

Multiword Expressions Dataset for Indian Languages

Dhirendra Singh, Sudha Bhingardive, Pushpak Bhattacharyya

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

Abstract

Multiword Expressions (MWEs) are used frequently in natural languages, but understanding the diversity in MWEs is one of the open problem in the area of Natural Language Processing. In the context of Indian languages, MWEs play an important role. In this paper, we present MWEs annotation dataset created for Indian languages *viz.*, Hindi and Marathi. We extract possible MWE candidates using two repositories: 1) the POS-tagged corpus and 2) the IndoWordNet synsets. Annotation is done for two types of MWEs: *compound nouns* and *light verb constructions*. In the process of annotation, human annotators tag valid MWEs from these candidates based on the standard guidelines provided to them. We obtained 3178 compound nouns and 2556 light verb constructions in Hindi and 1003 compound nouns and 2416 light verb constructions in Marathi using two repositories mentioned before. This created resource is made available publicly and can be used as a gold standard for Hindi and Marathi MWE systems.

Keywords: Multiword Expressions, MWEs, WordNet, Hindi WordNet, Compound Nouns, Light Verb Constructions

1. Introduction

Recently, various approaches have been proposed for the identification and extraction of MWEs (Calzolari et al., 2002; Baldwin et al., 2003; Guevara, 2010; Al-Haj and Wintner, 2010; Kunchukuttan and Damani, 2008; Chakrabarti et al., 2008; Sinha, 2011; Singh et al., 2012; Reddy et al., 2011). The quality of such approaches depends on the use of algorithms and also on the quality of resources used. Various standard MWEs datasets¹ are available for languages like English, French, German, Portuguese, *etc* and can be used for evaluation of MWE approaches. But for Indian languages, no such standard datasets are available publicly. Our goal is to create MWEs annotation for Indian languages *viz.*, Hindi and Marathi and make it available publicly. We have explored two types of MWEs: *compound nouns* (CNs) and *light verb constructions* (LVCs), since they are used very frequently in the text data in comparison to other MWEs. The created resource can be useful for various natural language processing applications like information extraction, word sense disambiguation, machine translation, *etc*.

The rest of the paper is organized as follows. Section 2 gives detail about the compound nouns and light verb constructions. Section 3 describes the extraction process of possible MWE candidates. Section 4 gives the statistics of MWEs annotation for Hindi and Marathi. MWEs guidelines are given in Section 5 followed by discussions in Section 6. Section 7 concludes the paper and points to the future work.

2. Compound Nouns and Light Verb Constructions

In the context of Indian languages, MWEs are quite varied and many of these are borrowed from other languages like English, Urdu, Arabic, Sanskrit, *etc*. For Hindi, there are limited investigations on MWE extraction. Venkatapathy et. al., (2006) worked on syntactic and semantic features for N-V collocation extraction using MaxEnt classifier. Mukerjee et al., (2006) proposed Parts-of-Speech projection from English to Hindi with corpus alignment for extracting complex predicates. Kunchukuttan et. al., (2008) presented a method for extracting compound nouns in Hindi using statistical co-occurrence. Sinha (2009) uses linguistic property of light verbs in extraction of complex predicates using Hindi-English parallel corpus. All the work mentioned above have considered only limited aspects of Hindi MWE. In this paper, we focus on creating gold standard data for CNs and LVCs.

- **Compound Nouns:** A word-pair forms a CN if its meaning cannot be composed from the meanings of its constituent words. CNs are formed by either Noun+Noun (N+N) or Adjective+Noun (Adj+N) word combinations. For example, बाग बगीचा (*baaga bagiiichaa*, garden) (N+N), काला धन (*kaalaa dhana*, black money) (Adj+N), *etc*. are examples of CNs in Hindi.
- **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).
 - **Conjunct Verbs:** CjVs are formed by Noun+Verb (N+V) or Adjective+Verb

¹http://multiword.sourceforge.net/PHITE.php?sitesig=FILES&page=FILES_20_Data_Sets

(Adj+V) or Adverb+Verb (Adv+V) word combinations. For example, काम करना (*kaama karanaa*, to work) (N+V), ठीक करना (*thik karanaa*, to repair) (Adj+V), वापस आना (*vaapas aanaa*, to come back) (Adv+V), etc. are examples of CjVs in Hindi.

