Multiword Expressions in the CLIA Project

Anoop Kunchukuttan, Munish Minia, Pushpak Bhattacharyya
Department of Computer Science and Engineering
IIT Bombay
{anoop.k,pb,munishminia}@cse.iitb.ac.in

Abstract:- ***A multiword expression (MWE) can be said to be a word collocation that exhibits characteristics of a single syntactic word. In the recent past, the important role played by multiword expressions in language has been recognized by the natural language processing community. They pose significant challenges for NLP since they lie somewhere between words and larger syntactic units, and hence break many of the assumptions about the analysis of language. Specifically, within cross lingual information retrieval, multiword expressions are encountered very frequently and it is essential to provide correctly process MWE in order to provide improved information retrieval results.***

In this paper, we analyze these challenges provided by MWEs in the context on cross-lingual information retrieval and discuss our efforts to tackle the problem in the 'India Search' CLIA (Cross Lingual Information Access) system.

I. Introduction

A multiword expression (MWE) can be said to be a word collocation that exhibits characteristics of a single syntactic word. In the recent past, the important role played by multiword expressions in language has been recognized by the natural language processing community. It has been described as a 'pain in the neck' of NLP. They pose significant challenges for NLP since they lie somewhere between words and larger syntactic units, and hence break many of the assumptions about the analysis of language. Specifically, within cross lingual information, multiword expressions are encountered pretty frequently and it is essential to provide correctly process MWE in order to provide improved information retrieval results.

Section II provides an overview of multiword expressions and their impact on Natural Language Processing. Section III discusses the importance of MWE processing to cross-lingual information retrieval and specifically, their integration into the CLIA system. Section IV discusses the various challenges in processing MWEs in the CLIA system. Section V describes our current investigations into the handling of MWEs.

II. Multiword Expressions

A. What are Multiword Expressions?

It is generally understood that a 'word' in a language is the smallest unit of meaning that can stand in isolation. For the native speaker of a language, a word like book, card and blackboard would conjure association to a particular concept (or concepts). Now consider the compound noun green card, which is used in the sense of authorization for permanent residency in the United States. Here, the compound green card acts as though it stands for a single concept, and its meaning cannot be understood from those of the constituent words green and card. Similarly, 'petrol pump' would signal a specific concept, and this usage has been institutionalized in everyday use and it would be unlikely for equally valid alternative representations of the concept like 'petrol shop' or 'petrol stations'. In Hindi, गभ गृह *(garBh grih, sanctum sanctorum)* and जल परपात *(jal prapaat, waterfall)* are examples of collocations that exhibit similar behaviour. Such collocations take the nature of words-with-spaces and are generally referred to as Multiword Expressions.

B. Analysis of Multiword Expressions

Carrying the words-with-spaces interpretation forward, MWE can be said 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'*. This psycholinguistic interpretation is not sufficient for studying properties of MWEs, and an analysis from syntactic, semantic and empirical viewpoints will be useful to understand them.

1) Semantic Compositionality

Non-compositionality of the meaning of collocations from the constituent meanings is the key criteria for identifying MWEs. MWEs may be completely non-compositional or partially compositional. In the former case, the semantics are totally opaque as illustrated by *promise one the moon* or उंगली उठाना (ungalI uthanA,accuse). On the other hand, idioms like *spill the beans* are partially compositional because *spill* is being used in the sense of reveal and *beans* metaphorically represent a secret.

2) Empirical Observation

Some collocations are used together even though they are perfectly compositional, and there exist alternatives for the constituent words. This suggests that the usage of that collocation has been frozen and institutionalized. The statistical evidence would make them strong candidates for being multiwords, although a linguistic grounding for explain them is still lacking. E.g. *traffic signal*, समुद्र तट (samudra taT, sea shore)

3) Syntactic Analysis

MWEs, either non-compositional or institutionalized, can also be divided into various syntactic categories.

