Automatic Identification of Bengali Noun-Noun Compounds Using Random Forest

Vivekananda Gayen
Department of Computer Science and Technology
Central Calcutta Polytechnic
Kolkata-700014, India
vivek3gayen@gmail.com

Kamal Sarkar
Department of Computer Science and Engineering
Jadavpur University
Kolkata, India
jukamal2001@yahoo.com

Abstract

This paper presents a supervised machine learning approach that uses a machine learning algorithm called Random Forest for recognition of Bengali noun-noun compounds as multiword expression (MWE) from Bengali corpus. Our proposed approach to MWE recognition has two steps: (1) extraction of candidate multi-word expressions using Chunk information and various heuristic rules and (2) training the machine learning algorithm to recognize a candidate multi-word expression as Multi-word expression or not. A variety of association measures, syntactic and linguistic clues are used as features for identifying MWEs. The proposed system is tested on a Bengali corpus for identifying noun-noun compound MWEs from the corpus.

1 Introduction

Automatic identification of multiword expression (MWE) from a text document can be useful for many NLP (natural language processing) applications such as information retrieval, machine translation, word sense disambiguation. According to Frank Samadja (1993), MWEs are defined as "recurrent combinations of words that co-occur more often than expected by chance". Timothy Baldwin et al. (2010) defined multiword expressions (MWEs) as lexical items that: (a) can be decomposed into multiple lexemes; and (b) display lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity. Most real world NLP applications tend to ignore MWE, or handle them simply by listing, but successful applications will need to identify and treat them appropriately.

As Jackendoff (1997) stated, the magnitude of this problem is far greater than has traditionally been realized within linguistics. He estimates that the number of MWEs in a native speakers's lexicon is of the same order of magnitude as the number of single words. In WordNet 1.7 (Fellbaum, 1999), for example, 41% of the entries are multiword.

MWEs can be broadly classified into lexicalized phrases and institutionalized phrases (Ivan A. sag et al., 2002). In terms of the semantics, compositionality is an important property of MWEs. Compositionality is the degree to which the features of the parts of a MWE combine to predict the features of the whole. According to the compositionality property, the MWEs can take a variety of forms: complete compositionality (also known as institutionalized phrases, e.g. many thanks, 'রাজ্য সরকার' (Rajya Sarkar, state government)), partial compositionality (e.g. light house, 'শপিং মল' (shopping mall), 'আম আদমি' (aam admi, common people)), idiosyncratically compositionality (e.g. spill the beans (to reveal)) and finally complete non-compositionality (e.g. hot dog, green card, 'উভয় সঙ্কট' (ubhoy sangkat, on the horns of a dilemma)).

Compound noun is a lexical unit. It is a class of MWE which is rapidly expanding due to the continuous addition of new terms for introducing new ideas. Compound nouns fall into both groups: lexicalized and institutionalized. A noun-noun compound in English characteristically occurs frequently with high lexical and semantic variability. A summary examination of the 90 million-

Proceedings of the 9th Workshop on Multiword Expressions (MWE 2013), pages 64–72,
Atlanta, Georgia, 13-14 June 2013.

word written component of the British National Corpus (BNC) uncover the fact that there are over 400,000 NN (Noun-Noun) compound types, with a combined token frequency of 1.3 million, that is, over 1% of words in the BNC are NN compounds (Timothy Baldwin et al., 2003). Since compound nouns are rather productive and new compound nouns are created from day to day, it is impossible to exhaustively store all compound nouns in a dictionary

It is also common practice in Bengali literature to use compound nouns as MWEs. Bengali new terms directly coined from English terms are also commonly used as MWEs in Bengali (e.g., 'ডেং থ্রি' (dengue three), 'ন্যানো সিম' (nano sim), 'ভিলেজ টুযরিজম' (village tourism), 'অ্যালার্ট মেসেজ' (alert message)).

