Semantic Clustering: an Attempt to Identify Multiword Expressions in Bengali

Tanmoy Chakraborty **Dipankar Das** **Sivaji Bandyopadhyay**
Department of Computer Science and Engineering
Jadavpur University, Kolkata 700 032, India
its_tanmoy@yahoo.co.in, dipankar.dipnil2005@gmail.com
sivaji_cse_ju@yahoo.com

Abstract

One of the key issues in both natural language understanding and generation is the appropriate processing of Multiword Expressions (MWEs). MWE can be defined as a semantic issue of a phrase where the meaning of the phrase may not be obtained from its constituents in a straightforward manner. This paper presents an approach of identifying bigram noun-noun MWEs from a medium-size Bengali corpus by clustering the semantically related nouns and incorporating a vector space model for similarity measurement. Additional inclusion of the English WordNet::Similarity module also improves the results considerably. The present approach also contributes to locate clusters of the synonymous noun words present in a document. Experimental results draw a satisfactory conclusion after analyzing the Precision, Recall and F-score values.

1 Introduction

Over the past two decades or so, Multi-Word Expressions (MWEs) have been identified with an increasing amount of interest in the field of Computational linguistics and Natural Language Processing (NLP). The term MWE is used to refer the various types of linguistic units and expressions including idioms (*kick the bucket*, 'to die'), noun compounds (*village community*), phrasal verbs (*find out*, 'search') and other habitual collocations like conjunction (*as well as)*, institutionalized phrases (*many thanks*) etc. They can also be grossly defined as "idiosyncratic interpretations that cross the word boundaries" (Sag *et al.*, 2002).

MWE is considered as a special issue of semantics where the individual components of an expression often fail to keep their meanings intact within the actual meaning of the expression. This opaqueness in meaning may be partial or total depending on the degree of compositionality of the whole expression. In Bengali, an analogous scenario can be observed when dealing with the expressions like compound nouns (*taser ghar*, 'house of cards', 'fragile'), complex predicates such as conjunct verbs (*anuvab kara*, 'to feel') and compound verbs (*uthe para*, 'to arise'), idioms (*matir manus*, 'down to the earth'), Named Entities (NEs) (*Rabindranath Thakur*, 'Rabindranath Tagore') etc.

In this paper, we analyze MWEs from the perspective of semantic interpretation. We have focused mainly on the fact that the individual meanings of the components are totally or partially diminished in order to form the actual semantics of the expression. A constellation technique has been employed to group all nouns that are somehow related to the meaning of the component of any expression in the corpus and hence to build cluster for that component. Two types of vector space based similarity techniques are applied to make a binary classification of the candidate nouns. The intuition was that more the similarity of the components of an expression, less the probability of the candidate to become a MWE. We have also shown the results using WordNet::Similarity module.

The remainder of the paper is organized as follows. In the next section, we review the related work on MWE and graph-clustering approach for detecting compositionality. Section 3 proposes a brief description of the semantic clustering approach. The system framework is elaborated in Section 4. Experimental results and the various observations derived from our research are discussed in Section 5. Finally, Section 6 concludes the paper.

Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE 2011), pages 8–13,
Portland, Oregon, USA, 23 June 2011.

2 Related Work

A number of research activities regarding MWE identification have been carried out in various languages like English, German and many other European languages. The statistical co-occurrence measurements such as Mutual Information (MI) (Church and Hans, 1990), Log-Likelihood (Dunning, 1993) and Salience (Kilgarriff and Rosenzweig, 2000) have been suggested for identification of MWEs. An unsupervised graph-based algorithm to detect the compositionality of MWEs has been proposed in (Korkontzelos and Manandhar 2009).

In case of Indian languages, an approach in compound noun MWE extraction (Kunchukuttan and Damani, 2008) and a classification based approach for Noun-Verb collocations (Venkatapathy and Joshi, 2009) have been reported. In Bengali, the works on automated extraction of MWEs are limited in number. One method of automatic extraction of Noun-Verb MWE in Bengali (Agarwal *et al.*, 2004) has been carried out using significance function. In contrast, we have proposed a clustering technique to identify Bengali MWEs using semantic similarity measurement. It is worth noting that the conducted experiments are useful for identifying MWEs for the electronically resource constrained languages.

