A System for Compound Noun Multiword Expression Extraction for Hindi

Anoop Kunchukuttan and Om P. Damani
Department of Computer Science and Engineering
Indian Institute of Technology Bombay, India
anoopk@cse.iitb.ac.in, damani@cse.iitb.ac.in

Abstract

Identifying compound noun multiword expressions is important for applications like machine translation and information retrieval. We describe a system for extracting Hindi compound noun multiword expressions (MWE) from a given corpus. We identify major categories of compound noun MWEs, based on linguistic and psycholinguistic principles. Our extraction methods use various statistical co-occurrence measures to exploit the statistical idiosyncrasy of MWEs. We make use of various lexical cues from the corpus to enhance our methods. We also address the extraction of reduplicative expressions using lexical, semantic and phonetic knowledge. We have also built an evaluation resource of compound noun MWEs for Hindi. Our methods give a recall of 80% and precision of 23% at rank 1000.

1. Introduction

Multiword expressions (MWE) can be understood as concepts which cross word boundaries or alternatively, are "words with spaces". For instance, the collocation *wheel chair or pick pocket* denotes a single concept. The interpretation of the word sequence is done as a whole. A grammatical analysis is not done while interpreting multiword expressions, but the entire expression is treated as a single unit. Thus, an MWE can be considered to be *a sequence, continuous or discontinuous, of words or other elements, which is or appears to be prefabricated: that is stored and retrieved whole from memory at the time from use, rather than being subject to generation or analysis by language grammar* [1]. Psycholinguistic and phonological studies [2] point to the representation of MWEs in the mental lexicon as a single entity. Some examples of Hindi multiwords are जल प्रपात (*jal prapaat, waterfall*), गर्भ गृह (*garBh grih, sanctum sanctorum*), and अंगुलि उठाना (*ungalee uThaanaa, accuse).*

MWEs are characterized by lexical, statistical, syntactic, semantic or, pragmatic idiosyncrasies. Of these, semantic non-compositionality has gotten special attention. MWEs span a continuum in terms of the semantics: from complete compositionality (*traffic signal*) to partial compositionality (*light house*) to complete non-compositionality (*green card*). Over time, MWEs get institutionalized and become lexicalized. For instance, *petrol pump* in India and *gas station* in United States have been institutionalized and are far more likely to be used than the potential synonym *petrol station.*

1.1. Motivation for Identifying Compound Noun MWEs

While MWE is an umbrella term covering syntactic categories like compound nouns (*wheel chair*), phrasal verbs (*put off*), verb phrase idioms (*kick the bucket*), light verb constructions *(make a demo*), etc., we have focused our efforts on the extraction of compound nouns MWEs from a text corpus. Compound noun is a class of MWE which is rapidly expanding due to the continuous need for coinage of new terms for describing new concepts, such as *multi word expression, gold standard,* and *web page*.

Identification of compound noun MWE can particularly help parsing, and dictionary based applications like machine translation, and cross lingual information retrieval, since such word

Proceedings of ICON-2008: 6th International Conference on Natural Language Processing
Macmillan Publishers, India. Also accessible from http://ltrc.iiit.ac.in/proceedings/ICON-2008

sequences should be treated as a single unit. The purpose of our work is to come up with a list of potential MWEs which a lexicographer can look at and decide whether a given word sequence should be added to the lexicon. This will aid the construction of a quality lexicon which incorporates MWE entries. Hence we err on the side of increasing recall when faced with a precision-recall tradeoff.

1.2. Our Contribution

In this work we have developed a system for Hindi compound noun multiword expressions (MWE) extraction from a given corpus. Our extraction methods utilize the statistical idiosyncrasy of MWEs, using statistical co-occurrence measures. We use lexical cues like hyphenation from the corpus and the use of rank aggregation to enhance the statistical methods. We also address the extraction of reduplicative expressions using lexical, semantic, and phonetic knowledge.

Due to the absence of the linguistic resources, we are not able to explore the semantic non-compositionality aspect directly. However, non-compositional compounds also exhibit statistical idiosyncrasies. Hence we believe that, statistical techniques can perform reasonably well without heavy linguistic resources.

We have also built an evaluation resource of compound noun MWEs for Hindi. Our methods give 80% recall and 23% precision at rank 1000.

A serious limitation of our approach is the use of a very small corpus – 160,000 words Hindi corpus. Particularly limiting is the use of PMI scores on such a corpus. In line with the claim by Dunning [6], we find that LLR is a much better association measure than PMI when dealing with very low collocation counts. In future, we need to work with a much bigger corpus.