The main focus of our work is to develop a machine learning approach based on a set of statistical, syntactic and linguistic features for identifying Bengali noun-noun compounds.

To date, not much comprehensive work has been done on Bengali multiword expression identification.

Different types of compound nouns in Bengali are discussed in section 2. Related works are presented in section 3. The proposed noun-noun MWE identification method has been detailed in section 4. The evaluation and results are presented in section 5 and conclusions and feature work are drawn in section 6.

2 Classification of Bengali Compound Nouns

In Bengali, MWEs are quite varied and many of these are of types that are not encountered in English. The primary types of compound nouns in Bengali are discussed below.

Named-Entities (NE): Names of people ('তীর্থ দাস' (Tirtha Das), 'নয়ন রায়' (Nayan Roy)). Name of the location ('হুগলি স্টেশন' (Hooghly Station), 'অশোক বিহার' (Ashok Bihar)). Names of the Organization ('আইডিয়াল কেব্ল অপারেটর্স অ্যাসোসিয়েশন' (Ideal cable operators association), 'রিবক ইন্ডিয়া' (Reebok India)). Here inflection can be added to the last word.

Idiomatic Compound Nouns: These are characteristically idiomatic and unproductive. For example, 'মা বাবা' (maa baba, father mother), 'কল কারখানা' (kaal karkhana, mills and workshops) are MWEs of this kind.

Idioms: These are the expressions whose meanings can not be recovered from their component words. For example, 'তাসের ঘর' (taser ghar, any construction that may tumble down easily at any time), 'পাখির চোখ' (pakhir chokh, target), 'সবুজ বিপ্লব' (sabuj biplab, green revolution) are the idioms in Bengali.

Numbers: These are productive in nature and little inflection like syntactic variation is also seen in number expression. For example, 'সোয়া তিন ঘন্টা' (soya teen ghanta, three hours and fifteen minutes), 'আড়াই গুণ' (arawi guun, two and a half times), 'সাড়ে তিনটে' (sharre teenta, three hours and thirty minutes), 'দেড় বছর' (der bachar, one and a half year) are MWEs of this kind.

Relational Noun Compounds: These are generally consists of two words, no word can be inserted in between. Some examples are: 'পিচতুতো ভাই' (pistuto bhai, cousin), 'মেজ মেয়ে' (majo meyya, second daughter).

Conventionalized Phrases (or Institutionalized phrases):

Institutionalized phrases are conventionalized phrases, such as ('বিবাহ বার্ষীকি' (bibaha barshiki, marriage anniversary, 'চাক্কা জ্যাম' (chakka jam, standstill), 'শেয়ার বাজার' (share bazar, share market)). They are semantically and syntactically compositional, but statistically idiosyncratic.

Simile terms: It is analogy term in Bengali and semi-productive ('হাতের পাঁচ' (hater panch, last resort), 'কথার কথা' (kather katha, a word for word's sake)).

Reduplicated terms: Reduplicated terms are non-productive and tagged as noun phrase. Namely Onomatopoeic expression ('খট খট' (khhat khhat, knock knock), 'হু হু' (hu hu, the noise made by a strong wind)), Complete reduplication ('বাড়ি বাড়ি' (bari bari, door to door), 'ব্লকে ব্লকে' (blocke blocke, block block)), Partial reduplication ('যন্তর মন্তর' (jantar mantar)), Semantic reduplication ('মাথা মুন্ডু' (matha mundu, head or tail)), Correlative reduplication ('মারামারি' (maramari, fighting)).

Administrative terms: These are institutionalized as administrative terms and are non-productive in nature. Here inflection can be added with the last word ('স্বরাষ্ট্র মন্ত্রক' (sarastra montrak, home ministry)), 'স্বাস্থ্য সচিব' (sastha sachib, health secretary)).

One of the component of MWE from English literature: Some examples of Bengali MWEs of this kind are 'মাদ্রাসা বোর্ড' (madrasha board), 'মেট্রো শহর' (metro sahar, metro city).

