Literature Survey on Multi-Lingual Multiword Expressions

By

Munish Minia

Roll No: 07D05016

Under the guidance of

Prof. Pushpak Bhattacharyya

Abstract

The literature survey defines the Multiword Expressions and aims to provide a principled way to identify Multiword Expressions in different Indian Languages. Three types of Multiword Expressions namely, Noun + Noun (compound noun), Noun + Verb (conjunct verb) and Verb + Verb (compound verb) sequences are examined. It concentrates on the linguistic methods, like chunker, part-of-speech tagging, and the statistical methods like, point wise mutual information, log-likelihood, to extract the Multiword Expressions.

The study focuses on the following:
A. isolation Noun + Noun combinations, which are compound nouns, from other noun sequences
B. isolating Verb + Verb and Noun + Verb combinations, which are complex predicates, from other verb-verb and noun-verb sequences

Along with this, different approaches to extract Multiword Expressions are also discussed at the end.

Contents

Chapter 1: MWE and its Survey

In this chapter we firstly define the MWE, and the necessary and sufficient condition to be a MWE. Then we move on to the different characteristics of MWE. In next part we introduce types of MWE that we are concentrating on. After that we move on to the ongoing projects that are going on MWE and the methods to extract them which include both statistical and Linguistic methods.

1.1 Definition of MWE

Various Researchers has defined MWE differently, but all leading to a one concept only.

- **[Sag02]** defines MWE as: A Multiword Expression is an idiosyncratic expression which crosses word boundaries, the idiosyncrasy being lexical, statistical, syntactic, semantic or pragmatic.
- **[Bal02]** defines MWE as: A sequence of words that acts as a single unit at some level of linguistic analysis.
- **[Ren09]** defines MWE as: A well thought-out as sequence with comparatively fixed structure which represents a special meaning. The exact meaning of an MWE is not directly obtained from its component parts.

1.2 Necessary and Sufficient Condition

1.2.1 Necessary Condition

For a word sequence to be a MWE, it has to be separated by space/delimiter. This condition was decided in the Kashmir Multiword Workshop 2011.

Example: இந்திய கிரிக்கெட் அணி (Tamil)

- **Transliteration:** Indhiya kirikket ani
- **Gloss:** India cricket team
- **Translation:** Indian Cricket Team

1.2.2 Sufficient Condition

The sufficient condition to be a MWE is:

1. The non-compositionality of meaning of the MWE, i.e. meaning of a MWE cannot be derived from its constituents. Examples:
    a. पोटांत चाबता **(**Konkani**)**
        - **Transliteration:** potamta cabata
        - **Gloss:** biting in the stomach
        - **Meaning:** to feel jealous

    b. చెట్టు కిందికి ప్లీడరు (Telugu)
        - **Transliteration**: cevttu kimda plidaru
        - **Gloss:** a lawyer sitting under the tree
        - **Translation:** an idle person

2. The fixity of expression, i.e. the constituents of MWE cannot be replaced by its synonyms or other words. Examples:
    a. life imprisonment
        - We do not say: lifelong imprisonment
    b. Many thanks
        - We do not say: Plenty thanks

1.3 Idiosyncrasies observed in MWE

A glance at the types of idiosyncrasies observed in MWEs follows:

- **Semantic**: A very important class of MWEs is one where the meaning/semantics are not obvious from the composition of the meanings of the constituent words. For example, in *"show him the door"*, the individual constituent words have no connotation to the actual meaning of the phrase, which is to ask somebody to leave or step down from a position. Contrast this with *"show him the table"* which has no other interpretation than the literal one obtained from the composition of the constituent words. These kinds of collocations

are very common in human language due to prolific metaphorical and figurative usage. Handling these kinds of MWEs is crucial to robust natural language processing. *"Spill the beans", "run for office", "green card"* are other instances.

- **Statistical**: Some collocations are perfectly compositional semantically and well formed syntactically. But, these collocations tend to co-occur together more than what can be attributed to chance. For instance, it will be more likely to encounter *"traffic signal"* than "*traffic lamp*" or *"traffic lights"*, though they all mean the same thing. Such an idiosyncrasy arises because of the association of the collocation with a concept and the conventional acceptance or institutionalization of the collocation. Thus, statistical significance of the collocation is a criterion to judge these collocations as MWEs. Some instances of such collocations are *"good morning", "nail cutter"*.

- **Lexical**: Collocations not generally observed in the language, probably borrowed from other languages and institutionalized due to usage e.g. *"ad hoc"*.

