 'স'.

For **onomatopoeic expression,** mainly words are repeated twice and may be with some matra (mainly 'এ'-matra is added with the first word to make second word). In this case, after removing inflection, words are divided equally and then the comparison is done.

For **correlative reduplication**, the formative affixes '–আ' and '-ই' are added with the root to form w1 and w2 respectively and agglutinated together to make a single word.

For **semantic reduplication,** a dictionary based approach has been taken. List of inflections identified for the semantic reduplication is shown in Table 1.

Set of identified inflections and matra
০(শূন্য), এ(–য়ে, –য়), –তে(–এতে), –কে, রে(–এরে), –র, –এর(য়ের), এরা, –দের, –টা, –টি, –গুলো , –ও, –ই,

Table 1. Inflections identified for semantic reduplication.

This system has identified those consecutive words having same part-of-speech. Then, morphological analysis has been done to identify the roots of both components. In synonymous reduplication, w2 is the synonym of w1. So, at first in Bengali monolingual dictionary, the entry of w1 is searched to have any existence of w2. For antonym words, they are mainly *gradable opposite* (**pap-purna,** *Vice and Virtue*) where the word and its antonyms are entirely different word forms. The *productive opposites* (**garraji**, *disagree* is the opposite of **raji,** *agree)* are easy to identify because the opposite word is generated by adding some fixed number of prefixes or suffixes with the original. In dictionary based approach, English meaning of both w1 and w2 are extracted and opposite of w1 is searched in English WordNet[3] for any entry of w2. The first model for identifying the five types of reduplications is shown in Figure 1.

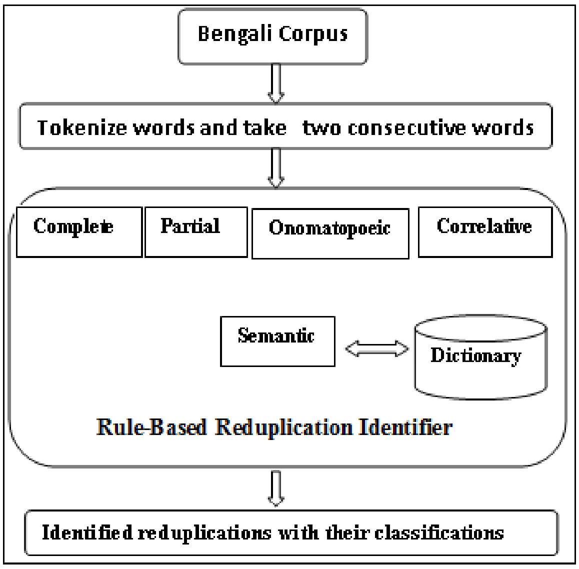
Bengali Corpus
Tokenize words and take two consecutive words
Complete
Partial
Onomatopoeic
Correlative
Semantic
Dictionary
Rule-Based Reduplication Identifier
Identified reduplications with their classifications

Figure 1. System Architecture of first phase.

4.2 Semantics (Sense) Analysis

Mainly eight types of semantic classifications are identified in Section 3.2. If the reduplication is an onomatopoeic expression, its sense is easily identified as the sense of onomatopoeia. When infinite verb with complete reduplication is identified in a sentence, it obviously expresses the sense of incompleteness. The semantic or partial reduplicated words belong to the sense of completion. The correlative word is classified as the sense of corresponding correlative word because it is generally associated with the full verb in the sentence. The problem arises when grouping the complete reduplication. Sometime they are used as sense of repetition, plurality and sometime they express some kind of hesitation, incompleteness or softness. Sense disambiguation for this case has been identified as a future work.