Both of the component of MWE from English literature: Some examples of Bengali MWEs of this kind are 'রোমিং চার্জ' (roaming charge), 'ক্রেডিট কার্ড' (credit card).

3 Related Work

The earliest works on Multiword expression extraction can be classified as: Association measure based methods, deep linguistic based methods, machine learning based methods and hybrid methods.

Many previous works have used statistical measures for multiword expression extraction. One of the important advantages of using statistical measures for extracting multiword expression is that these measures are language independent. Frank Smadja (1993) developed a system, Xtract that uses positional distribution and part-of-speech information of surrounding words of a word in a sentence to identify interesting word pairs. Classical statistical hypothesis test like Chi-square test, t-test, z-test, log-likelihood ratio (Ted Dunning, 1993) have also been employed to extract collocations. Gerlof Bouma (2009) has presented a method for collocation extraction that uses some information theory based association measures such as mutual information and pointwise mutual information.

Wen Zhang et al (2009) highlights the deficiencies of mutual information and suggested an enhanced mutual information based association measures to overcome the deficiencies. The major deficiencies of the classical mutual information, as they mention, are its poor capacity to measure association of words with unsymmetrical co-occurrence and adjustment of threshold value. Anoop et al (2008) also used various statistical measures such as point-wise mutual information (K. Church et al., 1990), log-likelihood, frequency of occurrence, closed form (e.g., blackboard) count, hyphenated count (e.g., black-board) for extraction of Hindi compound noun multiword extraction. Aswhini et al (2004) has used co-occurrence and significance function to extract MWE automatically in Bengali, focusing mainly on Noun-verb MWE. Sandipan et al (2006) has used association measures namely salience (Adam Kilgarrif et al., 2000), mutual information and log likelihood for finding N-V collocation. Tanmoy (2010) has used a linear combination of some of the association measures namely co-occurrence, Phi, significance function to obtain a linear ranking function for ranking Bengali noun-noun collocation candidates and MWEness is measured by the rank score assigned by the ranking function.

The statistical tool (e.g., log likelihood ratio) may miss many commonly used MWEs that occur in low frequencies. To overcome this problem, some linguistic clues are also useful for multiword expression extraction. Scott Songlin Paul et al (2005) focuses on a symbolic approach to multiword extraction that uses large-scale semantically classified multiword expression template database and semantic field information assigned to MWEs by the USAS semantic tagger (Paul Rayson et al.,2004). R. Mahesh et al (2011) has used a stepwise methodology that exploits linguistic knowledge such as replicating words (ruk ruk e.g. stop stop), pair of words (din-raat e.g. day night), samaas (N+N, A+N) and Sandhi (joining or fusion of words), Vaalaa morpheme (jaane vaalaa e.g. about to go) constructs for mining Hindi MWEs. A Rule-Based approach for identifying only reduplication from Bengali corpus has been presented in Tanmoy et al (2010). A semantic clustering based approach for indentifying bigram noun-noun MWEs from a medium-size Bengali corpus has been presented in Tanmoy et al (2011). The authors of this paper hypothesize that the more the similarity between two components in a bigram, the less the probability to be a MWE. The similarity between two components is measured based on the synonymous sets of the component words.

Pavel Pecina (2008) used linear logistic regression, linear discriminant analysis (LDA) and Neural Networks separately on feature vector consisting of 55 association measures for extracting MWEs. M.C. Diaz-Galiano et al. (2004) has applied Kohonen's linear vector quantization (LVQ) to integrate several statistical estimators in order to recognize MWEs. Sriram Venkatapathy et al. (2005) has presented an approach to measure relative compositionality of Hindi noun-verb MWEs using Maximum entropy model (MaxEnt). Kishorjit et al (2011) has presented a conditional random field (CRF) based method for extraction and transliteration of Manipuri MWEs.

