Detection of Compound Nouns and Light Verb Constructions using IndoWordNet

Dhirendra Singh Sudha Bhingardive Pushpak Bhattacharyya

Department of Computer Science and Engineering,
Indian Institute of Technology Bombay.
{dhirendra,sudha,pb}@cse.iitb.ac.in

Abstract

Detection of MultiWord Expressions (MWEs) is one of the fundamental problems in Natural Language Processing. In this paper, we focus on two categories of MWEs - *Compound Nouns* and *Light Verb Constructions*. These two categories can be tackled using knowledge bases, rather than pure statistics. We investigate usability of IndoWordNet for the detection of MWEs. Our *IndoWordNet based approach* uses semantic and ontological features of words that can be extracted from IndoWordNet. This approach has been tested on Indian languages *viz.,* Assamese, Bengali, Hindi, Konkani, Marathi, Odia and Punjabi. Results show that ontological features are found to be very useful for the detection of *light verb constructions*, while use of semantic properties for the detection of *compound nouns* is found to be satisfactory. This approach can be easily adapted by other Indian languages. Detected MWEs can be interpolated into WordNets as they help in representing semantic knowledge.

1 Introduction

MultiWord Expressions or MWEs can be described as idiosyncratic interpretations that crosses word boundaries or spaces (Sag et al., 2002). MWE is formed by atleast two words which are syntactically and/or semantically idiosyncratic in nature. For example, *swimming pool*, *telephone booth*, *strong coffee*, *pay attention*, *fast food*, etc. are some of the MWEs in English, while धन दौलत (*Dhana daulata*, wealth), वादा करना (*vaadaa karanaa*, to promise), मार डालना (*maara Daalanaa*, to kill), धीरे धीरे (*dhiire dhiire*, slowly), *etc.* are some of the MWEs in Hindi. In past, ample number of approaches have been proposed in literature for the detection of MWEs (Calzolari et al., 2002),(Baldwin et al., 2003), (Guevara, 2010), (Al-Haj and Wintner, 2010), (Tsvetkov and Wintner, 2012). However, for Indian languages, many researchers have proposed statistical and rule based approaches (Sinha, 2009), (Kunchukuttan and Damani, 2008), (Chakrabarti et al., 2008), (Mukerjee et al., 2006), (Sinha, 2011), (Singh et al., 2012), (Sriram et al., 2007).

This paper focuses on Indian languages *viz.,* Assamese, Bengali, Hindi, Konkani, Marathi, Odia and Punjabi for the detection of MWEs. These languages are part of the IndoWordNet[1]. To the best of our knowledge, the IndoWordNet based approach is being used for the first time for detecting MWEs. This approach is restricted for two categories of MWEs: *compound nouns* (Noun+Noun) and *light verb constructions* (Noun+Verb, Adjective+Verb, Verb+Verb). Semantic features of words are used for detecting *compound nouns*, while ontological features are used for detecting *light verb constructions*. The motivation behind this work is that,

- If we add suitable amount of MWEs in WordNet, its coverage will be increased in terms of vocabulary and linguistic phenomenon.
- Improper handling of MWEs is one of the

[1]IndoWordNet is available in following Indian languages: Assamese, Bodo, Bengali, English, Gujarati, Hindi, Kashmiri, Konkani, Kannada, Malayalam, Manipuri, Marathi, Nepali, Punjabi, Sanskrit, Tamil, Telugu and Urdu. These languages cover three different language families, Indo-Aryan, Sino-Tibetan and Dravidian. http://www.cfilt.iitb.ac.in/indowordnet/