- **Syntactic**: In these cases, certain collocations don't follow the rules of the conventional grammar, thus defying attempts at a successful parse and a hence a meaningful interpretation. *"By and large", "ad hoc", "wine and dine"* illustrate.

1.4 Characteristics of MWE

It is important to take note of some characteristics that MWEs exhibit due to their idiosyncratic behavior. The extent to which these characteristics are exhibited may vary, as we saw in the case of non-compositionality.

- **Institutionalization**: MWEs tend to have acceptance in conventional usage and therefore the collocations show statistical significance in their occurrence. This fact can be utilized to identify potential MWEs. Example: *"traffic signal", "prime minister", "nail cutter"*.

- **Paraphrasability**: Since an MWE may stand for a single concept, it might be possible to paraphrase the MWEs with a single word. For example, leave out means omit. Paraphrasability can thus be used as a rule of thumb test for MWEness.
- **Substitutability**: Due to their institutionalized usage, MWEs typically resist substitution of a constituent word by a similar word. For example, many thanks cannot be replaced by *"several thanks" or "many gratitudes"*.

- **Non-compositionality**: Non-compositionality of the MWE into its constituents is one of the key characteristics of MWEs. This is illustrated in examples like *"blow hot and cold", "spill the beans"*.

- **Syntactic fixedness**: Like any other language structure, MWE can undergo inflections (for tense, pluralization, etc.), insertions, passivization, etc. For example, *"traffic signal(s)", "promise(s/d/ing) (him/her/one) the moon"*. The semantic compositionality affects the amount of lexical variation that the collocations can undergo. For example, *"promising the moon"* would be resistant to any insertion, but a literal phrase like promise the pastry would allow for an insertion like promise the chocolate pastry.

1.5 Types of MWE

MWEs are categorized as:

1. **Compound Noun**
2. **Conjunct Verb**
3. **Compound Verb**
4. **Reduplication**

1.5.1 Compound Noun

: Nouns are parts of the language which provide the vocabulary to describe things and concepts. New concepts result in new nouns being added to the language. One of the ways in which new words are generated in language is by combining existing nouns to form new compound nouns. A compound noun is a noun consisting of more than one free morpheme. Examples:

1. चलचित्र प्रदर्शन (Hindi)
   - **Transliteration**: chalchitra pradarshan
   - **Gloss**: movie display
   - **Translation**: movie show

2. திராவிட மொழி (Tamil)
   - **Transliteration:** Thiraavida mozhi
   - **Gloss:** Dravidian language
   - **Translation:** Dravidian Language

Compound nouns can be names (i.e. proper nouns) or common nouns that have become institutionalized by usage or display semantic non-compositionality. Compound nouns consist of concatenated nouns, but they exhibit an internal hierarchical structure. Each compound noun can generally be expressed recursively in terms of head-modifier relationships, giving rise to bracketed structures

Example: (((संयुक्त अरब अमिरात) संवाददाता) जुलिया व्हीलर) (Hindi)

- **Transliteration:** sanyukt arab amirat samvaddata Julia wheeler
- **Gloss:** United Arab Emirates correspondent Julia wheeler
- **Translation:** United Arab Emirates correspondent Julia wheeler

This combination of noun to form compound noun is found among all the Indian languages.

Example: சிதம்பரம் தில்லை நடராஜர் கோயில் (Tamil)

- **Transliteration:** Chidambaram thillai nadarajar koil
- **Gloss:** chithambaram thillai natarajar temple
- **Translation:** chithambaram thillai natarajar koil

1.5.2 Conjunct Verb: Consider the word help in English, which appears in two different constructions:

1. He helped me with the housework.
2. He gave me help with the housework.

The equivalent in Hindi for give help is मदद (help) करना (to do). In Hindi, structures such as the one in 2 are possible while the one in 1 is not (i.e., direct noun to verb derivation). The question, therefore, is which word(s) is (are) to be stored in a lexical knowledge base (such as the Hindi Wordnet), मदद or मदद करना or both?

There are two possible solutions:

a. Enter मदद in the lexicon and then link it with the verb with which it co-occurs, i.e., करना

b. Enter मदद करना as a single entry in the lexicon and then specify its syntactic and semantic features separate from either मदद or करना individually specified.

The first approach is the simplest. Syntactically there is no problem in terms of the argument structure of the associated verb and its subject-verb agreement properties. But, the meaning of the conjoined term is not strictly compositional. Consider, for example, "छलांग मारना" 'dive'. "मारना" may mean either 'to beat' or 'to kill'. But neither meaning of the verb surfaces in "छलांग मारना".