Hybrid methods combine statistical, linguistic and/or machine learning methods. Maynard and Ananiadou (2000) combined both linguistics and statistical information in their system, TRUCK, for extracting multi-word terms. Dias (2003) has developed a hybrid system for MWE extraction, which integrates word statistics and linguistic information. Carlos Ramisch et al. (2010) presents a hybrid approach to multiword expression extraction that combines the strengths of different sources of information using a machine learning algorithm. Ivan A. Sag et al (2002) argued in favor of maintaining the right balance between symbolic and statistical approaches while developing a hybrid MWE extraction system.

4 Proposed Noun-Noun compound Identification Method

Our proposed noun-noun MWE identification method has several steps: preprocessing, candidate noun-noun MWE extraction and MWE identification by classifying the candidates MWEs into two categories: positive (MWE) and negative (non-MWE).

4.1 Preprocessing

At this step, unformatted documents are segmented into a collection of sentences automatically according to Dari (in English, full stop), Question mark (?) and Exclamation sign (!). Typographic or phonetic errors are not corrected automatically. Then the sentences are submitted to the chunker [1] one by one for processing. The chunked output is then processed to delete the information which is not required for MWE identification task. A Sample input sentence and the corresponding chunked sentence after processing are shown in figure 1.

Sample input sentence:

পরিবহণ একটি অত্যাবশ্যক শিল্প ।(paribhan ekti attyaboshak shilpo, Communication is a essential industry.)

Processed output from the chunker:

((NP পরিবহণ NN)) ((NP একটি QC অত্যাবশ্যক JJ শিল্প NN SYM))

Figure 1: A Sample input sentence and processed output from the chunker.

[1] http//ltrc.iiit.ac.in/analyzer/bengali

4.2 Candidate Noun-Noun MWE Extraction

The chunked sentences are processed to identify the noun-noun multi-word expression candidates. The multiword expression candidates are primarily extracted using the following rule:

Bigram consecutive noun-noun token sequence within same NP chunk is extracted from the chunked sentences if the Tag of the token is NN or NNP or XC (NN: Noun, NNP: Proper Noun, XC: compounds) (Akshar Bharati et al., 2006).

We observed that some potential noun-noun multi-word expressions are missed due to the chunker's error. For example, the chunked version of the sentence is ((NP কবেকার NN)) ((NP বিএসএ NN)) ((NP সাইকেল NN, SYM)). Here we find that the potential noun-noun multi-word expression candidate "বিএসএ সাইকেল" (BSA Cycle) cannot be detected using the first rule since "বিএসএ" (BSA) and সাইকেল (Cycle) belong to the different chunk.

To identify more number of potential noun-noun MWE candidates, we use some heuristic rules as follows:

Bigram noun-noun compounds which are hyphenated or occur within single quote or within first brackets or whose words are out of vocabulary (OOV) are also considered as the potential candidates for MWE.

4.3 Features

4.3.1 Statistical features: We use the association measures namely phi, point-wise mutual information (pmi), salience, log likelihood, poisson stirling, chi and t-score to calculate the scores of each noun-noun candidate MWE. These association measures use various types of frequency statistics associated with the bigram. Since Bengali is highly inflectional language, the candidate noun-noun compounds are stemmed while computing their frequencies.

The frequency statistics used in computing association measures are represented using a typical contingency table format (Satanjeev Banerjee et al., 2003). Table 1 shows a typical contingency table showing various types of frequencies associated with the noun-noun bigram <word, word2> (e.g., রাজ্য সরকার). The meanings of the entries in the contingency table are given below:

n_{11} = number of times the bigram occurs, joint frequency.

n_{12} = number of times word1 occurs in the first position of a bigram when word2 does not occur in the second position.