1. Compound Nouns: A sequence of words acts as a single noun. These form part of noun phrases and could be proper nouns or common nouns. E.g. *traffic signal, George Bush, green card.*

2. Phrasal Verbs: These are collocations containing a verb followed by a preposition or adverb, acting as a participle, which in fact stands for a single concept. E.g. *broke up, gobble up, look round, get over.*

3. Light Verb Constructions (Conjunct Verbs): These consist of a verb and a noun/adjective/adverb collocation as in *fall asleep* or *make a demo*, where the semantics is not completely non-compositional. It also includes collocations in Hindi using the verb करना *(karnA, to do),* like हमला करना (*hamlA karnA, to attack*). The verb does not provide semantic content, instead performing only a syntactic function for the host word (noun/adjective/adverb).

4. Verb Phrase Idioms: These consist of a verb phrase whose semantics are highly non-compositional. E.g. *promise him the moon*, *blow hot and cold*.

5. Grammaticalized Verb Sequences: The are Verb+Verb sequences, where the

main verb carries the semantic content while the other verbs indicates syntactic information with some assisting semantic role. Examples would include constructions in Hindi like बोल उठा (bola uXaa), गरि पडा (gir padaa). Here the main verb interacts with a closed class of light verbs like पड़ना *(panda)*, लगना *(lagnA)*, उठना *(uXnA)*.

6. Reduplications: Entire words or parts of words are repeated. The whole sequence denotes a concept. e.g. In Hindi अलग ठलग *(alag thalag, separated)*.

C. The role of MWEs in NLP

These multiword expressions pose challenges for Natural Language Processing due to this characteristic of being lexically represented as independent words while have strong semantic cohesion. We discuss the role that MWEs can play in NLP:

- Incorporating MWEs into lexicons will provide a greater coverage and truer representation of language vocabulary. This will need a revision of the way lexicons are organized today to account for the syntactic and morphological characteristics of MWEs.
- Parsers accounting for MWEs will be able to generate correct parse trees. Having truer representation of parsing and lexicons will help downstream NLP applications like information extraction, sentiment analysis, etc.
- MWE analysis is imperative for Machine translation since a word-by-word translation will not be adequate to capture the semantics of the source language sentence.
- In information retrieval, analysis of MWE will give a better indication of the user intent. Perhaps of greater value is the aid that multiword expressions can provide in query translation for cross lingual information retrieval.

III. The Role of Multiword Expressions in CLIA

A. Benefits of MWE Analysis in CLIA

While using any information retrieval systems, users are generally looking for ‘things’ to satisfy their information needs. Therefore, most IR queries tend to nouns forming about 30-40% of the queries. New compound nouns tend to be created in a language, and there is a high incidence of compound nouns in IR queries. Many of these compound nouns may be MWEs. It is therefore important that CLIA systems give primary importance to compound noun MWEs.

Identifying and handling MWEs can help to understand the user intent better in an IR system. Query expansion can be effective if MWE lexicon and synonyms are available. IR systems make use of a number of ‘signals’ which go into determining the importance of a result record to the query. Being a multiword may be an important ‘signal’ which would be useful for re-ranking top-k results and for ensuring the diversity of top-k results. For instance, *Dutch treat* has literal and idiomatic meanings, and identifying this MWE can ensure that results pertaining to both senses are retrieved.

Identification of MWEs assumes more criticality in the context of cross-lingual information access. Due to their non-compositional nature, query translation without identification of multiwords will not be effective. In the absence of any context for query translation, identification of multiwords can help to narrow the search space for query translation.

