Detection of Multiword Expressions for Hindi Language using Word Embeddings and WordNet-based Features

Dhirendra Singh Sudha Bhingardive Kevin Patel Pushpak Bhattacharyya

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

Abstract

Detection of Multiword Expressions (MWEs) is a challenging problem faced by several natural language processing applications. The difficulty emanates from the task of detecting MWEs with respect to a given context. In this paper, we propose approaches that use Word Embeddings and WordNet-based features for the detection of MWEs for Hindi language. These approaches are restricted to two types of MWEs *viz., noun compounds and noun+verb compounds*. The results obtained indicate that using linguistic information from a rich lexical resource such as WordNet, help in improving the accuracy of MWEs detection. It also demonstrates that the linguistic information which word embeddings capture from a corpus can be comparable to that provided by WordNet. Thus, we can say that, for the detection of above mentioned MWEs, word embeddings can be a reasonable alternative to WordNet, especially for those languages whose WordNets does not have a better coverage.

1 Introduction

Multiword Expressions or MWEs can be understood as idiosyncratic interpretations or *words with spaces* wherein concepts cross the word boundaries or spaces (Sag et al., 2002). Some examples of MWEs are *ad hoc, by and large, New York, kick the bucket, etc.* Typically, a multiword is a noun, a verb, an adjective or an adverb followed by a light verb (LV) or a noun that behaves as a single unit (Sinha, 2009). Proper detection and sense disambiguation of MWEs is necessary for many Natural Language Processing (NLP) tasks like machine translation, natural language generation, named entity recognition, sentiment analysis, etc. MWEs are abundantly used in Hindi and other languages of Indo Aryan family. Common part-of-speech (POS) templates of MWEs in Hindi language include the following: noun+noun, noun+LV, adjective+LV, adjective+noun, etc. Some examples of Hindi multiwords are पुण्य तिथि (*puNya tithi*, death anniversary), वादा करना (*vaadaa karanaa*, to promise), आग लगाना (*aaga lagaanaa*, to burn), धन दौलत (*dhana daulata*, wealth), etc.

WordNet (Miller, 1995) has emerged as crucial resource for NLP. It is a lexical structure composed of synsets, semantic and lexical relations. One can look up WordNet for information such as synonym, antonym, hypernym, etc. of a word. WordNet was initially built for the English language, which is then followed by almost all widely used languages all over the world. WordNets are developed for different language families *viz.* EuroWordNet[1] (Vossen, 2004) was developed for Indo-European family of languages and covers languages such as German, French, Ital-

[1] http://www.illc.uva.nl/EuroWordNet/

D S Sharma, R Sangal and E Sherly. Proc. of the 12th Intl. Conference on Natural Language Processing, pages 295–302,
Trivandrum, India. December 2015. ©2015 NLP Association of India (NLPAI)

ian, *etc.* Similarly, IndoWordNet[2] (Bhattacharyya, 2010) covers the major families of languages, *viz.,* Indo-Aryan, Dravidian and Sino-Tibetian which are used in the subcontinent. Building WordNets is a complex task. It takes lots of time and human expertise to build and maintain WordNets.

A recent development in computational linguistics is the concept of distributed representations, commonly referred to as *Word Vectors* or *Word Embeddings*. The first such model was proposed by Bengio et al. (2003), followed by similar models by other researchers *viz.,* Mnih et al. (2007), Collobert et al. (2008), Mikolov et al. (2013a), Pennington et al. (2014). These models are extremely fast to train, are automated, and rely only on raw corpus. Mikolov et al. (2013c; 2013b) have reported various linguistic regularities captured by such models. For instance, vectors of synonyms and antonyms will be highly similar when evaluated using cosine similarity measure. Thus, these models can be used to replace/supplement WordNets and other such resources in different NLP applications (Collobert et al., 2011).

The roadmap of the paper is as follows, Section 2 describes the background and related work. Our approaches are detailed in section 3. The description of the datasets used for the evaluation is given in section 4. Experiments and results are presented in Section 5. Section 6 concludes the paper and points to the future work.