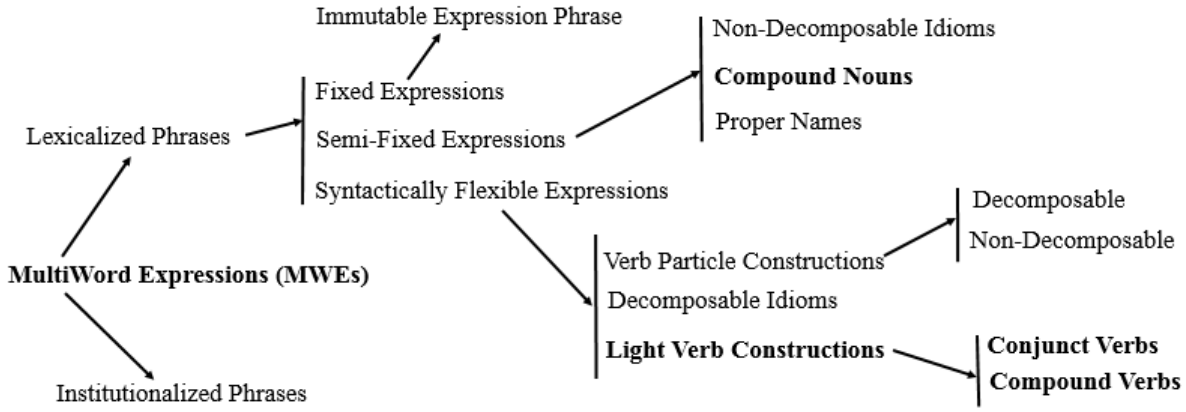

Figure 1: Classification of MWEs

major sources of error in various NLP applications. Hence, correct detection of MWEs will show improvement in performance of these applications, as reported by Finlayson et al. (2011) for word sense disambiguation, Ren et al. (2009) and Bouamor et al. (2011) for machine translation, etc.

The roadmap of the paper is as follows. Section 2 covers the classification of MWEs. The IndoWordNet based approach is explained in Section 3. Section 4 details the experimental setup. Results are presented in section 5 and discussed in section 6. Related work is given in section 7, followed by conclusion and future work.

2 MWEs Classification

MWEs are classified based on their lexical and semantic characteristics (Sag et al., 2002). This has been further studied from Indian language perspective and expanded as shown in Figure 1. As we can see in figure 1, we modified the Sag et al., (2002) classification by adding Light Verb Constructions and its further classification which is needed for Indian languages. MWEs are classified into two broad categories. They are Lexicalized Phrases and Institutional Phrases. The meaning of lexicalized phrases cannot be construed from its individual units that make up the phrase, as they exhibit syntactic and/or semantic idiosyncrasy. On the other hand, the meaning of institutional phrases can be construed from its individual units that make up the phrase. However, they exhibit statistical idiosyncrasy. Institutional phrases are not in the scope of this paper. Lexicalized phrases are further classified into three sub-classes *viz.,* Fixed, Semi-fixed and Syntactically flexible expressions.

In this paper, we focus on *compound nouns* and *light verb constructions* which fall under the semi-fixed and syntactically flexible categories respectively.

2.1 Compound Nouns

Compound Nouns (CNs) are syntactically-unalterable units that inflect for number. A word-pair forms CN if its meaning cannot be composed from the meanings of its constituent words. CNs are formed by either Noun+Noun or Adj+Noun word combinations. For example, पेड़ पौधे (*peda paudhe*, flora), बाग बगीचा (*baaga bagiichaa* , garden), काला धन (*kaalaa dhana* , black money), *etc.*

2.2 Light Verb Constructions

Light Verb Constructions (LVCs) show high idiosyncratic constructions with nouns. It is difficult to predict which light verb chooses which noun and why the light verb cannot be substituted with another. LVCs are further classified into Conjunct Verbs (CjVs) and Compound Verbs (CpVs). *CjVs* are formed by Noun+Verb and Adj+Verb word combinations, while *CpVs* are formed by Verb+Verb