3 Semantic Clustering Approach

Semantic clustering aims to cluster semantically related tokens present in a document. Identifying semantically related words for a particular token is carried out by looking the surrounding tokens and finding the synonymous words within a fixed context window. Statistical idiomaticity demands frequent occurrence of a particular expression as one or few occurrences of a particular word cannot infer all its meaning. However, the semantics of a word may be obtained by analyzing its similarity sets called *synset*. Higher value of the similarity coefficient between two synonymous sets of the multi-word components indicates more affinity of the components to each other.

For individual component of a bigram expression, semantically related words of the documents are extracted by using a monolingual dictionary (as discussed in Section 4.4). Count of elements in an intersection of two synsets indicates the commonality of the two sets and its absolute value stands for their commonality measure. Considering the common elements as the dimensions of the vector space, similarity based techniques are applied to measure the semantic affection of the two components present in a bigram.

4 System Framework

4.1 Corpus Preparation and Candidate Selection

The system uses a large number of Bengali articles written by the noted Indian Nobel laureate Rabindranath Tagore[1]. We are primarily interested in single document term affinity rather than document information and document length normalization. Merging all of the articles, a medium size raw corpus consisting of 393,985 tokens and 283,533 types has been prepared. Basic pre-processing of the crawled corpus is followed by parsing with the help of an open source shallow parser[2] developed for Bengali. Parts-of-Speech (POS), chunk, root, inflection and other morphological information for each token have been retrieved. Bigram noun sequence within a noun chunk is extracted and treated as candidates based on their POS, chunk categories and the heuristics described as follows.

1. **POS**: POS of each token is either 'NN' or 'NNP'
2. **Chunk**: w1 and w2 must be in the same 'NP' chunk
3. **Inflection**: Inflection [3] of w1 must be '- শুন্য'*(null)*, '-র'(*-r*), '-এর'(*-er*), '-এ'(*-e*), '-য়'(*-y*) or '-য়ের'(*-yr*) and for w2, any inflection is considered.

4.2 Dictionary Restructuring

To the best of our knowledge, no full-fledged WordNet resource is available for Bengali. Hence, the building of Bengali synsets from a monolingual Bengali dictionary not only aims to identify the meaning of a token, but also sets up the framework towards the development of Bengali WordNet. Each word present in the monolingual dictionary (Samsada Bengali Abhidhana)[4] contains its POS,

[1] http://www.rabindra-rachanabali.nltr.org
[2] http://ltrc.iiit.ac.in/analyzer/bengali
[3] Linguistic study (Chattopadhyay, 1992) reveals that for compound noun MWE, considerable inflections of first noun are only those which are mentioned above.
[4] http://dsal.uchicago.edu/dictionaries/biswas-bangala/

phonetics and synonymous sets. An automatic technique has been devised to identify the synsets of a particular word based on the clues (“,” comma and “;” semi-colon) provided in the dictionary to distinguish words of similar and different sense from the synonymous sets. The symbol tilde (~) indicates that the suffix string followed by the tilde (~) notation makes another new word concatenating with the original entry word. A partial snapshot of the synsets for the Bengali word “অংশু” (Angshu) is shown in Figure 1. In Table 1, the frequencies of different synsets according to their POS are shown.

Dictionary Entry:
অংশু [aṃśu] বি. 1 কিরণ, রশ্মি, প্রভা; ~ ক বি. বস্ত্র , সূক্ষ্ম বস্ত্র ; রেশম পাট ইত্যাদিতে প্রস্তুত বস্ত্র। ~ জাল বি. কিরণরাশি, কিরণমালা।
Synsets:
অংশু কিরণ/রশ্মি/প্রভা_বি.#25_1_1
অংশুক বস্ত্র/সূক্ষ্ম_বস্ত্র_বি.#26_1_1
অংশুক রেশম_পাট_ইত্যাদিতে_প্রস্তুত_বস্ত্র_বি.#26_2_2
অংশুজালকিরণরাশি/কিরণমালা_বি.#27_1_1

Figure 1: A partial snapshot of the Bengali monolingual dictionary entry (word and synsets)

Total #Word	Total #Synset	Noun	Adjective	Pronoun	Verb
33619	63403	28485	11023	235	1709

Table 1: Total number of words, synsets and Frequencies of different POS based synsets

4.3 Generating Semantic Clusters of Nouns

In the first phase, we have generated the synonymous sets for all nouns present in the corpus using the synset based dictionary whereas in the second phase, the task is to identify the semantic distance between two nouns. The format of the dictionary can be thought of as follows:

$W^1=n_1^1, n_2^1, n_3^1, \ldots\ldots\ldots\ldots = \{n_i^1\}$

.