2 Background and Related Work

Most of the proposed approaches for the detection of MWEs are statistical in nature. Some of these approaches use association measures (Church and Hanks, 1990), deep linguistics based methods (Bansal et al., 2014), word embeddings based measures (Salehi et al., 2015), etc.

The work related to the detection of MWEs has been limited in the context of Indian languages. The reasons are, unavailability of gold data (Reddy, 2011), unstructured classification of MWEs, complicated theory of MWEs, lack of resources, etc. Most of the approaches of Hindi MWEs have used parallel corpus alignment and POS tag projection to extract MWEs (Sriram et al., 2007) (Mukerjee et al., 2006). Venkatapathy et al. (2007) used a classification based approach for extracting noun+verb collocations for Hindi. Gayen and Sarkar et al. (2013) used Random Forest approach wherein features such as *verb identity, semantic type, case marker, verb-object similarity, etc.* are used for the detection of compound nouns in Bengali using MaxEnt Classifier. However, our focus is on detecting MWEs of the type *compound noun* and *noun+verb compounds* while verb based features are not implemented in our case. We have used word embeddings and WordNet based features for the detection of above MWEs.

Characteristics of MWEs

MWE has different characteristics based on their usage, context and formation. They are as follows-

Compositionality: Compositionality refers to the degree to which the meaning of MWEs can be predicted by combining the meanings of their components. *E.g.* तरण ताल (*taraNa taala*, swimming pool), धन लक्ष्मी (*dhana laxmii*, wealth), चाय पानी (*chaaya paanii*, snacks), etc.

Non-Compositionality: In non-compositionality, the meaning of MWEs cannot be completely determined from the meaning of its constituent words. It might be completely different from its constituents. *E.g.* गुजर जाना, (*gujara jaanaa*, passed away), नजर डालना, (*najara Daalanaa*, flip through). There might be some added elements or inline meaning

[2] 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/

to MWEs that cannot be predicted from its parts. *E.g.* नौ दो ग्यारह होना (*nau do gyaaraha honaa*, run away).

Non-Substitutability: In non substitutability, the components of MWEs cannot be substituted by its synonyms without distorting the meaning of the expression even though they refer to the same concept (Schone and Jurafsky, 2001). *E.g.* in the expression चाय पानी (*chaaya paanii*, snacks), the word *paanii* (water) cannot be replaced by its synonym जल (*jala*, water) or नीर (*niira*, water) to form the meaning 'snacks'.

Collocation: Collocations are a sequence of words that occur more often than expected by chance. They do not show either statistical or semantical idiosyncrasy. They are fixed expressions and appear very frequently in running text. *E.g.* कड़क चाय (*kaDaka chaaya*, strong tea), काला धन (*kaalaa dhana*, black money), etc.

Non-Modifiability: In non-modifiablility, many collocations cannot be freely modified by grammatical transformations such as *change of tense, change in number, addition of adjective, etc.* These collocations are frozen expressions which cannot be modified at any condition. *E.g.*, the idiom घाव पर नमक छिड़कना (*ghaava para namaka ChiDakanaa*, rub salt in the wound) cannot be replace to *घाव पर ज्यादा नमक छिड़कना (*ghaava para jyaadaa namaka ChiDakanaa*, rub more salt in the wound) or something similar.

Classification of MWEs

According to Sag et.al (2002) MWEs are classified into two broad categories *viz.,* 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 *noun compounds* and *noun+verb compounds* which fall under the semi-fixed and syntactically fixed categories respectively.

Noun Compounds: Noun compounds are MWEs which are formed by two or more nouns which behave as a single semantic unit. In the case of compositionality, noun compounds usually put the stress on the first component while the remaining components expand the meaning of the first component. *E.g.* बाग बगीचा (*baaga bagiichaa*, garden) is a noun compound where *baaga* is giving the full meaning of the whole component against the second component *bagiichaa.* However, in the case of non-compositionality, noun compounds do not put stress on any of the components. *E.g.* अक्षय तृतीया (*axaya tRitiiyaa*, one of the festival), पुण्य तिथि (*puNya tithi*, death anniversary).