In Section 2, we survey the related work. In Section 3, we describe our categorization of compound noun MWEs. Section 4 describes the methods used for compound noun MWE extraction. Section 5 describes the evaluation resources created and the methodology used. Section 6 presents the experimental details and results-discussion. Section 7 concludes the paper.

2. Related Work

Most MWE extraction methods are based on exploiting the various idiosyncrasies exhibited by MWEs. The variation in statistical distributional characteristics has been widely employed to test for evidence of a collocation being an institutionalized MWE. Pointwise Mutual Information is one of the earliest measures of association used for collocations [5]. Word association has also been measured using measures like Jaccard, Odds Ratio, etc [8]. Classical statistical hypothesis tests like Chi-square test, t-test, z-test, Log Likelihood Ratio [6] have also been employed to decide whether the constituents of a collocation are independent of each other. The variation in positional distribution of words in a collocation has also been used to identify significant collocations [7].

Lin [9] and Cruys et.al. [10] have used the principle of substitution to extract institutionalized collocations. They measure the difference between the distributional characteristics of the collocation and other similar collocations obtained by lexical substitution. For instance, *traffic signal* could have *traffic sign* and *traffic light* as similar collocations. If one of these collocations is highly preferred as compared to others, then it is likely to be an institutionalized MWE. The substitution tests measure this bias in preference for a collocation. While Lin uses PMI as the base association score, Cruys et.al. [10] use a strength of association measure motivated by the idea of selectional preference of a constituent word for another.

Linguistic properties of the MWE category under consideration are also a discriminating source of information. Fazly et.al. [16] extract MWEs by exploiting their syntactic fixedness. However, little work has been done to exploit linguistic features of compound nouns. This is probably because nouns are not richly inflected in English, and the internal structure and semantics is quite complex. Thus it is not easy to obtain hints for MWE extraction. Though many studies on semantic interpretation of compound nouns have been done [17], they have not been applied to the MWE extraction task.

In addition to the constituent words, the context in which the collocation is found can give clues about whether the collocation is a non-compositional MWE. Katz [11] and Baldwin [12] use the context as a bag of words and build context vectors for representing collocations and their constituents. Comparison of the collocation and constituent vectors helps determine if the collocation is non-compositional. In [13], Moiron et.al. have used the idea of translation ambiguity to extract non-compositional MWEs. The non-compositional collocations will have more translation candidates on account of more uncertainty in translation. This uncertainty is measured as translational entropy. Language modeling has been used to extract domain specific phrases, by comparing the distribution of collocations in a general and domain-specific corpus [14]. All the measures mentioned above have modeled the problem as a ranking problem, where the collocations more likely to be MWEs are ranked higher. If an annotated training set is available, the MWE extraction problem can be set up as a classification problem [15].

For Indian languages, automated MWE extraction work has been limited. In fact, both of the existing works [15, 18] use some kind of English translation for extracting Hindi MWEs. Mukerjee et.al. [18] have used parallel corpus alignment and POS tag projection with parallel English corpus to extract complex predicates. Venkatapathy et.al. [15] use a classification based approach for extracting N-V collocations for Hindi. They use identity of the verb, semantic type of the object, case marker with the object, similarity of the verb form of the object with the verb-object pair under consideration etc. as features in a MaxEnt classifier. In contrast, our focus is on extracting compound noun MWEs and many of their verb based features are not applicable in our case. We also focus on identifying reduplicative expressions using lexical, semantic and phonetic knowledge.

3. Categorization of Compound Noun MWEs

A compound noun is a noun consisting of more than one free morpheme. e.g. *black board, car driver, wheel chair*. Compound nouns can occur in open, closed, or hyphenated forms, e.g. *black board, blackboard, or black-board*. Such concepts in open form may be multiwords. However, not all compound nouns are MWEs. In the above examples, *black board* and *wheel chair* are MWE, while *car driver* is not. In this section, we discuss our work on developing criteria for identifying different kinds of compound noun MWEs. We first discuss how compound nouns satisfy the words-with-spaces paradigm of MWE. Then we discuss compound noun MWEs arising out of semantic, statistical, and linguistic criteria.