.

$W^m=n_1^m, n_2^m, n_3^m, \ldots\ldots\ldots\ldots = \{n_p^m\}$

where, W^1, W^2, ….,W^m are the dictionary word entries and n_j^m (for all j) are the elements of the synsets of W^m. Now, each noun entry identified by the shallow parser in the document is searched in the dictionary. For example, if a noun N present the corpus becomes an entry of the synsets, W^1, W^3 and W^5, the synset of N is as follows,

$$SynSet\ (N) = \{W^1, W^3, W^5\} \ldots\ldots\ldots (1)$$

To identify the semantic similarity between two nouns, we have applied simple intersection rule. The number of common elements between the synsets of the two noun words denotes the similarity between them. If N_i and N_j are the two noun words in the document and W^i and W^j are their corresponding synsets, the similarity of the two words can be defined as,

$$Similarity\ (N_i,N_j) = |W^i \cap W^j| \ldots\ldots\ldots (2)$$

We have clustered all the nouns present in the document for a particular noun and have identified the similarity score for every pair of nouns obtained using equation 2.

4.4 Checking of Candidate Bigram as MWE

The identification of candidates as MWE is done using the results obtained from the previous phase. The algorithm to identify the noun-noun bigram <M1 M2> as MWE is discussed below with an example shown in Figure 2.

ALOGRITHM: MWE-CHECKING
INPUT: Noun-noun bigram <*M1 M2*>
OUTPUT: Return true if MWE, or return false.
1. Extract semantic clusters of *M1* and *M2*
2. Intersection of the clusters of both *M1* and *M2* (Figure 2.1 shows the common synset entries of *M1* and *M2* using rectangle).
3. For measuring the semantic similarity between *M1* and *M2*:
   3.1. In an *n*-dimensional vector space (here *n*=2), the common entries act as the axes. Put *M1* and *M2* as two vectors and associated weights as their co-ordinates.
   3.2. Calculate cosine-similarity measurement and Euclidean distance (Figure 2.2).
4. Final decision taken individually for two different measurements-
   4.1 If cosine-similarity > *m*, return false; Else return true;
   4.2 If Euclidean-distance > *p*, return false; Else return true;

(Where *m* and *p* are the pre-defined cut-off values)

We have also employed English WordNet[5] to measure the semantic similarity between two

[5] http://www.d.umn.edu/tpederse/similarity.html

Cut-off	Cosine-Similarity			Euclidean Distance			WordNet Similarity		
	P	R	FS	P	R	FS	P	R	FS
0.6	70.75	64.87	67.68	70.57	62.23	66.14	74.60	61.78	67.58
0.5	**78.56**	**59.45**	**67.74**	72.97	58.79	65.12	**80.90**	**58.75**	**68.06**
0.4	73.23	56.97	64.08	**79.78**	**53.03**	**63.71**	75.09	52.27	61.63

Table 3: Precision (P), Recall (R) and F-score (FS) (in %) for various measurements

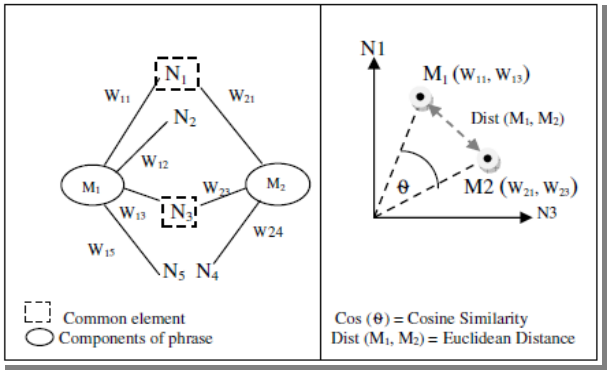

Figure 2.1: Intersection of the clusters of the constituents (left side); Figure 2.2: Similarity between two constituents Evaluation (right side)