Noun+Verb Compounds: Noun+ verb compounds are type of MWEs which are formed by sequence of words having noun followed by verb(s). These are type of conjunct verbs where noun+verb pattern behaves as a single semantic unit wherein *noun* gives the meaning for whole expression. *E.g.* वादा करना (*vaadaa karanaa*, to promise), मार डालना (*maar daalanaa*, to kill), etc.

3 Our Approach

The central idea behind our approach is that words belonging to a MWE co-occur frequently. Ideally, such co-occurrence can be computed from a corpus. However, no matter how large a corpus actually is, it cannot cover all possible usages of all words of a particular language. So, a possible workaround to address this issue can be as follows:

Given a word pair w_1 w_2 to be identified as a MWE,

1. Find the co-occurrence estimate of w_1 w_2 using the corpus alone.
2. Further refine this estimate by using co-occurrence estimate of w'_1 w'_2, where w'_1 and w'_2 are synonyms or antonyms of w_1 and w_2 respectively.

In order to estimate co-occurrence of w_1 w_2, one can use *word embeddings or word vectors.* Such techniques try to predict (Baroni et al., 2014), rather than count the co-occurrence patterns of different tuples of words. The *distributional* aspect of these representations enables one to estimate the co-occurrence of, say, *cat* and *sleeps*, using the co-occurrence of *dogs* and *sleep.* Such *word embeddings* are typically trained on raw corpora, and the similarity between a pair of words is computed by calculating the cosine similarity between the embeddings corresponding to the pair of words. It has been proved that such methods indirectly capture co-occurrence only, and can thus be used for the task at hand.

While exact co-occurrence can be estimated using word embeddings, substitutional co-occurrence cannot be efficiently captured using the same. More precisely, if w_1 w_2 is a MWE, but the corpus contains w_1 *synonym*(w_2) or *synonym*(w_1) w_2 frequently, then one cannot hope to learn that w_1 w_2 is indeed a MWE. Such paradigmatic (substitutional) information cannot be captured efficiently by word vectors. This has been established by the different experiments performed by (Chen et al., 2013), (Baroni et al., 2014) and (Hill et al., 2014). So one needs to look at other resources to obtain this information. We decided to use WordNet for the same. Similarity between a pair of words appearing in the WordNet hierarchy can be acquired using multiple means. For instance, two words are said to be synonyms if they belong in the same *synset* in the WordNet.

Having these two resources at our disposal, we can realize the above mentioned approach more concretely as follows:

1. Use WordNet to detect synonyms, antonyms.
2. Use similarity measures either facilitated by WordNet or by the word embeddings.

These options lead to the following three concrete heuristics for the detection of *noun compounds* and *noun+verb* compounds for word pair $w_1 w_2$.

3.1 Approach 1: Using WordNet-based Features

1. Let WNBag $= \{w' \mid w' = IsSynOrAnto(w_1)\}$, where the function $IsSynOrAnto$ returns either a synonym or an antonym of w_1, by looking up the WordNet.
2. If $w_2 \in$ WNBag, then w_1 w_2 is a MWE.

3.2 Approach 2: Using Word Embeddings

1. Let WEBag $= \{w' \mid w' = IsaNeighbour(w_1)\}$, where the function $IsaNeighbour$ returns *neighbors* of w_1, i.e, returns the top 20 words that are close to w_1 (as measured by cosine similarity of the corresponding word embeddings).
2. If $w_2 \in$ WEBag, then w_1 w_2 is a MWE.