	সরকার (government)	~সরকার (~ government)	
রাজ্য (state)	n_{11}	n_{12}	n1p
~রাজ্য (~state)	n_{21}	n_{22}	n2p
	np1	np2	npp

Table 1: Contingency table

n_{21} = number of times word2 occurs in the second position of a bigram when word1 does not occur in the first position.

n_{22} = number of bigrams where word1 is not in the first position and word2 is not in the second position.

n1p = the number of bigrams where the first word is word, that is, n1p =n_{11}+ n_{12}.

np1 = the number of bigrams where the second word is word2, that is np1=n_{11}+n_{21}.

n2p = the number of bigrams where the first word is not word1, that is n2p=n_{21}+n_{22}.

np2 = the number of bigrams where the second word is not word2, that is np2=n_{12}+n_{22}.

npp is the total number of bigram in the entire corpus.

Using the frequency statistics given in the contingency table, expected frequencies, $m_{11,}$ m_{12}, m_{21} and m_{22} are calculated as follows:

m_{11} = (n1p*np1/npp)
m_{12} = (n1p*np2/npp)
m_{21} = (np1*n2p/npp)
m_{22} = (n2p*np2/npp)

where:

m_{11}: Expected number of times both words in the bigram occur together if they are independent.

m_{12}: Expected number of times word1 in the bigram will occur in the first position when word2 does not occur in the second position given that the words are independent.

m_{21}: Expected number of times word2 in the bigram will occur in the second position when word1 does not occur in the first position given that the words are independent.

m_{22}: Expected number of times word1 will not occur in the first position and word2 will not occur in the second position given that the words are independent.

The following association measures that use the above mentioned frequency statistics are used in our experiment.

Phi, Chi and T-score: The Phi, Chi and T-score are calculated using the following equations:

$$phi = \frac{((n_{11}*n_{22})-(n_{12}*n_{21}))}{\sqrt{(n1p*np1*np2*n2p)}}$$

$$chi = 2*\left(\left(\frac{(n_{11}-m_{11})}{m_{11}}\right)^2+\left(\frac{(n_{12}-m_{12})}{m_{12}}\right)^2+\left(\frac{(n_{21}-m_{21})}{m_{21}}\right)^2+\left(\frac{(n_{22}-m_{22})}{m_{22}}\right)^2\right)$$

$$T-Score = \frac{(n_{11}-m_{11})}{\sqrt{n_{11}}}$$

Log likelihood, Pmi, Salience and Poisson Stirling: Log likelihood is calculated as:

$$LL = 2*(n_{11}*\log(n_{11}*m_{11})+n_{12}*\log(n_{12}*m_{12})+n_{21}*\log(n_{21}*m_{21})+n_{22}*\log(n_{22}*m_{22}))$$

Pointwise Mutual Information (pmi) is calculated as:

$$pmi = \log\left(\frac{n_{11}}{m_{11}}\right)$$

The salience is defined as:

$$salience = (\log(\frac{n_{11}}{m_{11}}))*\log(n_{11})$$

The Poisson Stirling measure is calculated using the formula:

$$Poisson-Stirling = n_{11}*((\log(\frac{n_{11}}{m_{11}})-1)$$

Co-occurrence: Co-occurrence is calculated using the following formula (Agarwal et al., 2004):

$$co(w1,w2) = \sum_{s\in S(w1,w2)} e^{-d(s,w1,w2)}$$

Where co(w1,w2)=co-occurrence between the words (after stemming).

S(w1,w2)= set of all sentences where both w1 and w2 occurs.

d(s,w1,w2)= distance between w1 and w2 in a sentence in terms of words.

Significance Function: The significance function (Aswhini Agarwal et al., 2004) is defined as:

$$sig_{w1}(w2) = \sigma[k1(1-co(w1,w2).\frac{f_{w1}(w2)}{f(w1)})].\sigma[k2.\frac{f_{w1}(w2)}{\lambda}-1]$$

$$sig(w1,w2) = sig_{w1}(w2).\exp[\frac{f_{w1}(w2)}{\max(f_{w1}(w2))}-1]$$

Where:

$sig_{w1}^{(w2)}$ = significance of w2 with respect to w1.