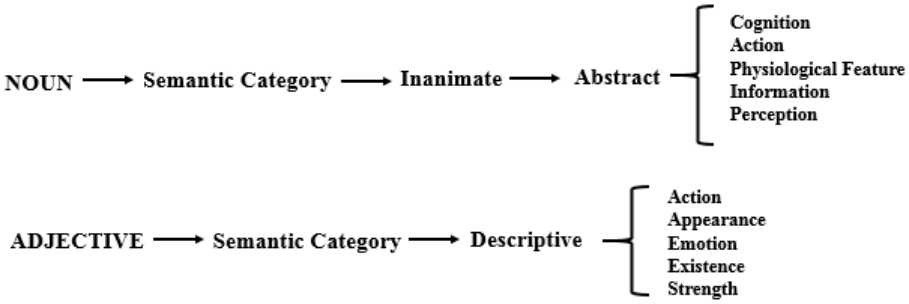

Figure 2: Noun and Adjective ontological features needed to form Conjunct Verbs

Figure 3: Verb ontological features needed to form Compound Verbs

word combinations. Examples of *CjVs* are गुजर जाना (*gujara jaanaa*, passed away), काम करना (*kaama karanaa*, to work), प्यार करना (*pyaara karanaa*, to love), etc. and examples of *CpVs* are भाग जाना (*bhaaga jaanaa*, run away), उठ जाना (*uTha jaanaa*, to wake up), खा लेना (*khaa lenaa*, to eat), etc.

3 IndoWordNet Based Approach

Our IndoWordNet based approach uses various semantic and ontological features from the IndoWordNet. The semantic features are used for *CN* detection while ontological features are used for *LVC* detection. Now, we explain the IndoWordNet based approach for each of these categories.

3.1 Detection of Compound Nouns

The semantic features of words such as *synonyms*, *definition/gloss*, *example sentence*, *hypernyms*, *antonyms, etc.* are used for detection of *CNs*.

The *bag of words* (*BOW*) for a word w_i is created using the semantic features of IndoWordNet, as follows.

$$BOW(w_i) = \{x | x \in WordNetFeatures(w_i)\}$$

where, $WordNetFeatures(w_i)$ contains all content words from *synonyms*, *gloss*, *example(s)*, *hypernyms*, *hyponyms*, *meronyms*, *antonyms* with respect to the word w_i. We considered only one level of hierarchy for extracting these semantic features.

Consider a word-pair w_1w_2 to be detected as a MWE. As per the IndoWordNet based approach, the given pair can be treated as *compound noun* MWEs when any one of the following condition holds -

- if $w_1 \in BOW(w_2)$, then w_1w_2 is a *CN*
- if $w_2 \in BOW(w_1)$, then w_1w_2 is a *CN*

For instance, consider a word-pair in Hindi, धन दौलत (*dhanaa daulata, wealth*). The *BOWs* for *dhana* and *daulata* are as follows,

$$BOW(dhana) = \{paisaa, daulata, vaibhava, ..\}$$

$$BOW(daulata) = \{sampatti, laxmi, dhana, ...\}$$

Since, $dhana \in BOW(daulata)$, the word-pair *dhana daulat* is considered as a *CN*.

3.2 Detection of Light Verb Constructions

The ontological features of words such as *abstract, inanimate, action, information, etc.* (refer figure 4) are used for detection of LVCs. There are two types of LVCs, Conjunct Verbs (CjVs) and Compound Verbs (CpVs).

3.2.1 Conjunct Verbs

As mentioned earlier, conjunct verbs are formed by Noun+Verb and Adj+Verb word combinations. However, it is very difficult to predict which type of nouns or adjectives form *CjVs*. Previous approaches tried to detect

such nouns or adjectives based on their statistical collocation with restricted sets of verbs (most frequently used, manually selected, etc.) (Sidhu et al., 2010). This limitation results in less coverage at *CjV* detection.

We claim that whether a noun or an adjective forms *CjVs* depends on its ontological properties. Figure 2 shows some ontological properties of nouns and adjectives that are available in IndoWordNet and needed to form *CjVs*. This removes the dependence on the restricted set of verbs, thereby increasing the upper bound of coverage that we can achieve. Algorithm 1 details the detection of *CjVs*.