3.3 Approach 3: Using WordNet and Word Embeddings with Exact match

1. Let WNBag $= \{w' \mid w' = IsSynOrAnto(w_1)\}$, where the function $IsSynOrAnto$ returns either a synonym or an antonym of w_1, by looking up the WordNet.
2. Let WEBag $= \{w' \mid w' = IsaNeighbour(w_2)\}$, where the function $IsaNeighbour$ returns *neighbors* of w_2, i.e, returns the top 20 words that are close to w_2 (as measured by cosine similarity of the corresponding word embeddings).

3. If WNBag $\cap$ WEBag $\neq \phi$, then w_1 w_2 is a MWE.

4 Datasets

MWE Gold Data

There is a dearth of datasets for Hindi MWEs. The ones that exists, have some shortcomings. For instance, (Kunchukuttan and Damani, 2008) have performed MWEs evaluation on their in-house dataset. However, we found this dataset to be extremely skewed, with only ∼300 MWEs out of ∼12500 phrases. Thus, we have created the in-house gold standard dataset for our experiments. While creating this dataset we extracted 2000 noun+noun and noun+verb word pairs each from the ILCI Health and Tourism domain corpus automatically. Further, three annotators were asked to manually check whether these extracted pairs are MWEs or not. They deemed 450 valid noun+noun and 500 noun+verb pairs to be MWEs. This process achieved an inter-annotator agreement of ∼0.8.

Choice of Word Embeddings

Since Bengio et. al. (2003) came up with the first word embeddings, many models for learning such word embeddings have been developed. We chose the Skip-Gram model provided by word2vec tool developed by (Mikolov et al., 2013a) for training word embeddings. The parameters for the training are as follows: Dimension = 300, Window Size = 8, Negative Samples = 25, with the others being kept at their default settings.

Data for Training Word Embeddings

We used Bojar Hindi MonoCorp dataset (Bojar et al., 2014) for training word embeddings. This dataset contains 44 million sentences with approximately 365 million tokens. To the best of our knowledge, this is the largest Hindi corpus available publicly on the internet.

Data for Evaluating Word Embeddings

Before commenting on the applicability of word embeddings to this task, one needs to evaluate the quality of the word embeddings. For evaluating word embeddings of the English language, many word-pair similarity datasets have emerged over the years (Lev Finkelstein and Ruppin, 2002), (Hill et al., 2014). But no such datasets exists for Hindi language. Thus, once again, we have developed an in-house evaluation dataset. We manually translated the English word-pairs in (Lev Finkelstein and Ruppin, 2002) to Hindi, and then asked three annotators to score them in the range [0,10] based on their semantic similarity and relatedness[3]. The inter-annotator agreement on this dataset was 0.73. This is obtained by averaging first three columns of Table 1.

5 Experiments and Results

5.1 Evaluation of Quality of Word Embeddings

Entities	Agreement
human1/human2	0.74
human1/human3	0.68
human2/human3	0.77
word2vec/human1	0.65
word2vec/human2	0.54
word2vec/human3	0.63

Table 1: Agreement of different entities on the translated similarity dataset for Hindi

We have evaluated word embeddings that are trained on Bojar corpus on the word-pair similarity dataset (which is mentioned in the previous section). It is observed that, the average agreement between word embeddings (*word2vec* tool) and human annotators was ∼0.61. This is obtained by averaging last three columns of Table 1.

[3] We are in the process of releasing this dataset publicly

Techniques	**Resources used**	**P**	**R**	**F-score**
Approach 1	WordNet	0.79	0.77	0.78
Approach 2	word2vec	0.75	0.64	0.69
Approach 3	word2vec+WordNet	0.76	0.68	0.72

Table 2: Results of noun compounds on Hindi Dataset

Techniques	**Resources used**	**P**	**R**	**F-score**
Approach 1	WordNet	0.75	0.82	0.78
Approach 2	word2vec	0.56	0.75	0.64
Approach 3	word2vec+WordNet	0.57	0.58	0.58