- **Compound Verbs:** CpVs are formed by Verb+Verb (V+V) word combinations. For example, भाग जाना (*bhaaga jaanaa*, run away) (V+V), उठ जाना (*uTha jaanaa*, to wake up) (V+V), etc. are examples of CpVs in Hindi.

3. MWE Candidates Extraction

We extracted possible MWE candidates using two resources: 1) the POS-tagged corpus and 2) the IndoWordNet synsets.

3.1. Candidate Extraction using POS-tagged Corpus

For Indian languages, standard POS-tagged corpora are publicly available². We used such corpora for extracting possible candidates for MWEs. For CNs, we extracted candidates of patterns *noun followed by noun* and *adjective followed by noun*. However, for LVCs, we extracted candidates of patterns *noun followed by verb*, *adjective followed by verb*, *adverb followed by verb* and *verb followed by verb*.

3.2. Candidate Extraction using IndoWordNet Synsets

IndoWordNet³ (Bhattacharyya, 2010) is the Indian language WordNet of 18 official languages of India. It consists of synsets and semantic and lexical relations. It also stores MWEs as they represent concepts (synsets). For example, it stores Hindi CNs like बाग बगीचा (*baaga bagiiichaa*, garden), धन दौलत (*dhana daulata*, wealth), काला धन (*kaalaa dhana*, black money), etc. and LVCs like गुजर जाना (*gujara jaanaa*, passed away), काम करना (*kaama karanaa*, to work), भाग जाना (*bhaaga jaanaa*, run away), etc.

We extracted possible MWE candidates from WordNet synsets from Hindi and Marathi. Synsets which consist of words of following patterns are extracted and used as possible candidates.

- *noun followed by noun*
- *adjective followed by noun*
- *noun followed by verb*

²<http://www.ldcil.org/resourcesTextCorp.aspx>

³Wordnets for Indian languages have been developed under the IndoWordNet umbrella. Wordnets are 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 3 different language families, Indo Aryan, SinoTibetian and Dravidian. <http://www.cfilt.iitb.ac.in/indowordnet/>

- *adjective followed by verb*
- *adverb followed by verb*
- *verb followed by verb*

All these MWEs candidates were given to three human annotators in both these languages. They were told to tag the valid MWEs based on the guidelines provided to them (Refer Section 5).

4. MWEs Annotation Statistics

This section gives statistics of annotated MWEs by three human annotators. Valid MWEs are obtained by taking the majority of votes. These MWE dataset has been made available on the CFILT website <http://www.cfilt.iitb.ac.in/Downloads.html>.

4.1. Annotation Statistics of MWEs obtained from the POS-tagged Corpus

For CNs, we extracted 12000 possible candidates from Hindi and 2000 possible candidates from Marathi POS-tagged corpus. For LVCs, we extracted 4000 possible candidates each from Hindi and Marathi POS-tagged corpus. The statistics of valid MWEs annotated by human annotators are as shown in Table 1 and Table 2 respectively.

MWEs type	Possible candidates	Valid MWEs
Compound Nouns	12000	2178
Light Verb Constructions	4000	1556

Table 1: Hindi MWEs annotation statistics obtained from pos-tagged corpus

MWEs type	Possible candidates	Valid MWEs
Compound Nouns	2000	503
Light Verb Constructions	4000	1916

Table 2: Marathi MWEs annotation statistics obtained from pos-tagged corpus

MWEs type	Possible candidates	Annotated MWEs
Compound Nouns	19326	1000
Light Verb Constructions	4017	1000

Table 3: Hindi MWEs annotation statistics obtained from the IndoWordNet Synsets

MWEs type	Possible candidates	Annotated MWEs
Compound Nouns	5327	500
Light Verb Constructions	1838	500

Table 4: Marathi MWEs annotation statistics obtained from the IndoWordNet Synsets

4.2. Annotation Statistics of MWEs obtained from the IndoWordNet Synsets

For Hindi, we extracted 19326 possible candidates for CNs and 4017 possible candidates for LVCs from the IndoWordNet synsets. For Marathi, we extracted 5327 possible candidates for CNs and 1838 possible candidates for LVCs from the WordNet synsets. Statistics of valid MWEs annotated by human experts for Hindi and Marathi languages are as shown in Table 3 and Table 4 respectively. The inter-annotator agreement was calculated using Cohen’s kappa index value. The inter-annotator agreement for the annotation is found to be 0.86 for Hindi and 0.82 for Marathi.