Bengali words translated into English. WordNet::Similarity is an open-source package for calculating the lexical similarity between word (or sense) pairs based on various similarity measures. Basically, WordNet measures the relative distance between two nodes denoted by two words in the WordNet tree which can vary from -1 to 1 where -1 indicates total dissimilarity between two nodes. The equation used to calculate this distance is mentioned below-

Normalized_Distance= minDistToCommonParent / (DistFromCommonParentToRoot + minDistToCommonParent) …..…..…..(3)

We have translated the root of the two components of a Bengali candidate into their English equivalents using a Bengali to English bilingual dictionary. They are passed into the WordNet based similarity module for measuring similarity between the components.

If we take an example of a Bengali idiom *hater panch* (*remaining resource*) to describe our intuition, we have seen that the WordNet defines two components of the idiom *hat* (*hand*) as 'a part of a limb that is farthest from the torso' and *panch* (*five*) as 'a number which is one more than four'. So from these two glosses it is quite clear that they are not at all semantically related in any sense.

The synonymous sets for these two components extracted from the formatted dictionary are shown below –

Synset (হাত '*hat*') = { হস্ত, কর, পাণি, বাহু, ভুজ, কৌশল, হস্তক্ষেপ, ধারণ, রেখা, লিখিত, হস্তাক্ষর, হস্তান্তর, হাজা }

Synset (পাঁচ '*panch*') = {পঞ্চ, সংখ্যা, কর্ম, গঙ্গা, গব্য, কন্যা, গুণ, গৌড়, তন্ত্র, তীর্থ, পঞ্চত্ব, পনেরো, পূর্ণিমা, পঞ্চাশ }

It is clearly seen from the above synonymous sets that there is no common element and hence its similarity score is obviously zero. In this case, the vector space model cannot be drawn using zero dimensions. For them, a marginal weight is assigned to show them as completely non-compositional phrase. To identify their non-compositionality, we have to show that their occurrence is not certain only in one case; rather they can occur side by side in several occasions. But this statistical proof can be determined better using a large corpus. Here, for those candidate phrases, which show zero similarity, we have seen their existence more than one time in the corpus. Taking any decision using single occurrence may give incorrect result because they can be unconsciously used by the authors in their writings. That is why, the more the similarity between two components in a bigram, the less the probability to be a MWE.

4.5 Annotation Agreement

Three annotators identified as A1, A2 and A3 were engaged to carry out the annotation. The annotation agreement of 628 candidate phrases is measured using standard Cohen's *kappa* coefficient (κ) (Cohen, 1960). It is a statistical measure of inter-rater agreement for qualitative (categorical) items. In addition to this, we also choose the measure of agreements on set-valued items (*MASI*) (Passonneau, 2006) that was used for measuring agreement in the semantic and pragmatic annotation. Annotation results as shown in Table 2 are satisfactory.

The list of noun-noun collocations are extracted from the output of the parser for manual checking. It is observed that 39.39% error occurs due to wrong POS tagging or extracting invalid collocations by considering the bigrams in a n-gram chunk where n > 2. We have separated these phrases from the final list.

MWEs [# 628]	Agreement between pair of annotators A1-A2	A2-A3	A1-A3	Avg
KAPPA	87.23	86.14	88.78	87.38
MASI	87.17	87.02	89.02	87.73

Table 2: Inter-Annotator Agreement (in %)

4.6 Experimental Results

We have used the standard IR matrices like Precision (P), Recall (R) and F-score (F) for evaluating the final results obtained from three modules. Human annotated list is used as the gold standard for the evaluation. The present system results are shown in Table 3. These results are compared with the statistical baseline system described in (Chakraborty, 2010). Our baseline system is reported with the precision of 39.64%. The predefined threshold has been varied to catch individual results in each case. Increasing Recall in accordance with the increment of cut-off infers that the maximum numbers of MWEs are identified in a wide range of threshold. But the Precision does not increase considerably. It shows that the higher cut-off degrades the performance. The reasonable results for Precision and Recall have been achieved in case of cosine-similarity at the cut-off value of 0.5 where Euclidean distance and WordNet Similarity give maximum precision at cut-off values of 0.4 and 0.5 respectively. In all cases, our system outperforms the baseline system.