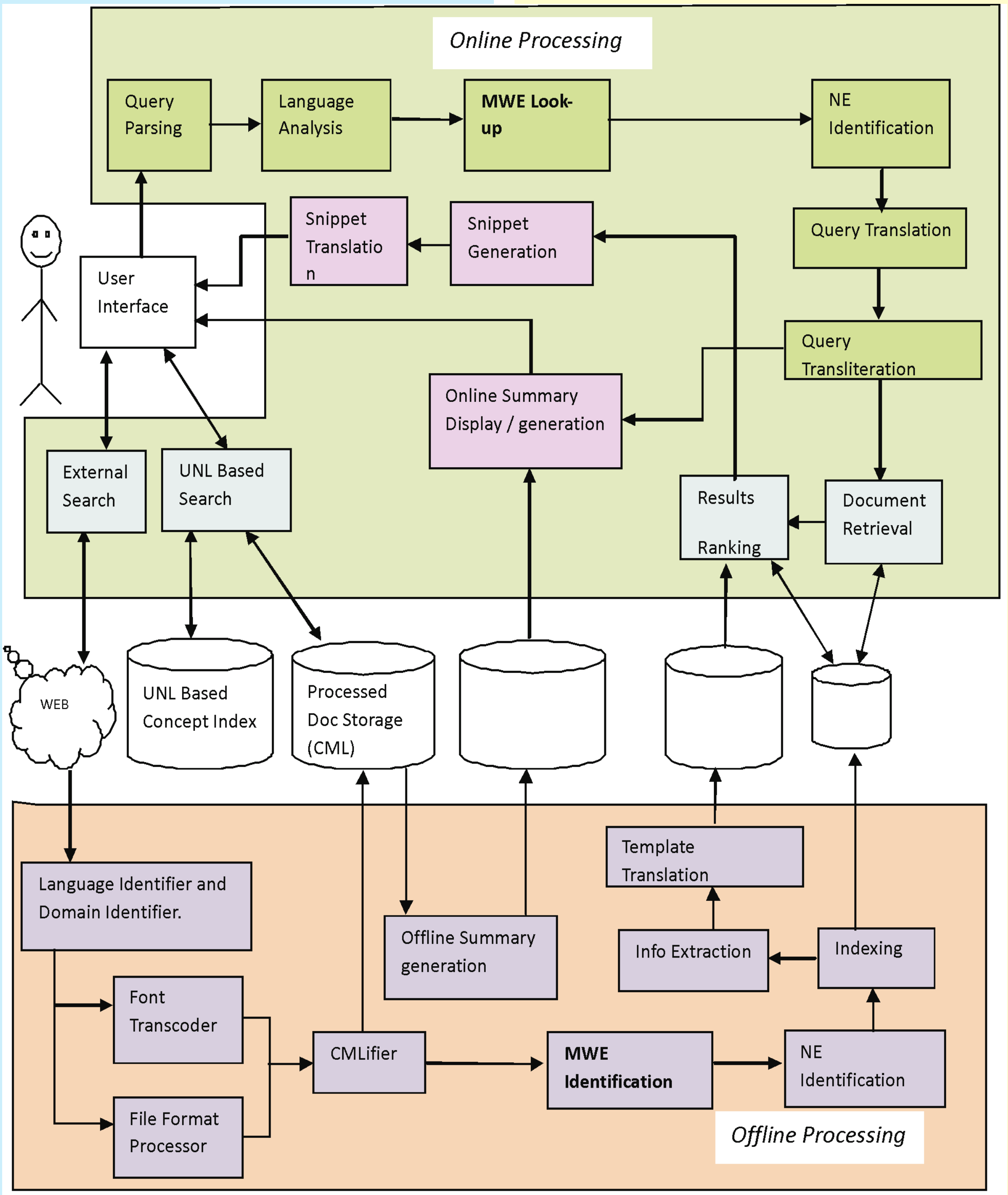


Figure 1 CLIA Architecture (Source: [3])

B. MWE Processing in the CLIA Architecture

The CLIA project is a CLIR system being developed by a consortium of academic and research organization to further research in cross-lingual information languages in Indian languages. It puts an emphasis of linguistic analysis of the data so as to be able to serve relevant search results. Fig 1 shows a high level architecture of the CLIA system.

Data is processed through the CLIA system through offline and online processes. Offline processes refer to all processing involved in getting the data ready to handle search requests, whereas online processing refers to the query processing. MWE analysis has a role to play in both these processes.

1) Offline Processing

1. An **MWE lexicon** is created through corpus analysis and linguistic validation. This lexicon is an important resource for MWE recognition in the corpus and queries.

2. While indexing, MWE units are identified in the text and indexed.

2) Online Processing

1. As part of query analysis, MWEs in the query are identified on the query language side.

2. The query is translated in the target language. At this stage, the MWE expressions identified are also translated into the target language by using a bi-lingual MWE dictionary. The source side MWE identification can also be used to disambiguate the query.

IV. Challenges in MWE Processing For CLIA

This section describes the challenges in processing of MWEs in a CLIA system:

A. Identifying Multiword Expressions in queries

Identifying MWEs during query processing is a key task. This is a difficult task because there is little context available to know if there is non-compositionality involved. e.g. in a query like Dutch treat Amsterdam the sense of Dutch treat is ambiguous. This query ambiguity may work with monolingual IR, which is based on keyword match.