5. MWE Annotation Guidelines

In this section, we describe guidelines given to human annotators to annotate MWEs from the possible candidates. Annotators have been told to check whether the candidate (word-pair) satisfy the following criteria of MWEs formation.

- **Reduplication:** Here, a root or stem of a word, or part of it is repeated. Reduplication can further be subdivided into:
 - **Onomatopoeic Expression:** In this case, the constituent words imitate a sound or a sound of an action. Generally, in this case, the words are repeated twice with the same ‘matra’. For example, टिक टिक (*tick tick*, the ticking sound of watch’s needle).
 - **Non-Onomatopoeic Expression:** Here, the constituent words have meaning but they are repeated to convey a particular meaning. For example, चलते चलते (*chhalate chhalate*, while walking).
 - **Partial Reduplication:** In this case, one of the constituent word is meaningful while the other is constructed by partially repeating the first word. For example, पानी वाणी (*paani vaani*, water).
 - **Semantic Reduplication:** Here, the constituent words have some semantic relationship among them. For example, धन दौलत (*dhana daulata*, wealth) shows [Synonymy], दिन रात (*dina raata*, always) shows [Antonymy].
- **Fixed Expression:** Fixed Expressions are immutable expressions, which do not undergo any

transformation or morphological inflections or possibility of insertion between two words. For example, कम से कम (*kam se kam*, atleast), ज्यादा से ज्यादा (*jyada se jyada*, maximal).

- **Semi-fixed Expression:** Semi-fixed expressions obey constraints on word order and composition. They might show some degree of lexical variation. For example, कार पार्क (*car park*, car park) can be used as कार पार्क्स (*car parks*, car parks).
- **Non-Compositional:** The meaning of a complete multiword expression can not be completely determined from the meaning of its constituent words. For example, अक्षय तृतीया (*akshaya Tritiyaa*, a festival in India)
- **Decomposable Idioms:** Decomposable idioms are syntactically flexible and behave like semantically linked parts. But it is difficult to predict exactly what type of syntactic expression they are. For example, आटे-दाल का भाव मालूम होना (*aate daal ka bhava maalum honaa*, to create a knowledge). Here in this example, we can replace the phrase 'आटे-दाल का भाव मालूम होना' (*aate daal ka bhava maalum honaa*) to 'आटे-दाल का दाम मालूम पड़ना' (*aate daal ka bhava maalum padanaa*).
- **Non-Decomposable Idioms:** Non-Decomposable idioms are those idioms, which do not undergo any syntactic variations but might allow some minor lexical modification. For example, नौ दो ग्यारह होना (*Nau do gyaraaha honaa*, to run away).
- **Name Entity Recognition(NER):** Named entities are phrases that contain the names of persons, organizations, locations, times, and quantities. NERs are syntactically highly idiosyncratic. These entities are formed based on generally a place or a person. For example, भारतीय प्रौद्योगिकी संस्थान (*Bhartiya Prodyogiki Sansthan*, Indian Institute of Technology) (Organization), सचिन तेंदुलकर (*Sachin Tendulkar*, Sachin Tendulkar) (Proper noun), ताज महल (*Taj Mahal*) (Location), etc.
- **Collocations:** A collocation is an expression consisting of two or more words that correspond to some conventional way of saying things. For example, कड़क चाय (*kadaka chai*, strong tea), पोस्ट ऑफिस (*post office*, post office), etc.
- **Foreign Words:** A set of words borrowed from another languages are called as foreign words. They can be treated as valid MWEs in the context of Indian languages. For example, रेलवे स्टेशन (*Railway station*, Railway Station), पोस्ट ऑफिस (*Post office*, post office), etc.