$f_{w1}^{(w2)}$ = number of w1 with which w2 has occurred.

Sig(w1,w2)= general significance of w1 and w2, lies between 0 and 1.

σ(x)= sigmoid function =exp(-x)/(1+exp(-x))]

k1 and k2 define the stiffness of the sigmoid curve (for simplicity they are set to 5.0)
λ is defined as the average number of noun-noun co-occurrences.

4.3.2 Syntactic and linguistic features: Other than the statistical features discussed in the above section, we also use some syntactic and linguistic features which are listed in the table 2.

Feature name	feature description	Feature value
AvgWordLength	average length of the components of a candidate MWE	Average length of the words in a candidate MWE
Whether-Hyphenated	Whether a candidate MWE is hyphenated	Binary
Whether-Within-Quote	Whether a candidate MWE is within single quote	Binary
Whether-Within-Bracket	Whether a candidate MWE is within first brackets	Binary
OOV	Whether candidate MWE is out of vocabulary	Binary
First-Word-Inflection	Whether the first word is inflected	Binary
Second-Word-Inflection	Whether second word is inflected	Binary
TagOf-FirstWord	Lexical category of the first word of a candidate.	XC (compound), NN (noun), NNP (proper noun)
TagOfSecondWord	Lexical category of the second word of a candidate	XC (compound), NN (noun), NNP (proper noun)

Table2. Syntactic and linguistic features

4.4 Noun-noun MWE identification using random forest

Random forest (Leo Breiman, 2000) is an ensemble classifier that combines the predictions of many decision trees using majority voting to output the class for an input vector. Each decision tree participated in ensembling chooses a subset of features randomly to find the best split at each node of the decision tree. The method combines the idea of "bagging" (Leo Breiman, 1996) and the random selection of features. We use this algorithm for our multiword identification task for several reasons: (1) For many data sets, it produces a highly accurate classifier (Rich Caruana et al, 2008), (2) It runs efficiently on large databases and performs well consistently across all dimensions and (3) It generates an internal unbiased estimate of the generalization error as the forest building progresses.

The outline of the algorithm is given in the figure 2.

Training Random Forests for noun-noun MWE identification requires candidate noun-noun MWEs to be represented as the feature vectors. For this purpose, we write a computer program for automatically extracting values for the features characterizing the noun-noun MWE candidates in the documents. For each noun-noun candidate MWE in a document in our corpus, we extract the values of the features of the candidate using the measures discussed in subsection 4.3. If the noun-noun candidate MWE is found in the list of manually identified noun-noun MWEs, we label the MWE as a "Positive" example and if it is not found we label it as a "negative" example. Thus the feature vector for each candidate looks like {<a_1 a_2 a_3 ….. a_n>, <label>} which becomes a training instance (example) for the random forest, where a_1, a_2 . . .a_n, indicate feature values for a candidate. A training set consisting of a set of instances of the above form is built up by running a computer program on the documents in our corpus.

For our experiment, we use Weka (www.cs.waikato.ac.nz/ml/weka) machine learning tools. The *random forest* is included under the panel Classifier/ trees of WEKA workbench.. For our work, the random forest classifier of the WEKA suite has been run with the default values of its parameters. One of the important parameters

is number of trees in the forest. We set this parameter to its default value of 10.

Random forest learning algorithm

Training phrase:
For each of N decision trees to be built

- Select a new bootstrap sample from training set
- Grow an un-pruned decision tree on this bootstrap.
- While growing a decision tree, at each internal node, randomly select m_{try} predictors (features) and determine the best split using only these predictors.
- Do not perform pruning. Save the decision tree.

Testing phase:
For an input vector, output the class that is the mode of the classes produced by the all individually trained decision trees.

Figure 2. Random forest learning algorithm

5 Evaluation and results

For evaluating the performance of our system the traditional precision, recall and F-measure are computed by comparing machine assigned labels to the human assigned labels for the noun-noun candidate MWEs extracted from our corpus of 274 Bengali documents.