But CLIR will need the right sense to translate the query correctly into the target language, and hence the MWE needs to be identified. One way to identifying MWE is to rely on knowledge learnt offline. This can be in the form of a MWE lexicon, which is then looked up during query processing to identify the MWE in the query. The CLIA project has taken this approach for compound nouns MWEs.

B. Multiword Expressions in Query Translation

The CLIA system relies on query translation to get the semantics of the source language query into the target language query, and is thus a key component. It is a difficult problem to address since the sense cannot be disambiguated easily from the short context available. For MWEs, even if the sense of each word is understood, the literal translation would not address the problem. CLIA proposes to use bilingual MWE dictionary to identify the target language translation for MWEs in the source language query.

C. Building Multiword Expression Lexicon

It is then obvious that the handling MWEs in the CLIA system requires a wide coverage MWE lexicon. A completely manual process of building this resource will not be scalable for a large scale IR system. There has been a lot of work in developing methods for automatic extraction of MWEs from corpora. The fundamental idea behind these techniques is to use the notion of statistical co-occurrence to identify potential MWEs. Thus, institutionalized MWEs can be detected effectively with these methods. However, many collocations extracted would be coincidental and completely compositional. While having these entries in the MWE dictionary wouldn't affect the coverage of MWE search during query processing, an unnecessarily large dictionary would have an impact on the query processing time. It is thus necessary to prune the MWE candidates identified through corpus analysis.

The CLIA project plans to employ linguistic validation of the corpus-extracted MWEs to achieve the right balance between coverage and processing efficiency. While the automated methods can alleviate the problem of identifying the institutionalized MWEs to some extent, non-compositional MWEs are still a challenge to identify if there aren't enough occurrences in the corpus to identify them.

Finally, as more data is crawled, it is possible to incorporate information from the new corpus into MWE analysis. This raises a couple of issues:

1. Will the entire corpus have to be completely processed to extract MWE? It would be interesting to investigate incremental algorithms to update the information on extracted MWEs based on the freshly crawled data.
2. What would be the policy for updating the MWE lexicon in when the MWE analysis is updated? In the face of new evidence will some MWEs now be removed from the lexicon?

Another key question is the representation of the MWE in the lexicon. How will information on the syntactic and morphological variations that the MWE can take be maintained in the system? This will be important for stemming and index representation of MWEs.

D. Building Multiword Expression Parallel Lexicon

The availability of a monolingual MWE lexicon alone is useful only in a monolingual search. However, it still does not represent an advancement in cross-lingual search if the target language query can be properly translated. Translating MWEs will require a parallel MWE lexicon between the source and target languages. Building such a lexicon manually would be daunting. Automated or even semi-supervised construction of parallel MWE lexicon is still an unaddressed problem.

V. An Approach to building a MWE Lexicon

We have developed an MWE extraction system, which can extract compound noun bigram MWEs from a corpus, which will create a ranked list of collocations, given a POS tagged corpus. Section V.A describes the system, whose components are shown in Figure 2. The collocations are then validated by linguistics before they are incorporated into the MWE dictionary. Section V.B describes the validation tool used for this process.

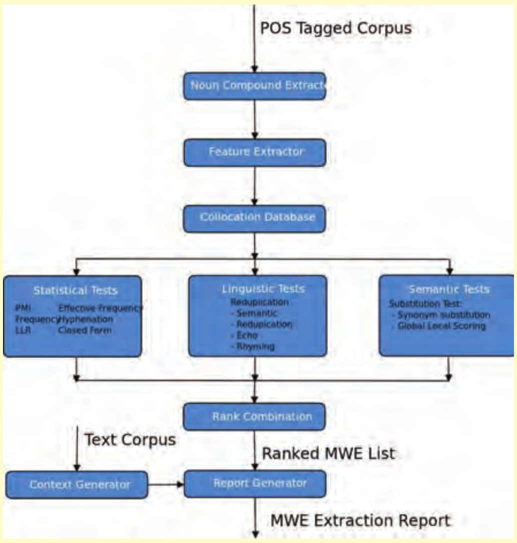


Figure 2 MWE Extraction System (Source : 1)

A. Multiword Expression Extraction System

A part-of-speech tagged corpus is fed to the system. First, the *Compound Noun Extractor* module extracts compound nouns and prepares a database of bigram candidates to be analyzed. This is done using a set of regular expressions over POS tags.