Algorithm 1 Conjunct Verb Detection

```
1: procedure CJV–DETECTION (w1,w2)
2:     if w1 is Noun and w2 is Verb then
3:         if w1 is abstract Noun then
4:             print "CjV detected"
5:         end if
6:     end if
7:
8:     if w1 is Adj and w2 is Verb then
9:         if w1 is descriptive Adj then
10:            print "CjV detected"
11:        end if
12:    end if
13: end procedure
```

3.2.2 Compound Verbs

As mentioned earlier, compound verbs are formed by Verb+Verb word combinations. The first verb gives lexical information whereas the second verb provides grammatical information about the expression. Just as in the case of *CjVs*, formation of *CpVs* also depends on the ontological properties of the constituent verbs. Figure 3 shows some ontological properties of verbs that are available in IndoWordNet and needed to form *CpVs*. Algorithm 2 details the detection of *CpVs*.

4 Experiments

We performed experiments on some Indian languages *viz.,* Hindi, Marathi, Bengali, Punjabi, Konkani, Odia, Assamese for the detection of *compound nouns* and *conjunct verbs*. However, for *compound verb* detection, we performed experiments only on Hindi and Marathi due to unavailability of gold data for other languages.

The gold data for these experiments is created by automatically extracting Noun+Noun, Noun+Verb, Adj+Verb and Verb+Verb word-pair combinations. These word-pairs are extracted from the generic domain in-house corpus. Out of these word-pairs, 1000 Noun+Noun word-pairs are detected as *CNs* for each of the seven languages mentioned above, while 399 and 504 Verb+Verb word-pairs are detected as *CpVs* for Marathi and Hindi respectively. Also, 457, 404, 797, 1017, 879, 832, 703 Noun+Verb and 577, 502, 303, 307, 269, 368, 259 Adj+Verb word-pairs are detected as *CjVs* for Hindi, Marathi, Bengali, Punjabi, Konkani, Odia, Assamese languages respectively. Three lexicographers were engaged in this activity and the inter-annotator agreement is found to be 0.8.

Algorithm 2 Compound Verb Detection

```
1: procedure CPV–DETECTION (w1,w2)
2:     if w1 is action verb then
3:         if w2 is action verb or
4:             w2 is occurrence verb then
5:             print "CpV detected"
6:         end if
7:     end if
8:
9:     if w1 is occurrence verb then
10:        if w2 is action verb then
11:            print "CpV detected"
12:        end if
13:    end if
14: end procedure
```

5 Results

In this section, results of the experiments are presented and discussed in detail. Table 1 shows the results obtained for the detection of *CNs*, while Table 2 and Table 3 show the results obtained for the detection of *CjVs* and *CpVs* respectively. It has been observed that results of *CN* detection are found to be considerably good only for Marathi as compared to other languages. However, the results of *CjV* and *CpV* detection are found to be promising for languages under consideration. Hence, we can say that, IndoWordNet based approach using ontological properties are found to be very effective for the detection of *light verb constructions* such as *CpVs* and *CjVs*.

6 Discussions

As we have observed that the results of *CN* detection are found to be unsatisfactory for languages other than Marathi. This may be be-

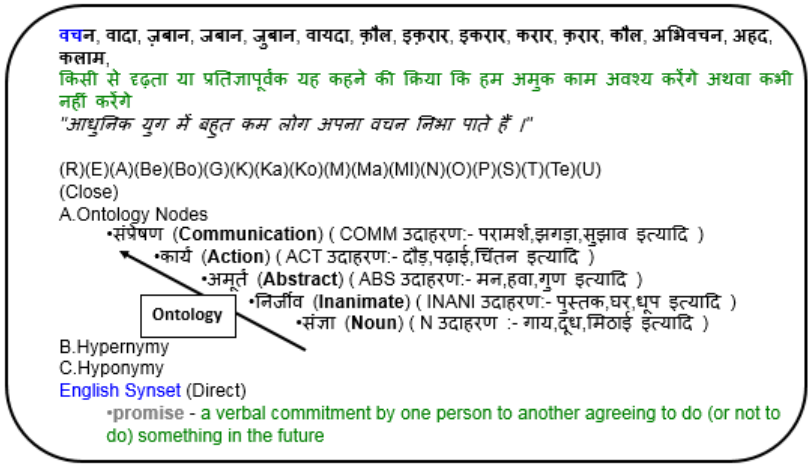

Figure 4: IndoWordNet ontological properties for a Hindi word 'vachanaa' (promise)