3.1. Compound Nouns as Words

A multiword expression is understood as a single word that happens to be written with spaces. Thus, for compound nouns to be MWEs, they must exhibit characteristics of a single word. The defining characteristics of a word [19] are:

- a part of speech specification.
- syntactic atomicity, meaning, words cannot be further analyzed by syntax; they are treated as a single unit for syntactic processing.
- one primary stress (usually).

Compound nouns exhibit these characteristics. The noun sequence denotes a nominal concept, hence it is a noun. In fact, in some POS tagsets, compound nouns have their separate tag. They generally act as a syntactic unit. Case markings and inflections are consistently applied to the head of the compound. The head represents the compound as a whole, and the inflections are not applicable for the head alone. This is evident if we compare headless compounds with one of the nouns having an irregular form. For example, the plural of *tooth* is *teeth,* which is an irregular form retained from old English, but the plural of *bluetooth* is *bluetooths*, and not *blueteeth.* Compound nouns also show a stress pattern, which is distinct from other noun phrases, the stress being left-prominent, at least in English and Hindi.

All these indicate that compound nouns are syntactic words. Thus, they satisfy a necessary condition for being MWEs. But this may not be sufficient for qualifying a compound noun as

MWE. The semantics and institutionalization of a compound noun plays a more important role in determining if it is an MWE. The next few sections explain the criteria for determining if a compound noun is indeed an MWE.

3.2. Semantic Non-Compositionality

A compound noun is an MWE if its meaning cannot be composed from the meanings of its constituent words. Such MWEs generally arise from figurative or metaphorical usage of the constituent words. e.g. *green card, wheel chair,* तरण ताल *(taraN taal, swimming pool)*. In general, MWEs span a continuum in terms of the semantics: from complete compositionality (*traffic signal*) to partial compositionality (*light house*) to complete non-compositionality (*green card*).

3.3. Statistical Co-Occurrence

An important question is whether compound nouns which are clearly compositional (e.g. *car driver, traffic signal,* समुद्र तट *(samudra taT, sea shore)*) are also MWEs. Current psycholinguistic models of morphological processing assume that compounds are processed in two ways - either by direct access or by the decomposition route and the faster route wins [19]. The access to a word depends on how frequently it is used, and the more frequently used words are accessed faster. This model of the mental lexicon suggests that not only non-compositional compounds, but highly frequent institutionalized compounds can also be MWEs. In addition, continued usage of a collocation in a particular context causes extra meaning to be associated with it. Hence, over time, institutionalized compound nouns acquire non-compositional semantics.

3.3. Linguistic Phenomena

Noun compounds generated by certain linguistic phenomena are also MWEs. Reduplication is one such linguistic phenomenon commonly found in many languages of India. The pair of words in a reduplication act as a single word syntactically and they denote a single concept. e.g. अस्त्र शस्त्र (*astra shastra*, *weapons)*. The meaning may be idiosyncratic as in दिन रात (*din raat, all the time*), साज सजावट (*saaj sajawat, decorations*). Reduplicative expressions are thus truly MWEs.

Following classes of reduplications commonly occur in Indian languages [20]:

Onomatopoeic expressions. The constituent words imitate a sound, and the unit as a whole refers to that sound. e.g. छन छन *(Chan Chan, sound of water falling on a hot surface)*, खट खट *(khat khat, knock knock)*.

Complete Reduplication. The individual words are meaningful, and they are repeated. e.g. कदम कदम *(kadam kadam, at every step)*, धीरे धीरे *(Dheere Dheere, slowly)*.

Partial Reduplication. Only one of the words is meaningful, while the other word is constructed by partially reduplicating the first word. There are various ways of constructing such reduplications, but the most common type in Hindi is one where the first syllable alone is changed. e.g. अलग थलग *(alag thalag, separated)*, रंग बिरंगा *(rang birangaa, colourful)*.

Semantic Reduplication. The two paired members are semantically related. The most common forms of relation between the words are synonymy (बाग़ बगीचा, *baag bagichaa, garden*), antonymy (लेन देन, *len den, dealing*), class representative (चाय पानी, *chaay paanee, snacks*)).

To summarize, there are three major criteria giving birth to compound noun MWEs, (1) semantic non-compositionality, (2) statistical co-occurrence, and (3) linguistic phenomena.

4. Compound Noun MWE Extraction

We have developed a system that extracts bigram compound nouns MWEs from a text corpus. It is an offline extraction system, which creates a ranked list of collocations. The higher a collocation is in the output list, the more likely it is to be an MWE.