For each collocation, the POS tagged corpus is a source of lexical and linguistic information like frequency, concatenation, etc. Statistical and lexical features about each collocation are gathered by the *Feature Extractor* module, which is used for analysis by the extraction algorithms.

Collocation information is stored in the *Collocation Database,* and provides the information for the MWE extraction algorithms. It is the central data repository, which manages all the collocation data. This includes the extracted features and various scores calculated for each collocation. It allows iteration through the collocation database and sorting on various scores.

Three sets of algorithms are then run on the collocations:

1. Statistical Co-occurrence Tests: These tests exploit the statistical idiosyncrasy exhibited by MWEs, by calculating various co-occurrence measures, to look for evidence for qualifying a collocation as an MWE.

2. Linguistic Tests: This module is dedicated to extracting MWE arising from language phenomena. Currently, we detect different kinds of reduplications using lexical and phonetic information. It handles phenomena like repetition, synonymy, antonymy, related words and rhyming.

3. Semantic Tests: This module uses semantic information to detect institutionalization or semantic non-compositionality. Presently, we make use of the substitution principle to search for evidence of institutionalization or semantic non-compositionality.

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. *The Rank Combination module* uses various rank aggregation strategies to combine these individual rankings, to give a better global ranking.

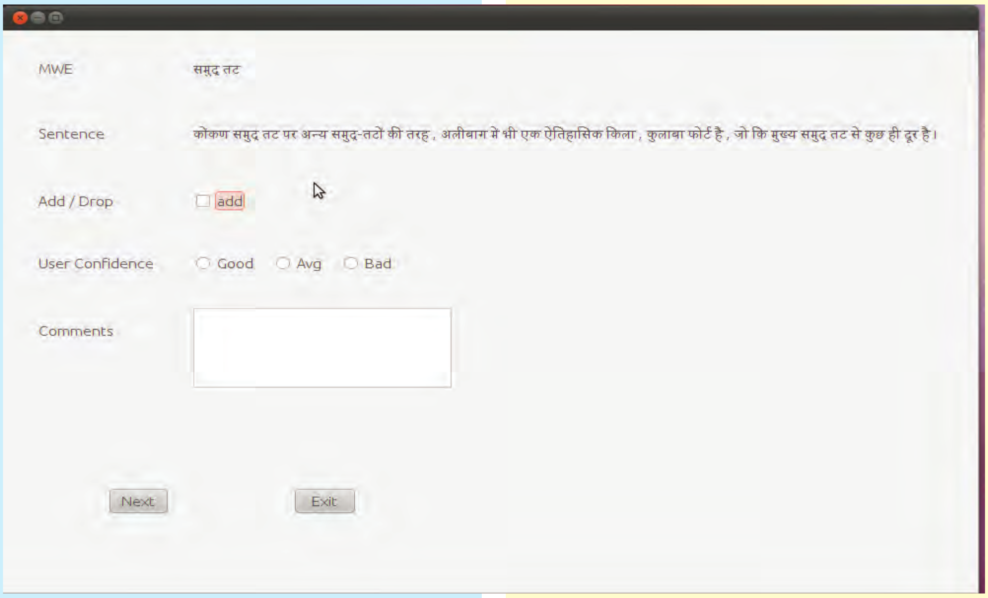
B. Multiword Expression Validation tool

The CLIA project has decided to use human validation in order to verify the MWEs generated by the extraction engine. To help lexicographers with the validation task, a validation tool has been developed.

This tool displays the ranked MWE list generated by the extraction engine to the lexicographer. For the candidate MWE, a contextual sentence is also shown. Using these the lexicographer can decide if the collocation is an MWE. The candidates that are approved are compiled into an MWE dictionary by the tool. Figure 3 shows a snapshot of the validation interface.

The MWE tool allows multiple lexicographers to work on the same dataset, and merge their respective dictionaries later. The lexicographer can also save her work at any point into a project and resume later from there.