	Compound Nouns (CNs)	
	Total pairs(N+N)	**F-score**
Hindi	1000	0.58
Marathi	1000	0.72
Bengali	1000	0.53
Punjabi	1000	0.43
Konkani	1000	0.52
Odia	1000	0.38
Assamese	1000	0.40

Table 1: Results of Compound Noun Detection

	Compound Verbs (CpVs)	
	Total pairs(V+V)	**F-score**
Hindi	399	0.99
Marathi	504	0.88

Table 2: Results of Compound Verb Detection

	Conjunct Verbs (CjVs)			
	Total pairs(N+V)	**F-score**	**Total pairs(Adj+V)**	**F-score**
Hindi	457	0.87	577	0.89
Marathi	404	0.86	502	0.88
Bengali	797	0.87	303	0.92
Punjabi	1017	0.8	307	0.9
Konkani	879	0.84	269	0.95
Odia	832	0.85	368	0.91
Assamese	703	0.84	259	0.94

Table 3: Results of Conjunct Verb Detection

cause, our IndoWordNet based approach completely depends on the semantic properties of words and do not rely on the statistical co-occurrence. Also, in IndoWordNet, there are some word-pairs which are not semantically related but can form *compound nouns* due to their high statistical co-occurrence in the corpus. For example, काला धन (*kaalaa dhana*, black money) is a *CN* even though काला (*kaalaa*,black) and धन (*dhana*, money) do not exhibit any semantic relation in the IndoWordNet.

Results of *CjV* detection for Noun+Verb and Adj+Verb combinations are found to be promising. This may be because, our IndoWordNet based approach uses ontological properties of words wherein coverage of nouns and adjectives is high in IndoWordNet. While, the results of the detection of *CpVs* are found to be almost 100% for Hindi and 88% for Marathi. This also used ontological properties of words. Hence, we can say that IndoWordNet based approach is very useful for the detection of *CjVs* and *CpVs*.

7 Related Work

Most of the proposed approaches for the detection of multiword expressions are statistical in nature. They are based on association methods (Church and Hanks, 1990), deep linguistics based methods (Bansal et al., 2014), word embeddings based methods (Salehi et al., 2015), *etc.* The detection of MWEs for Indian languages is not explored much by researchers due to the reasons such as unavailability of gold data (Reddy, 2011), unstructured classification of MWEs, improper universal theory, *etc.* In literature, Gayen and Sarkar et al. (2013) used Random Forest approach for Compound Noun detection for Bengali language. Sriram et al. (2007) used a classification based approach for extracting Noun-Verb collocations for Hindi language. Mukerjee et al. (2006) used parallel corpus alignment and Part-Of-Speech tag projection to extract complex predicates. However, our IndoWordNet based approach uses ontological and semantic features of words to detect MWEs. The focus is restricted for the detection of *compound nouns* and *light verb constructions.*

8 Conclusion and Future Work

Detection of MultiWord expressions is the fundamental problem and a challenging task in the area of NLP. To address this problem, an IndoWordNet based approach is proposed in this paper. The focus is restricted to the detection of *compound nouns* and *light verb constructions.* Semantic features of words from IndoWordNet are used for the detection of *compound nouns*, while ontological features of words are used for the detection of *light verb constructions.* The IndoWordnet based approach is tested on some Indian languages *viz.,* Assamese, Bengali, Hindi, Konkani, Marathi, Odia, punjabi. It has been observed that our approach gives encouraging results for the detection of light verb constructions as compared to compound nouns. In future, the detected MWEs can be incorporated in IndoWordNet as they can help to represent the lexical knowledge. This approach can be used in NLP applications *viz.,* word sense disambiguation, machine translation, information retrieval, question answering, sentiment analysis, *etc.* It can be implemented and tested for other Indian languages.