Table 3: Results of noun+verb compounds on Hindi Dataset

5.2 Evaluation of Our Approaches for MWEs detection

Table 2 shows the performance of the three different approaches at detecting noun compound MWEs. Table 3 shows the performance of the three different approaches at detecting noun+verb compound MWEs. As is evident from the Table 2 and Table 3, WordNet based approaches perform the best. However, it is also clear that results obtained by using word embeddings perform comparatively better. Thus, in general, these results can be favorable for word embeddings based approaches as they are trained on raw corpora. Also, they do not need much human help as compared to WordNets which require considerable human expertise in creating and maintaining them. In our experiments, we have used Hindi WordNet which is one of the well developed WordNet, and thus result obtained using this WordNet are found to be promising. However, for other languages with relatively underdeveloped WordNets, one can expect word embeddings based approaches to yield results comparable to those approaches which uses well developed WordNet.

6 Conclusion

This paper provides a comparison of Word Embeddings and WordNet-based approaches that one can use for the detection of MWEs. We selected a subset of MWE candidates *viz., noun compounds and noun+verb compounds*, and then report the results of our approaches for these candidates. Our results show that the WordNet-based approaches performs better than Word Embedding based approaches for the MWEs detection for Hindi language. However, word embeddings based approaches has the potential to perform at par with approaches utilizing well formed WordNets. This suggests that one should further investigate such approaches, as they rely on raw corpora, thereby leading to enormous savings in both time and resources.

References

Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2014. Tailoring continuous word representations for dependency parsing. Association for Computational Linguistics.

Marco Baroni, Georgiana Dinu, and German Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247. Association for Computational Linguistics.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3:1137–1155, March.

Pushpak Bhattacharyya. 2010. Indowordnet. In *In Proc. of LREC-10*. Citeseer.

Ondřej Bojar, Vojtěch Diatka, Pavel Rychlý, Pavel Straňák, Vít Suchomel, Aleš Tamchyna, and Daniel Zeman. 2014. HindMonoCorp 0.5.

Yanqing Chen, Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2013. The expressive power of word embeddings. In *ICML 2013 Workshop on Deep Learning for Audio, Speech, and Language Processing*, Atlanta, GA, USA, July.

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

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: deep neural networks with multitask learning. In William W. Cohen, Andrew McCallum, and Sam T. Roweis, editors, *ICML*, volume 307 of *ACM International Conference Proceeding Series*, pages 160–167. ACM.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 12:2493–2537, November.

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.

Felix Hill, Roi Reichart, and Anna Korhonen. 2014. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *arXiv preprint arXiv:1408.3456*.

Anoop Kunchukuttan and Om P Damani. 2008. A system for compound noun multiword expression extraction for hindi.

Yossi Matias Ehud Rivlin Zach Solan Gadi Wolfman Lev Finkelstein, Evgeniy Gabrilovich and Eytan Ruppin. 2002. Placing search in context: The concept revisited. *ACM Trans. Inf. Syst.*, 20(1):116–131, January.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013c. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, pages 746–751.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Andriy Mnih and Geoffrey E. Hinton. 2007. Three new graphical models for statistical language modelling. In Zoubin Ghahramani, editor, *ICML*, volume 227 of *ACM International Conference Proceeding Series*, pages 641–648. ACM.

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.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)*, 12.

Siva Reddy. 2011. An empirical study on compositionality in compound nouns. IJCNLP.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann A. Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing '02, pages 1–15, London, UK, UK. Springer-Verlag.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2015. A word embedding approach to predicting the compositionality of multiword expressions. In *Proceedings of the 2015 Conference of the North American Chapter of the Association of Computational Linguistics - Human Language Technologies (NAACL HLT)*.

Patrick Schone and Daniel Jurafsky. 2001. Is knowledge-free induction of multiword unit dictionary headwords a solved problem. Empirical Methods in Natural Language Processing.

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.

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.*

Piek Vossen. 2004. Eurowordnet: a multilingual database of autonomous and language-specific wordnets connected via an interlingualindex. *International Journal of Lexicography*, 17(2):161–173.