5.1 Experimental dataset

Our corpus is created by collecting the news articles from the online version of well known Bengali newspaper ANANDABAZAR PATRIKA during the period spanning from 20.09.2012 to 19.10.2012. The news articles published online under the section Rajya and Desh on the topics bandh-dharmoghat, crime, disaster, jongi, mishap, political and miscellaneous are included in the corpus. It consists of total 274 documents and all those documents contain 18769 lines of Unicode texts and 233430 tokens. We have manually identified all the noun-noun compound MWEs in the collection and labeled the training data by assigning positive labels to the noun-noun compounds and negative labels to the expressions which are not noun-noun compounds. It consists of 4641 noun-noun compound MWEs. Total 8210 noun-noun compound MWE candidates are automatically extracted employing chunker and using heuristic rules as described in subsection 4.2.

5.2 Results

To estimate overall accuracy of our proposed noun-noun MWE identification system, 10-fold cross validation is done. The dataset is randomly reordered and then split into *n* parts of equal size. For each of 10 iterations, one part is used for testing and the other *n*-1 parts are used for training the classifier. The test results are collected and averaged over all folds. This gives the cross-validation estimate of the accuracy of the proposed system. J48 which is basically a decision tree included in WEKA is used as a single decision tree for comparing our system. The table 2 shows the estimated accuracy of our system. The comparison of the performance of the proposed random forest based system to that of a single decision tree is also shown in table 2. Our proposed random forest based system gives average F-measure of 0.852 which is higher than F-measure obtained by a single decision tree for bigram noun-noun compound recognition task.

Systems	Precision	Recall	F-measure
Random Forest	0.852	0.852	0.852
Single Decision Tree	0.831	0.83	0.831

Table 2: Comparisons of the performances of the proposed *random forest* based system and a single decision tree based system for bigram noun-noun compound recognition task.

6 Conclusion and Future Work

This paper presents a machine learning based approach for identifying noun-noun compound MWEs from a Bengali corpus. We have used a number of association measures, syntactic and linguistic information as features which are combined

by a random forest learning algorithm for recognizing noun-noun compounds.

As a future work, we have planned to improve the noun-noun candidate MWE extraction step of the proposed system and/or introduce new features such as lexical features and semantic features for improving the system performance.

References

Adam Kilgarrif and Joseph Rosenzweig. 2000. Framework and Results for English Senseval. *Computer and the Humanities*, 34(1): pp 15-48.

Akshar Bharati, Dipti Misra Sharma, Lakshmi Bai, Rajeev Sangal. 2006. AnnCorra : Annotating Corpora Guidelines For POS And Chunk Annotation For Indian Languages.

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

Aswhini Agarwal, 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

Carlos Ramisch, Helena de Medeiros Caseli, Aline Villavicencio, André Machado, Maria José Finatto: *A Hybrid Approach for Multiword Expression Identification.* PROPOR 2010: 65-74

Fellbaum, Christine, ed.: 1998, WordNet: An Electronic Lexical Database, *Cambridge,MA: MIT Press*.

Frank Smadja 1993. "Retrieving Collocation from Text: Xtract." *Computational Linguistics*. 19.1(1993):143-177.

Gerlof Bouma. 2009. "Normalized (pointwise) mutual information in collocation extraction." *Proceedings of GSCL* (2009): 31-40.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake and Dan Flickinger.2002. Multi-worh expression: A Pain in the neck for NLP. *CICLing*, 2002.

Jackendoff, Ray: 1997, The Architecture of the Language Faculty, *Cambridge, MA: MIT Press*.

Kishorjit Nongmeikapam, Ningombam Herojit Singh, Bishworjit Salam and Sivaji Bandyopadhyay. 2011. Transliteration of CRF Based Multiword Expression (MWE) in Manipuri: From Bengali Script Manipuri to Meitei Mayek (Script) Manipuri. *International Journal of Computer Science and Information Technology*, vol.2(4) . pp. 1441-1447

K. Church and P. Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*. 16(1). 1990.

Leo Breiman . 1996. "Bagging predictors". *Machine Learning* 24 (2): 123–140.