To identify the different kinds of MWEs described in Section 3, our system relies mainly on the statistical co-occurrence information of the compound nouns. Statistical co-occurrence is a property exhibited by all kinds of MWEs.

Note that the existing discourse on MWE mostly centers on the semantic non-

compositionality aspect. However, determining semantic non-compositionality is a resource heavy process. It requires large amount of corpora, a knowledge of various semantic properties of words (for example, whether a given word is an abstract noun or a concrete noun), and a good parser. Due to the absence of the linguistic resources, we are not able to explore the compositionality aspect. However, we observe that non-compositional compounds also exhibit statistical idiosyncrasies. Hence we believe that, statistical techniques can perform reasonably well without heavy linguistic resources. Of course further improvement in performance will require us to look directly into compositionality aspect.

In our system, a POS tagger is run on the corpus and a list ofbigram compound noun candidates is prepared. Section 4.1 describes this process. For each candidate, statistical and lexical features like frequency, hyphenation, etc. are gathered. Using this information, statistical co-occurrence tests are run, as described in Section 4.2. In addition, linguistic tests determine MWEness arising from various language phenomena. These are described in Section 4.3.

Each extraction method creates a ranking of the collocations, the position indicating the confidence that the collocation is an MWE. These algorithms use different hints to determine whether a collocation is an MWE. We have implemented rank combination strategies to combine these individual rankings, to get a global ranking. Section 4.4 describes these methods.

4.1. Candidate Extraction

As the first step in the analysis, bigram noun sequences are extracted from a POS tagged corpus as MWE candidates. Ideally if the POS Tag set contains NNC tag, then one can just focus on all bigrams with the NNC tags. But with the present taggers, NNC tag can be quite unreliable for Indian languages. For example consider आम रस *(aam ras, mango juice)* in the following two sentences: आम रस से भरा है *(aam ras se bhara ha, the mango is full of juice)* and मुझे आम रस पीना है *(mujhe aam ras pina ha, I have to drink mango juice)*. In the first case, *aam ras* should get NN NN tags while in the second case, it should get NNC NNC.

But the tagger may give NN NN tag even in the second sentence. This unreliability results from the failure of phrase boundary detection. Given the unreliability of NNC (noun compound) tag, we err on the side of recall and consider bigrams consisting of all possible noun tags (NN, NNP, NNC, NNPC in our case). That is we try to ensure that all valid candidates are generated even if it means generating many invalid candidates. As a result, in a Subject-Object-Verb language like Hindi, the noun sequences detected by us may span phrases. For instance in लड़का आम खाता है *(laDakaa aam Kaataa hai, boy eats mango) laDakaa* and *aam* are in different phrases, yet it would be extracted as a bigram. A parser can help identify phrase boundaries and such errors can be avoided. Due to the unavailability of a robust Hindi parser, we are not able to eliminate such invalid candidates.

Some noun compounds may also be missed if the modifier is tagged as adjective. For instance, in *communist(JJ) national(NN) party(NN)*, *communist* is tagged as adjective. The solution can be to include the adjectival modifiers also in the candidate extraction. The choice depends upon the reliability of the POS taggers. The POS taggers we worked with were reasonably reliable in disambiguating the adjective-noun cases, and hence we restricted ourselves to extracting only noun sequences.

4.2. Statistical Co-Occurrence Tests

Statistical co-occurrence measures are calculated on each of the extracted candidates, and the candidate collocations are ranked by these measures. The following are the measures that have been used:

Frequency. Since MWEs generally get institutionalized, the frequency is a good first indicator of MWEness, given a large enough corpus. Hence candidate collocations are ranked by the frequency of occurrence in the corpus.

Pointwise Mutual Information. PMI measures the ratio of the joint distribution of the two

constituent words, assuming independence and otherwise [5]. Its value for a given bigram (x,y) is

$$\log \frac{P(x,y)}{P(x)P(y)}$$

PMI is prone to highly overestimating the occurrence of rare events.

