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 create MWEs annotation in Indian languages *viz.*, Hindi and Marathi. We extract possible MWEs 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 experts tag valid MWEs from these candidates based on the guidelines given. For Hindi language, we obtained 3,178 compound nouns and 2,556 light verb constructions and for Marathi language, we obtained 1,003 compound nouns and 2,416 light verb constructions. This created resource is made available publicly and can be used as a gold standard for Hindi and Marathi MWEs systems.

Keywords: Multiword Expressions, MWE, 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 datasets[1] 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 in Indian languages (Hindi and Marathi) and make it available publicly. We have explored two types of MWEs that are *compound nouns* (CNs) and *light verb constructions* (LVCs). Since, CNs and LVCs are used very frequently in the text data in comparison to other MWEs, we have considered only these MWEs in this paper. 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 MWEs candidates. Section 4 gives the statistics of MWEs annotation for Hindi and Marathi languages. MWEs guidelines are given in Section 5 followed by discussions in Section 6. Section 7 concludes the paper and points to the future work.

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

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, Arabi, Sanskrit and other ones. 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 POS 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 of these work 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 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 e.g. बाग बगीचा (*baaga bagichaa*, garden) (N+N), काला धन (*kaalaa dhana*, black money) (Adj+N), *etc.* are examples of Hindi CNs.
- **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

(Adj+V) or Adverb+Verb (Adv+V) word combinations. For e.g. 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 Hindi CjVs.

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

3. MWEs Candidate Extraction

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

3.1. Candidate Extraction using Pos-tagged Corpus

For Indian languages, standard pos-tagged corpora are publicly available[2]. 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

IndoWordNet[3] (Bhattacharyya, 2010) is the Indian language WordNets of 17 official languages of India. It consists of synsets and semantic relations. It also stores MWEs as they also represent concepts (synsets). For example, for Hindi it stores CNs like बाग बगीचा (*baaga bagiichaa*, 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 MWEs candidates from IndoWordNet synsets in Hindi and Marathi languages. 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*
- *adjective followed by verb*
- *adverb followed by verb*
- *verb followed by verb*

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

4. MWEs Annotation Statistics

This section gives statistics of annotated MWEs by three human experts. Valid MWEs are obtained by taking the majority of votes. These MWEs 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 12,000 possible candidates from Hindi pos-tagged corpus and 2,000 possible candidates from Marathi pos-tagged corpus. For LVCs, we extracted 4,000 possible candidates each from Hindi and Marathi pos-tagged corpus. The statistics of valid MWEs annotated by human experts 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

4.2. Annotation Statistics of MWEs obtained from IndoWordNet Synsets

For Hindi, we extracted 19,326 possible candidates for CNs and 4,017 possible candidates for LVCs from

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

Table 3: Hindi MWEs annotation statistics obtained from IndoWordNet Synsets

[2]http://www.ldcil.org/resourcesTextCorp.aspx

[3]Wordnets 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, SinoTebetian and Dravidian. http://www.cfilt.iitb.ac.in/indowordnet/

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

Table 4: Marathi MWEs annotation statistics obtained from IndoWordNet Synsets

IndoWordNet synsets. For Marathi, we extracted 5,327 possible candidates for CNs and 1,838 possible candidates for LVCs from IndoWordNet 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 possile candidates. Annotaotrs has 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'. E.g. 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. E.g. चलते चलते (*chalate chalate*, 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. E.g: पानी वाणी (*paani vaani*, water).
  - Semantic Reduplication: Here, the constituent words have some semantic relationship among them. E.g. धन दौलत (*dhana daulata*, Wealth) (Synonymy), दिन रात (*dina raata* always) (Antonymy) .
- Fixed Expression: Fixed Expressions are immutable expressions, which do not undergo any transformation or morphological inflections or possibility of insertion between two words. E.g. 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. E.g. कार पार्क (*car park, It can be car park(s)*).
- Non-Compositional: The meaning of a complete multiword expression can not be completely determined from the meaning of its constituent words. E.g. अक्षय ततीया (*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. E.g. आटे–दाल का भाव मालूम होना (*aate daal ka bhava maalum honaa*, to creat a knowledge). Here in this example, we can replace the phrase 'आटे–दाल का भाव मालूम होना' to 'आटे–दाल का दाम मालूम पड़ना'.
- Non-Decomposable Idioms: Non-Decomposable idioms are those idioms, which do not undergo any syntactic variations but might allow some minor lexical modification. E.g. नौ दो ग्यारह होना (*Nau do gyaraaha honaa*, to run off).
- 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. E.g. भारतीय प्रोद्योगिकी 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. E.g. 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. E.g. रेलवे स्टेशन (*Railway station*, Railway Station), पोस्ट ऑफिस (*Post office*, post office), *etc.*

6. Discussions

While annotating CNs and LVCs, annotators faced the several problems. Some of them 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 candiadtes. Most of the time these candidates behave as MWEs when they used as a metaphoric usage. For e.g.,

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 tag these candidates as valid MWEs.

- **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 languages. The annotation has been done for compound nouns and light verb constructions. MWEs candidates were extracted from pos-tagged corpus and IndoWordNet synsets. The annotation process involved three annotators in both the languages and the validation of MWEs is done using a majority vote decision. For Hindi language, we obtained 3,178 compound nouns and 2,556 light verb constructions as valid MWEs and for Marathi language, we obtained 1,003 compound nouns and 2,416 light verb constructions as valid MWEs. These MWEs can be used as a gold standard 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 consider the other types of MWEs.

8. Acknowledgments

We thank CFILT members at IIT Bombay for their valuable comments and suggestions. We also acknowledge the support of the Department of Information Technology (DIT), 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.