The obvious problem with the second solution is one of proliferating lexical items, redundantly; we cannot store every single Noun + Verb combination in the lexicon. Generally, the noun is a true object and there is no need to store it as a lexical unit along with a co-occurring verb.

Thus it is necessary to separate true conjunct verbs from other similar looking Noun + Verb sequences. Consider the two combinations चाय (tea) लेना (to take) meaning 'to take tea' and जम्हाई (yawn) लेना (to take) meaning 'to yawn'. In the former case चाय (tea) is an overt object of the verb whereas in the latter, जम्हाई (yawn) is not.

Examples:

1. விசாரணை செய் (Tamil)
   - **Transliteration:** Vicaranai cey
   - **Gloss:** Enquiry do
   - **Translation:** interrogate

2. સલાહ આપવી (Gujarati)
   - **Transliteration:** salaah aapavI
   - **Gloss:** advice give
   - **Translation:** advise

3. ଚେଷ୍ଟା କରିବା (Odiya)
   - **Transliteration:** Chesta Karibaa
   - **Gloss:** Effort Doing
   - **Translation:** Striving

1.5.3 Compound Verb: This type of MWE is composed of Polar Verb and Vector Verb Combination. Verb + Verb constructions are also difficult to deal with and describe, since there are many serial verb sequences to be found in the language.

1. राम किताब पड़ रहा है |
   - **Transliteration**: raam kitab pad raha hai.
   - **Gloss:** Ram book read (V1) stay (V2) is
   - **Translation:** 'Ram is reading the book.'

2. राम ने किताब पड़ डाली |
    - **Transliteration**: raam ne kitab pad daali.
    - **Gloss:** Ram book read (V1) pour (V2)
    - **Translation:** ‘Ram (somehow) read (and finished) the book.’

रहा in first marks the progressive aspect, whereas डाली in second indicates the attitude (somehow finished). The Verb + Verb sequence in first should not be stored in the lexicon, whereas the one in second should probably be.

Examples:

3. எழுந்து வா (Tamil)
    - **Transliteration:** Ezunthu vaa
    - **Gloss:** Up come
    - **Translation:** Come up

4. କାନ୍ଦି ପକେଇଲା (Odiya)
    - **Transliteration:** Kaandi Pakeilaa
    - **Gloss:** Cry Throw
    - **Translation:** Crying

1.5.4 Reduplication: Reduplication is a linguistic phenomenon commonly found in languages across all the language families in India. Reduplication is a morphological process by which the root or stem of a word, or part of it, is repeated [wiki]. Example:

Although no standard classification exists, the following are the major classes of reduplications that commonly occur in Indian languages [Kea01]:

Onomatopoeic expressions: The constituent words imitate a sound, and the unit as a whole refers to that sound. Example:

1. టక్ టక్ (Telugu)
    - **Transliteration:** tak tak
    - **Gloss:** sound observed while walking
    - **Translation:** tuck tuck

2. মিট মিট (Bengali)
    - **Transliteration:** mit mit
    - **Gloss:** twinkle twinkle
    - **Translation:** twinkle twinkle

Complete Reduplication: The individual words are meaningful, and they are repeated. Example:

3. మెల్లి మెల్లి గ (Telugu)
    - **Transliteration: melli melli ga**
    - **Gloss: slowly**
    - **Translation: slowly slowly**

4. খাওয়া দাওয়া (Bengali)
    - **Transliteration:** khaoa daoa
    - **Gloss:** eat rock
    - **Translation:** eat

Partial Reduplication: In this, only one of the words is meaningful, while the other word is constructed by partial 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. Example:

5. पानी वाणी (Hindi)
    - **Transliteration:** pani vani
    - **Gloss:** water voice
    - **Translation:** water

The pair of words in reduplication acts as a single word syntactically and generally denotes a single concept. Reduplicate expressions are thus truly MWEs.

1.6 Survey

1.6.1 Different ongoing projects

As basis for helping to determine whether a given sequence of words is in fact an MWE (e.g. ad hoc vs the small boy) some of these works employ linguistic knowledge for the task, while others employ statistical methods or combine them with some kinds of linguistic information such as syntactic and semantic properties or automatic word alignment.

Statistical measures of association have been commonly used for this task, as they can be democratically applied to any language and MWE type. However, there is no consensus about which measure is best suited for identifying MWEs in general.