Log Likelihood Ratio. The LLR test is a general test of significance [6]. In the context of statistically significant collocations, LLR is the log of ratio of the likelihood of observations assuming that the occurrence of the words in a collocation depend on each other to the likelihood assuming that the words occur independent of each other. Formally, it is the log of ratio of likelihood of observing given instances of bigram (x,y) under the following two hypotheses:

Hyp 1: P(y|x) = p = P(y|~x)
Hyp 2: P(y|x) = p1 ~= p2 = P(y|~x)

The probabilities are computed by modeling the frequencies of words in a corpus of size **N** as a binomial distribution and are shown to be equivalent to the following formulae in [23]:

$$2N\left(\sum_{x?\in\{x,\neg x\}}\sum_{y?\in\{y,\neg y\}} p(x?,y?)\log\frac{p(x?,y?)}{p(x?)p(y?)}\right)$$

Hyphen and Closed form count. Orthographic representation of a collocation may provide clues about the collocation being a MWE. Words joined with hyphens (*black-board*) or occurring in closed form (*blackboard*) are likely to denote a single concept or may be non-compositional. We therefore rank collocations according to their close-form count and hyphen-count. For the closed form count, we have considered the simple concatenation of words and have not taken into account any change in internal morphology of the concatenated words. e.g. नील (*neel, blue*) and अम्बर (*ambar, sky*) gives नीलाम्बर (*neelaamba, blue sky*), where the internal morphology is different from simple concatenation. Hence we do not treat these forms as equivalent.

Effective Frequency. The combined frequency of the open, closed and hyphenated form is referred to as the effective frequency of the collocation. We use effective frequency instead of simple frequency while computing LLR and PMI.

4.3. Identifying Linguistically Motivated MWEs

As described in Section 3.3, we use lexical, semantic and phonetic information for identifying the following kinds of reduplicative expressions:

Repetition: This category of reduplications is simple to identify, and we simply check if the two constituent words are the same.

Synonyms: We check if the two constituents are synonyms of each other. For this we have used the Hindi WordNet [21].

Antonyms: We use the antonymy lexical relation in the Hindi WordNet to check if the two words are antonyms of each other.

Partial Reduplication: We have handled only one kind of partial reduplication, commonly found in Hindi. Examples like अलग थलग *(alag thalag, separated)* and आर पार *(aar paar, right across*) illustrate this type. There is a clear pattern here. The first syllables of the words differ, while the other syllables are identical. Any collocation matching this criterion is a multiword. e.g. In the collocation अलग थलग *(alag thalag, separated)*, the first syllables of the words, अ and थ, are different. However, they share the remaining syllables, लग (*lag*). Devanagari, being a phonetic script, the syllable boundaries can be identified from the script. The first syllables and the remaining syllables of both words were identified. The above rule was then used to verify whether the candidate is a reduplicative expression.

4.4. Rank Combination

Each of the above methods gives a ranked list. We tried following two approaches to combining these ranked lists:

Weighted Combination. Different features are combined by assigning different weights to each feature and calculating a weighted sum of the individual scores. Before calculating the weighted sum, the individual scores are normalized so that they are in the range 0 to 1. It is bit debatable if such a normalization is meaningful. Luckily for

us, the next method of Rank Aggregation obviates the need for weighted combination.

Rank Aggregation. The aim is to combine ranked lists using information of the ordinal ranks of the elements in each list. No other information or score is used. Given multiple ordered lists l1, l2...l3 of a given set of elements, the rank aggregation problem is to *combine the individual rankings in a single ranked list.* This can be done by finding a consensus ranking that is at minimal distance from each of the individual rankings. This is a NP-complete problem [22].

Hence we use a popular rank aggregation heuristic called Borda's positional ranking [22]. Given lists t1, t2, t3 ... tk, for each candidate c and list ti, the score $B_{ti}(c)$ is the number of candidates ranked below c in ti. The total Borda score is $B(c) = \sum_i B_{ti}(c)$. The candidates are then sorted by descending Borda scores.

5. Evaluation Setup

To create an evaluation gold standard, manual identification of MWEs was done on an 80,000-word Tourism domain Hindi corpus. A total of 350 words bigram compound noun MWEs were identified, and categorized using following criteria: (1) semantic non-compositionality (2) statistical co-occurrence (3) linguistic phenomena. The collocation statistics were collected from a larger corpus of 160,000 words, containing 50,000 compound noun collocations. Using a larger corpus provided more evidence for the statistical measures we used.

We have used the standard IR metrics of Precision, Recall and F-1 score to evaluate the ranking methods. We calculate these metrics at different ranks, called Evaluation Points (EP).