It is interesting to observe that English WordNet becomes a very helpful tool to identify Bengali MWEs. WordNet detects maximum MWEs correctly at the cut-off of 0.5. Baldwin *et al.*, (2003) suggested that WordNet::Similarity measure is effective to identify empirical model of Multiword Expression Decomposability. This is also proved in this experiment as well and even for Bengali language. There are also candidates with very low value of similarity between their constituents (for example, *ganer gajat* (*earth of song, affectionate of song*), yet they are discarded from this experiment because of their low frequency of occurrence in the corpus which could not give any judgment regarding collocation. Whether such an unexpectedly low frequent high decomposable elements warrant an entry in the lexicon depends on the type of the lexicon being built.

5 Conclusions

We hypothesized that sense induction by analyzing synonymous sets can assist the identification of Multiword Expression. We have introduced an unsupervised approach to explore the hypothesis and have shown that clustering technique along with similarity measures can be successfully employed to perform the task. This experiment additionally contributes to the following scenarios - (i) Clustering of words having similar sense, (ii) Identification of MWEs for resource constraint languages and (iii) Reconstruction of Bengali monolingual dictionary towards the development of Bengali WordNet. However, in our future work, we will apply the present techniques for other type of MWEs (e.g., adjective-noun collocation, verbal MWEs) as well as for other languages.

Acknowledgement

The work reported in this paper is supported by a grant from the "Indian Language to Indian Language Machine Translation (IL-ILMT) System Phrase II", funded by Department of Information and Technology (DIT), Govt. of India.

References

Agarwal, Aswini, Biswajit Ray, Monojit Choudhury, Sudeshna Sarkar and Anupam Basu. 2004. Automatic Extraction of Multiword Expressions in Bengali: An Approach for Miserly Resource Scenario. In *Proceedings of International Conference on Natural Language Processing (ICON)*, pp. 165-174.

Baldwin, Timothy, Colin Bannard, Takaaki Tanaka and Dominic Widdows. 2003. An Empirical Model of Multiword Expression Decomposability. *Proceedings of the Association for Computational Linguistics-2003, Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, Sapporo, Japan, pp. 89–96.

Ckakraborty, Tanmoy, 2010, Identification of Noun-Noun (N-N) Collocations as Multi-Word Expressions in Bengali Corpus. *Student Session, International Conference of Natural Language Processing (ICON)*, IIT Kharagpur, India

Chakraborty, Tanmoy and Sivaji Bandyopadhyay. 2010. Identification of Reduplication in Bengali Corpus and their Semantic Analysis: A Rule Based Approach. In *proceedings of the Workshop on Multiword Expressions: from Theory to Applications (MWE 2010), 23rd International Conference on Computational Linguistics (COLING 2010)*, pp.73-76, Beijing, China.

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

Church, Kenneth Wrad and Patrick Hans. 1990. Word Association Norms, Mutual Information and Lexicography. *Proceedings of 27th Association for Computational Linguistics (ACL)*, 16(1). pp. 22-29.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, vol. 20, pp. 37–46.

Dunning, T. 1993. Accurate Method for the Statistic of Surprise and Coincidence. In *Computational Linguistics*, pp. 61-74.

Kilgarriff, Adam and Joseph Rosenzweig. 2000. Framework and results for English SENSEVAL. *Computers and the Humanities*. Senseval Special Issue, 34(1-2). pp. 15-48.

Korkontzelos,Ioannis and Suresh Manandhar. 2009. Detecting Compositionality in Multi-Word Expressions. *Proceedings of the Association for Computational Linguistics-IJCNLP*, Singapore, pp. 65-68.

Kunchukuttan F. A. and Om P. Damani. 2008. A System for Compound Noun Multiword Expression Extraction for Hindi. *Proceeding of 6th International Conference on Natural Language Processing (ICON)*. pp. 20-29.

Passonneau, R.J. 2006. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. *Language Resources and Evaluation.*

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. In *Proceedings of Conference on Intelligent Text Processing and Computational Linguistics (CICLING)*, pp. 1-15.

Venkatapathy, Sriram and Aravind Joshi. 2005. Measuring the relative compositionality of verb-noun (V-N) collocations by integrating features. *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), Association for Computational Linguistics*. pp. 899 - 906.