Leo Breiman . 2001. "Random Forests". *Machine Learning* 45 (1): 5–32.

M.C. Diaz-Galiano, M.T. Martin-Valdivia, F. Martinez-Santiago, L.A. Urea-Lopez. 2004. Multiword Expressions Recognition with the LVQ Algorithm. *Workshop on methodologies and evaluation of Multiword Units in Real-word Applications associated with the 4th International Conference on Languages Resources and Evaluation*, Lisbon, Portugal. pp.12-17

Paul Rayson, Dawn Archer, Scott Piao and Tony McEnery. 2004. The UCREL semantic analysis system. *In Proceedings of the LREC-04 Workshop, beyond Named Entity Recognition Semantic Labelling for NLP Tasks*, Lisbon, Portugal, pp.7-12.

Pavel Pecina. 2008. Reference data for czech collocation extraction. *In Proc. Of the LREC Workshop Towards a Shared Task for MWEs (MWE 2008*). pp. 11-14, Marrakech, Morocco, Jun.

Rich Caruana, Nikos Karampatziakis and Ainur Yessenalina (2008). "An empirical evaluation of supervised learning in high dimensions". *Proceedings of the 25th International Conference on Machine Learning (ICML).*

R. Mahesh and K. Sinha. 2011. Stepwise Mining of Multi-Word Expressions in Hindi. *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE 2011)* pp. 110-115

Sandipan Dandapat, Pabitra Mitra and Sudeshna Sarkar. 2006. Statistical Investigation of Bengali Noun-Verb (N-V) Collocations as Multi-word expressions. *In the Proceedings of MSPIL*, Mumbai, pp 230-233.

Santanjeev Banerjee and Ted Pedersen. 2003. "The Design, Implementation and Use of the Ngram Staistics Package." *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics.* Pp. 370-381

Scott Songlin Piao, Paul Rayson, Dawn Archer, Tony McEnery. 2005. Comparing and combining a semantic tagger and a statistical tool for MWE extraction. *Computer Speech and Language (ELSEVIER) 19 (2005)* pp. 378-397

Sriram Venkatapathy, Preeti Agrawal and Aravind K. Joshi. Relative Compositionality of Noun+Verb Multi-word Expressions in Hindi. *In Proceedings of ICON-2005*, Kanpur.

Takaaki Tanaka, Timothy Baldwin. 2003. "Noun-Noun Compound Machine Translation: a Feasibility Study on Shallow Processing." *Proceeings of the ACL 2003 workshop on Multiword expressions.* pp. 17-24

Tanmoy Chakraborty. 2010. Identification of Noun-Noun(N-N) Collocations as Multi-Word Expressions in Bengali Corpus. *8th International Conference on Natural Language Processing (ICON 2010)*.

Tanmoy Chakraborty and Sivaji Bandyopadhyay. 2010. Identification of Reduplication in Bengali Corpus and their Semantic Analysis: A Rule-Based Approach. *Proceedings of Workshop on Multiword Expressions: from Theory to Applications* (MWE 2010) pp. 72-75

Tanmoy Chakraborty, Dipankar Das and Sivaji Bandyopadhyay. 2011. Semantic Clustering: an Attempt to Identify Multiword Expressions in Bengali. Proceedings of Workshop on Multiword Expressions: from Parsing and Generation to the Real World(MWE 2011*)*. *Association for Computational Linguistics*. Portland, Oregon, USA, 23 June 2011.

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

Timothy Baldwin and Su Nam Kim (2010), in Nitin Indurkhya and Fred J. Damerau (eds .) Handbook of Natural Language Processing, *Second Edition, CRC Press,* Boca Raton, USA, pp. 267-292.