Precision at evaluation point k is defined as:

$$\text{Precision}_k = \frac{|I_k|}{k}$$

Recall at evaluation point k is defined as:

$$\text{Recall}_k = \frac{|I_k|}{|M|}$$

F-1 score at evaluation point k is defined as:

$$\text{F-1}_k = \frac{2 \times \text{Precision}_k \times \text{Recall}_k}{\text{Precision}_k + \text{Recall}_k}$$

where,

M = MWE gold standard list

$\mathbf{I_k}$ = MWEs in the top k members of ranked list

6. Experimental Results

Tables 1 and 4 summarize the results of our experiments for Hindi.

Log-likelihood ratio performs best among the statistical co- occurrence tests. Frequency is also an important indicator of whether a compound is an MWE. However, PMI proves to be a bad measure due to the very small size corpus size. The entries at the top of the ranked list are dominated by low frequency collocations, proper nouns, and rare collocations, e.g. राव जोधा (raav jodhaa, Raav Jodha), आदर्श मरुस्थल (aadarsh marusthal, ideal desert). In these cases, the probabilities of the words are very small, inflating the PMI score. Therefore, we apply the PMI only in the case where collocation frequency is greater than two. In this case, quantitatively PMI performs as well as frequency, but qualitatively its behavior is very different, since it mostly picks reduplicate expressions towards the top. We want to emphasize that the bad performance of PMI is due to the small frequencies being encountered in our small corpus, including the gold standard, and not because it is inherently unsuitable for the task.

The performance metrics clearly indicate that the hyphenation and closed form count features are strong indicators of a compound being an MWE. This agrees with our conjecture that such surface cues can aid MWE extraction. These are high precision, low coverage cues. Significantly, there is less overlap between the rankings of LLR and these features. This suggests that it might be fruitful to combine the statistical co-occurrence and the lexical cue based rankings. The use of effective frequency for ranking also gives significantly better performance as compared to the original frequency. MWEs like तट रेखा (*taTa rekha, coast-line*) and द्वीप समूह (*dweep samuh, archipelago*) had their effective frequencies boosted by use of the hyphenation and closed

form counts, providing stronger evidence for them being MWEs.

For the rank combination experiments, we combined the best co-occurrence measure, LLR, with hyphen count and closed form count. For the weighted combination method we tried various weights. The results are reported for the weight triple (0.33, 0.33, 0.33). The weighted combination based approach improves upon each of the individual methods. The rank aggregation based combination also performs equally well, but did not require any empirical setting of weights. The rank aggregation method can thus serve as an effective automated MWE extraction technique.

Reduplication extraction is a low coverage, high accuracy method. As more kinds of reduplications are handled, the system's accuracy will improve. Echo words and synonym reduplications were extracted accurately. Coverage of antonyms is low in the Hindi WordNet [21], hence antonym reduplicates are not easily found.

We obtain a combined ranking by concatenating the two rankings, the reduplication and the rank aggregation ranking. We are confident of the high accuracy of the reduplication extraction, so we put the reduplicate expressions ahead in the combined rankings. This gave the best extraction system for Hindi in all our experiments.

The presence of named entities in the top ranked results also affects the performance. While conceptually all named entities are multiwords, we do not include them in our gold standard. Hence we deliberately underreport our performance. Elimination of these named entities should further improve the accuracy of the system.

6.1. Applicability to Other Languages

We also applied our techniques to the Marathi and English. We used a Tourism domain corpora for English and Marathi too. In fact, these corpora are parallel to the Hindi corpora used. Compared to the 160,000 words in Hindi, Marathi corpora has 140,000 words while the English corpora has 210,000 words. Tables 2 and 3 summarize precision results for the different methods experimented. Since we do not have the Gold Standard for English and Marathi, we are not able to compute Recall. Precision is computed by manually evaluating the accuracy for the reported results. We observe that closed form counts are useful for Hindi and English, but not for Marathi. The Marathi orthographic convention allows all compound nouns to be written without spaces regardless of the compositionality of the meaning. However, hyphen counts still seem useful for Marathi. We did not have enough instances of hyphen count for English in our corpus