References

Hassan Al-Haj and Shuly Wintner. 2010. Identifying multi-word expressions by leveraging morphological and syntactic idiosyncrasy. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 10–18. Association for Computational Linguistics.

Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment-Volume 18*, pages 89–96. Association for Computational Linguistics.

Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2014. Tailoring continuous word representations for dependency parsing. In *Proceedings of ACL.*

Dhouha Bouamor, Nasredine Semmar, and Pierre Zweigenbaum. 2011. Improved statistical machine translation using multiword expressions. International Workshop on Using Linguistic Information for Hybrid Machine Translation (LIHMT 2011).

Nicoletta Calzolari, Charles J. Fillmore, Ralph Grishman, Nancy Ide, Alessandro Lenci, Catherine

MacLeod, and Antonio Zampolli. 2002. Towards best practice for multiword expressions in computational lexicons. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002)*, Las Palmas, Canary Islands - Spain, May. European Language Resources Association (ELRA). ACL Anthology Identifier: L02-1259.

Debasri Chakrabarti, Hemang Mandalia, Ritwik Priya, Vaijayanthi M Sarma, and Pushpak Bhattacharyya. 2008. Hindi compound verbs and their automatic extraction. In *COLING (Posters)*, pages 27–30.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Comput. Linguist.*, 16(1):22–29, March.

Mark Alan Finlayson and Nidhi Kulkarni. 2011. Detecting multi-word expressions improves word sense disambiguation. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 20–24. Association for Computational Linguistics.

Vivekananda Gayen and Kamal Sarkar. 2013. Automatic identification of Bengali noun-noun compounds using random forest. In *Proceedings of the 9th Workshop on Multiword Expressions*, pages 64–72, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Emiliano Guevara. 2010. A regression model of adjective-noun compositionality in distributional semantics. In *Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics*, pages 33–37. Association for Computational Linguistics.

Anoop Kunchukuttan and Om Prakash Damani. 2008. A system for compound noun multiword expression extraction for hindi. In *6th International. Conference on Natural Language Processing*, pages 20–29.

Amitabha Mukerjee, Ankit Soni, and Achla M Raina. 2006. Detecting complex predicates in hindi using pos projection across parallel corpora. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 28–35. Association for Computational Linguistics.

Siva Reddy. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP-11)*.

Zhixiang Ren, Yajuan Lü, Jie Cao, Qun Liu, and Yun Huang. 2009. Improving statistical machine translation using domain bilingual multiword expressions. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, MWE '09, pages 47–54, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ivan A Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Computational Linguistics and Intelligent Text Processing*, pages 1–15. Springer.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2015. A word embedding approach to predicting the compositionality of multiword expressions. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pages 977–983.

Brahmaleen K Sidhu, Arjan Singh, and Vishal Goyal. 2010. Identification of proverbs in hindi text corpus and their translation into punjabi. *Journal of Computer Science and Engineering*, 2(1):32–37.

Smriti Singh, Om P Damani, and Vaijayanthi M Sarma. 2012. Noun group and verb group identification for hindi. In *COLING*, pages 2491–2506. Citeseer.

R Mahesh K Sinha. 2009. Mining complex predicates in hindi using a parallel hindi-english corpus. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, pages 40–46. Association for Computational Linguistics.

R Mahesh K Sinha. 2011. Stepwise mining of multi-word expressions in hindi. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 110–115. Association for Computational Linguistics.

V Sriram, Preeti Agrawal, and Aravind K Joshi. 2007. Relative compositionality of noun verb multi-word expressions in hindi. In *published in Proceedings of International Conference on Natural Language Processing (ICON)-2005, Kanpur*.

Yulia Tsvetkov and Shuly Wintner. 2012. Extraction of multi-word expressions from small parallel corpora. *Natural Language Engineering*, 18(04):549–573.