5 Evaluation Metrics

The corpus is collected from some selected articles of Rabindranath Tagore[4]. Standard IR metrics like Precision, Recall and F-score are used to evaluate the system. Total number of relevant

[3] http://wordnetweb.princeton.edu/perl/webwn

[4] http://www.rabindra-rachanabali.nltr.org

reduplication is identified manually. For each type of expression level classification, Precision, Recall and F-score are calculated separately. The overall system score is the average of these scores. Statistical co-occurrence measures like frequency, hyphen and closed form count are calculated on each of the types as an evidence of their MWEhood.

6 Experimental Results

The collected corpus includes 14,810 tokens for 3675 distinct word forms at the root level. Precision, Recall, F-score are calculated for each class as well as for the reduplication identification system and are shown in Table 2.

Reduplications	Precision	Recall	F-Score
Onomatopoeic	99.85	99.77	99.79
Complete	99.98	99.92	99.95
Partial	79.15	75.80	77.44
Semantic	85.20	82.26	83.71
Correlative	99.91	99.73	99.82
System	92.82	91.50	92.15

Table 2. Evaluation results for various reduplications (in %).

The scores of partial and semantic evaluation are not satisfactory because of some wrong tagging by the shallow parser (adjective, adverb and noun are mainly interchanged). Some synonymous reduplication (**ধীরে-সুস্থে**, *dhire-susthe, slowly and steadily, leisurely*) implies some sense of the previous word but not its exact synonym. These words are not identified properly. Figure 2 shows that the use of complete reduplication is more in this corpus. In this corpus, only 8.52% reduplications are hyphened. It shows that the trend of writing reduplications is to use the space as separator. Also the percentage of closed reduplications is 33.09% where maximum of them are onomatopoeic, correlative and semantic reduplications. 100% of correlative reduplications are closed.

7 Conclusion

The reduplication phenomenon has been studied for Bengali at the expression as well as at the semantic levels. The semantics of the reduplicated words indicate some sort of sense disambiguation that cannot be handled by only rule-based approach. More works need to be done for identifying semantic reduplication using statistical and morphological approaches.

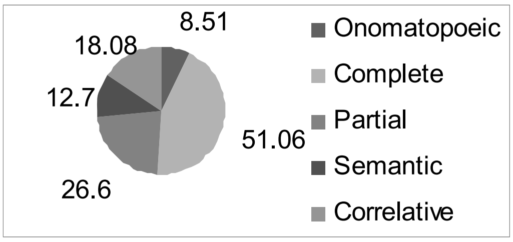

Figure 2. Frequencies (in %) of different reduplications.

References

Bhaskararao, Peri. 1977. Reduplication and Onomatopoeia in Telugu. Deccan College Post-Graduate and research Institute, Pune, India.

Chattopadhyay Suniti Kumar. 1992. Bhasa-Prakash Bangala Vyakaran, Third Edition.

Diab, Mona and Pravin Bhutada. 2009. Verb Noun Construction MWE Token Supervised Classification, *In Proceedings of the Joint conference of Association for Computational Linguistics and International Joint Conference on Natural Language Processing, Workshop on Multiword Expression.*, Singapore, pp.17-22.

Enivre, Joakim and Jens Nilson. 2004. Multiword Units in Syntactic Parsing. *In Proceedings of Methodologies and Evaluation of Multiword Units in Real-World Applications, 2004 Workshop*, Lisbon, pp. 39-46.

Kunchukuttan, Anoop and Om Prakash Damani, 2008. A System for Compound Noun Multiword Expression Extraction for Hindi. *6th International. Conference on Natural Language Processing*, Pune, pp. 20-29.

Nongmeikapam, Kishorjit and Sivaji Bandyopadhyay. 2010. Identification of Reduplication MWEs in Manipuri, a rule-based approach, *In Proceedings of the 23rd International Conference on the Computer Processing of Oriental Languages*, California, USA, pp. 49-54.

Thoudam, Doren Singh and Sivaji Bandyopadhyay. 2008. Morphology Driven Manipuri POS Tagger. *In workshop on NLP for Less Privileged Languages*, *International Joint conference of Natural Language Processing*, Hyderabad, pp. 91-98