Evaluation Point	Frequency			PMI (Freq > 2)			Effective Frequency			Hyphen Count			Closed Form Count		
10	60.0	2.1	4.05	40.0	1.4	2.7	60.0	2.1	4.0	80.0	2.8	5.4	70.0	2.5	4.7
50	38.0	6.6	11.3	34.0	5.9	10.1	54.0	9.4	16.1	74.0	12.9	22.0	72.0	12.6	21.4
100	31.0	10.8	16.1	26.0	9.0	13.5	38.0	13.3	19.7	69.0	24.1	35.8	60.0	21.0	31.1
200	32.5	22.7	26.7	24.5	17.1	20.1	40.5	28.3	33.3	52.0	36.4	42.8	53.7	25.2	34.3
500	22.2	38.8	28.2	22.8	39.9	29.0	25.4	44.4	32.3	38.2	58.0	42.2	NA	NA	NA
1000	15.1	52.8	23.5	NA	NA	NA	16.8	58.7	26.1	29.7	65.0	40.8	NA	NA	NA
Evaluation Point	LLR			Rank Aggregation			Weighted Combination			Reduplication			Best Performing Method		
10	80.0	10.4	2.75	90.0	3.2	6.1	70.0	2.5	4.7	90.0	2.8	5.41	90.0	2.8	5.41
50	50.0	8.7	14.9	78.0	13.6	23.2	72.0	12.6	21.4	68.0	11.9	20.24	68.0	11.9	20.24
100	41.0	14.3	21.2	64.0	22.9	33.2	65.0	22.7	33.7	66.18	15.8	25.42	66.18	15.8	25.42
200	39.5	27.6	32.5	57.0	39.9	46.9	54.0	37.8	44.4	NA	NA	NA	57.5	40.2	15.8
500	26.0	45.6	33.1	36.2	63.3	47.1	34.0	59.4	43.2	NA	NA	NA	35.8	62.6	16.5
1000	16.4	57.3	25.5	22.7	79.4	35.3	22.5	78.7	35.0	NA	NA	NA	22.7	79.4	17.3

Table 1: MWE Extraction Results for Hindi. The three columns for each method correspond to the Precision, Recall, and F-Score in that order

Evaluation Point	Freq.	Effective Freq.	Hyphen Count	Closed Form Count	LLR	Rank Aggr.
10	20	50	NA	80	40	70
50	20	32	NA	44	28	38
100	15	29	NA	34	23	27
200	11	18.5	NA	NA	19.5	27

Table 2: Precision Results for English

Evaluation Point	Freq.	Effective Freq.	Hyphen Count	Closed Form Count	LLR	Rank Aggr.
10	10	10	30	10	10	50
50	2	6	18	10	6	22
100	4	7	19	10	6	17
200	3.5	4.5	19	NA	5	11.5

Table 3: Precision Results for Marathi

Effective Frequency	LLR	Hyphen Count	Closed Form Count	Rank Aggregation	Reduplication	PMI (Freq. > 2)	Marathi (Rank Aggr.)	English (Rank Aggr.)
किलो मीटर	किलो मीटर	समुद्र तट	किलो मीटर	प्रवेश द्वार	अस्त्र शस्त्र	रहन सहन	साहस पर्यटन	kilo meters
समुद्र तट	समुद्र तट	खान पान	जल प्रपात	समुद्र तट	आकार प्रकार	तडक भडक	इसवी सन	wild life
राष्ट्रीय उद्यान	प्रवेश द्वार	प्रवेश द्वार	वास्तु शिल्प	जल प्रपात	आचार व्यवहार	वेश भूषा	आखीव रेखीव	sand stone
जल प्रपात	राष्ट्रीय उद्यान	भीड भाड	भू दृश्य	द्वीप समूह	आतिथ्य सत्कार	ताम झाम	शंख शिंपला	water falls
प्रवेश द्वार	जल प्रपात	उत्तर पश्चिम	तट रेखा	उत्तर पूर्व	आमना सामना	चमक दमक	मंदिर संकुल	court yard
वर्ग किलोमीटर	खान पान	रंग बिरंगे	प्रवेश द्वार	भू दृश्य	आमोद प्रमोद	रीवर क्रूस	ध्वनि प्रकाश	north east
संयुक्त राज्य	वर्ग किलोमीटर	उत्तर पूर्व	वन्य जीवन	शासन काल	आर पार	वॉल हैंगिंग	जिल्हा मुख्यालय	back drop
खान पान	वास्तु शिल्प	भित्ति चित्र	द्वीप समूह	तट रेखा	आस पास	उथल पुथल	शहर विराम	south east
भू दृश्य	संयुक्त राज्य	आर पार	कार्य कलाप	भित्ति चित्र	उथल पुथल	सास बहू	पायर्या पायरा	country side
वास्तु शिल्प	भीड भाड	प्रवाल भित्ति	समुद्र तट	दक्षिण पूर्व	उलट पलट	ऊबड खाबड	मौज मजा	light house

Table 4: Top 10 Hindi MWEs extracted by different methods (except last two columns)

7. Conclusions

We have developed a compound noun MWE extraction system which ranks collocations using statistical methods. We use lexical cues like hyphenation from the corpus and the use of rank aggregation to enhance the statistical methods.