6. Discussions

While annotating CNs and LVCs, annotators faced some difficulties which are mentioned below.

- **Polysemous candidates:** Sometimes extracted candidates were found to be polysemous. As we did not mention the context in which these candidates occurs, annotators confused while annotating these candidates. Most of the time these candidates behave as MWEs due to their frequent metaphoric usage. For example,

1. आग लगाना (*aag lagaana*) has two senses in Hindi: 1) *destroy by fire* and 2) *to provoke*. It forms MWEs when used in its second sense which is metaphoric in nature.
2. पर्दा उठाना (*pardaa uthanaa*) has two senses in Hindi: 1) *reveal secret information* and 2) *make visible*. It forms MWEs when used in its first sense.

For such polysemous candidates, annotators tagged them as valid MWEs based on their knowledge and context.

- **Infrequent candidates:** Sometimes candidates are not tagged as MWEs even though they satisfy some of the guidelines. This is because of their infrequent usage. For example, नीला पीला (*neela piila*) is not considered as a valid MWEs even though it looks similar to a valid MWEs लाल पीला (*lala piila*). Such infrequent candidates are not annotated as MWEs.

7. Conclusion

In this paper, we presented manually annotated dataset for MWEs in Hindi and Marathi. The annotation has been done for compound nouns and light verb constructions. MWEs candidates were extracted from the POS-tagged corpus and the IndoWordNet synsets. The annotation process involved three annotators in each languages and the validation of MWEs is done using a majority vote decision. For Hindi, we obtained 3178 compound nouns and 2556 light verb constructions as valid MWEs and for Marathi, we obtained 1003 compound nouns and 2416 light verb constructions as valid MWEs. This MWEs dataset has been made publicly available and now it can be used as a gold standard dataset for MWE systems and its applications. In future, we would like to work on annotating MWEs in the running text and will also try to explore the other types of MWEs and other languages also.

8. Acknowledgments

We would like to thank Jaya Saraswati, Rajita Shukla, Laxmi Kashyap, Nilesh Joshi and Irawati Kulkarni from CFILT lab at IIT Bombay for giving their valuable contribution in gold standard data creation. We also acknowledge the support of the Department of Electronics and Information technology (DeitY), Ministry of Communication and Information Technology, Government of India and also of Ministry of Human Resource Development.

9. Bibliographical References

- Al-Haj, H. and Wintner, S. (2010). Identifying multiword expressions by leveraging morphological and syntactic idiosyncrasy. In *Proceedings of the 23rd International conference on Computational Linguistics*, pages 10–18. Association for Computational Linguistics.
- Baldwin, T., Bannard, C., Tanaka, T., and Widdows, D. (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.
- Bhattacharyya, P. (2010). Indowordnet. In *Language Resources and Evaluation Conference (LREC)*, Malta.
- Calzolari, N., Fillmore, C. J., Grishman, R., Ide, N., Lenci, R., Macleod, C., and Zampolli, A. (2002). Towards best practice for multiword expressions in computational lexicons. In *In Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002), Las Palmas, Canary Islands*. Citeseer.
- Chakrabarti, D., Mandalia, H., Priya, R., Sarma, V. M., and Bhattacharyya, P. (2008). Hindi compound verbs and their automatic extraction. In *COLING (Posters)*, pages 27–30.
- Guevara, E. (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.
- Kunchukuttan, A. and Damani, O. P. (2008). A system for compound noun multiword expression extraction for hindi. In *6th International. Conference on Natural Language Processing*, pages 20–29.
- Mukerjee, A., Soni, A., and Raina, A. M. (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.
- Reddy, S., McCarthy, D., and Manandhar, S. (2011). An empirical study on compositionality in compound nouns. In *IJCNLP*, pages 210–218.
- Singh, S., Damani, O. P., and Sarma, V. M. (2012). Noun group and verb group identification for hindi. In *COLING*, pages 2491–2506. Citeseer.
- Sinha, R. M. K. (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.
- Sinha, R. M. K. (2011). Stepwise mining of multiword 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.

Venkatapathy, S. and Joshi, A. K. (2006). Using information about multi-word expressions for the word-alignment task. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 20–27. Association for Computational Linguistics.