The Multiword Expression Dictionary can be searched and browsed through the MWE Dictionary Browser. The browser also allows addition and deletion of MWE entries from the dictionary. Figure 4 shows a snapshot of the dictionary browser.



MWE समुद्र तट

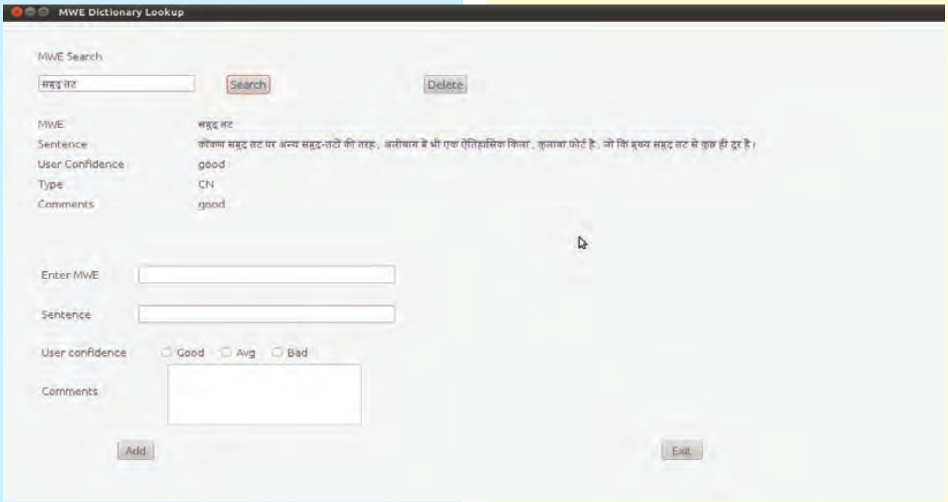
Sentence कोंकण समुद्र तट पर अन्य समुद्र-तटों की तरह , अलीबाग में भी एक ऐतिहासिक किला , कुलाबा फोर्ट है , जो कि मुख्य समुद्र तट से कुछ ही दूर है।

Add / Drop □ add

User Confidence ○ Good ○ Avg ○ Bad

Comments

Figure 3 MWE Validation Tool (Source: [4])



MWE Dictionary Lookup

MWE Search

समुद्र तट

MWE समुद्र तट

Sentence कोंकण समुद्र तट पर अन्य समुद्र-तटों की तरह , अलीबाग में भी एक ऐतिहासिक किला , कुलाबा फोर्ट है , जो कि मुख्य समुद्र तट से कुछ ही दूर है।

User Confidence good

Type CN

Comments good

Enter MWE

Sentence

User confidence ○ Good ○ Avg ○ Bad

Comments

Figure 4 MWE Dictionary Browser (Source: [4])

VII. Acknowledgements

The research was supported by Department of Information Technology, Govt. of India through the sponsored project Cross Lingual Information Access, code 10DIT012.

VIII. Bibliography

[1] A. Kunchukuttan, “Multiword Expression recognition for compound nouns”, M.Tech Thesis, IIT Bombay, 2008.

[2] A. Kunchukuttan and O. Damani, “A

system for compound noun multiword expression extraction for Hindi", 6th International Conference on Natural Language Processing, 2008.

[3] Y. Kakade, A. Atreya, P. Bhattacharyya,"Cross-Lingual Information Access", unpublished.

[4] A. Kunchukuttan, M. Minia, "User Manual for MWE Tool", unpublished.

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

[6] I. Plag, "Word Formation in English", Cambridge University Press, 2003.

[7] K. Church and P. Hanks, "Word association norms, mutual information, and lexicography", Computational Linguistics, 1990.

[8] T. Dunning, "Accurate methods for the statistics of surprise and coincidence", Computational Linguistics, 1993.

[9] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar,

[10] "Rank aggregation methods for the web", 10th WorldWide Web Conference (WWW) 2001.

[11] D. Lin. "Automatic identification of noncompositional phrases" ACL 1999.

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

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

[14] A. Fazly and S. Stevenson, "Automatically constructing a lexicon of verb phrase idiomatic combinations", EACL, 2006.

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

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

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