Complete automation of the MWE extraction is still a difficult task. Our methods however can improve the lexicographer productivity by providing them with a list to select MWEs. A precision of 23% at rank 1000 means that one in four-five collocations observed by the lexicographer will be an MWE. A recall of 79%

means that most of the MWEs in the corpus are in the top 1000.

Some serious limitations of our approach are the use of a very small corpus and the absence of a Name-Entity recognizer.

While the current work was focused largely on Hindi, we would like to evaluate the effectiveness of our methods for MWE extraction in other languages more thoroughly. We would also like to extract MWEs by exploiting the semantic non-compositionality characteristics.

Acknowledgements

We would like to thank all the CFILT members who spent lot of time wondering what an MWE is and what it is not. In particular, we want to thank Prabhakar Pandey and Subodh Kembhavi for help with Hindi and Marathi evaluations. We also want to thank anonymous referees for many valuable suggestions which helped improve the presentation a lot.

References

[1] A. Wray. *Formulaic Language and the Lexicon.* Cambridge University Press. 2002.

[2] I. Dahlmann and S. Adolphs. *Pauses as an indicator of psycholinguistically valid multi-word expressions (MWEs)?*. ACL-2007 Workshop on Multiword Expressions, 2007.

[3] C. Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press. 1998.

[4] I. Sag, T. Baldwin, F. Bond, A. Copestake, and D. Flickinger. *Multi-word expressions: A Pain in the neck for NLP*. CICLing, 2002.

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

[6] T. Dunning. *Accurate methods for the statistics of surprise and coincidence*. Computational Linguistics. 19(1), 1993.

[7] F. Smadja. *Retrieving collocations from text: Xtract*. Computational Linguistics. 19(1), 1993.

[8] P. Pecina. *An extensive empirical study of collocation extraction methods*. ACL Student Research Workshop. 2005.

[9] D. Lin. *Automatic identification of non-compositional phrases*. ACL 1999.

[10] T. de Cruys and B. V. Moiron. *Semantics-based multiword expression extraction*. ACL-2007 Workshop on Multiword Expressions, 2007.

[11] G. Katz and E. Giesbrechts. *Automatic identification of noncompositional multi-word expressions using Latent Semantic Analysis*. ACL-2006 Workshop on Multiword Expressions. 2006.

[12] T. Baldwin, C. Bannard, T. Tanaka, and D.Widdow. *An empirical model of multiword expressions decomposability*. ACL-2003 Workshop on Multiword Expressions. 2003.

[13] B.V. Moiron and J. Tiedemann. *Identifying idiomatic expressions using automatic word alignment*. EACL 2006 Workshop on Multiword Expressions in a multilingual context. 2006.

[14] T. Tomokiyo and M. Hurst. *A language model approach to keyphrase extraction*. ACL-2003 Workshop on Multiword Expressions. 2003.

[15] S. Venkatapathy and A. Joshi. *Relative Compositionality of Noun+Verb Multi-word Expressions in Hindi*. ICON-2005.

[16] A. Fazly and S. Stevenson. *Automatically constructing a lexicon of verb phrase idiomatic combinations*. EACL. 2006.

[17] M. Lauer. *Designing Statistical Language Learners: Experiments on Noun Compounds*. PhD thesis, Macquarie University. 1995.

[18] A. Mukerjee, A. Soni, and A. Raina. *Detecting Complex Predicates in Hindi using POS Projection across Parallel corpora.* Proceedings of the Workshop on Multiword Expressions at ACL-2006.

[19] I. Plag. *Word Formation in English*. Cambridge University Press, 2003.

[20] E. Keane. *Echo Words in Tamil*. PhD thesis, Meriton College, Oxford, 2001.

[21] D. Narayan, D. Chakrabarti, P. Pandey, and P.Bhattacharyya. *An experience in building the Indo WordNet - a WordNet for Hindi*. Global WordNet Conference, 2002.

[22] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. *Rank aggregation methods for the web*. 10th World Wide Web Conference (WWW) 2001.

[23] R. C. Moore, 2004. On Log-likelihood-Ratios and the Significance of Rare Events. EMNLP 2004.