Though the theory of MWEs is underdeveloped and the importance of the problem is underappreciated in the field at large, there is ongoing work on MWEs within various projects that are developing large-scale, linguistically precise computational grammars, including the

1. ParGram Project at Xerox parc (http://www.parc.xerox.com/istl/groups/nltt/pargram/)
2. the XTAG Project at the University of Pennsylvania (http://www.cis.upenn.edu/˜xtag/)
3. work on Combinatory Categorial Grammar at Edinburgh University
4. the LinGO Project (a multi-site collaboration including CSLI's English Resource Grammar Project — http://lingo.stanford.edu)
5. the FrameNet Project http://www.icsi.berkeley.edu/˜framenet/), which is primarily developing large-scale lexical resources

All of these projects are currently engaged (to varying degrees) in linguistically informed investigations of MWEs.

As far as idiom identification is concerned, the work is classified into two kinds: one is for idiom types and the other is for idiom tokens. With the former, phrases that can be interpreted as idioms are found in text corpora, typically for lexicographers to compile idiom dictionaries. Previous studies have mostly focused on the idiom type identification [Bal03]. However, there has been a growing interest in idiom token identification recently [Kat06; Has06]. The idiom token identification is in an early stage of its development.

For Hindi, there have been limited investigations on MWE Extraction. [Ven05] considered N-V collocation extraction with certain syntactic and semantic features.

1. [Muk06] used POS projection from English to Hindi with corpus alignment for extracting complex predicates.

2. [Cha08] present a method for extracting Hindi V+V compound verbs using linguistic features.

3. [Sin09] use linguistic property of light verbs in extracting complex predicates using Hindi-English parallel corpus.

For machine translation, it has been noted that the issue of MWE identification and accurate interpretation from source to target language remained an unsolved problem for existing MT systems. This problem is more severe when MT systems are used to translate domain-specific texts, since they may include technical terminology as well as more general fixed expressions and idioms. Although some MT systems may employ a machine-readable bilingual dictionary of MWE, it is time-consuming and in-efficient to obtain this resource manually.

1.6.2 Extraction Approaches:

1.6.2.1 Linguistic Approach:

This approach totally depends on the characteristics of the language.

1.6.2.1.1 POS tagging:

POS tags helps in recognizing the patterns that the language is following. Assuming POS tagging to be fully correct, different patterns following the different types of MWE can be extracted.

Example:

- Noun + Noun: Compound Noun Bigrams
    - ప్రభుత్వ ఆస్పత్రి (Telugu)
        - **Transliteration** : prahutva aspatri
        - **Gloss**: government hospital
        - **Translation** : government hospital
    - **அண்ணா சாலை** (Tamil)
        - **Transliteration:** anna salai
        - **Gloss:** anna road
        - **Translation:** anna salai

- Verb + Verb: Compound Verb
    - எழுந்து வா (Tamil)
        - **Transliteration:** Ezunthu vaa
        - **Gloss:** up come
        - **Translation:** come up

    - ప్రభుత్వ ఆస్పత్రి (Telugu)
        - **Transliteration:** prahutva aspatri
        - **Gloss:** government hospital
        - **Translation:** government hospital

- Noun + Verb: Conjunct Verb
    - రాము చేతికి (Telugu)
        - **Transliteration :** ramu chetiki
        - **Gloss :** ramu hand
        - **Translation :** to Ramu's hand

    - ମନେ ପକେଇବା (Odiya)
        - **Transliteration:** Mane pakeibaa
        - **Gloss:** Mind Place
        - **Translation:** Remembering

1.6.2.1.2 Chunker

Chunker identifies non-recursive noun phrases (Noun-chunks) and verb groups.

- सबसे जादा पसंदीदा मुल्क
    - **Transliteration:** sabse jada pasandida muluk
    - **Gloss:** most favored nation
    - **Translation:** The most favored nation

The most favoured nation (has become a fixed expression; most favourite nation or maximally favoured nation or most favoured country not is use)

Possible chunker output

- सबसे_RB जादा_RB पसंदीदा_JJ मुल्क_NN

  Label: "पसंदीदा मुल्क" as N-chunk

Chunker can be user to restrict the Noun phrase boundaries so that false positives can be decreased while extracting the compound noun and compound verb candidates.

1.6.2.1.3 Using Language Resources

This includes the Conjunct Verb and the Compound Verb Extraction.

Conjunct Verb: Noun + Verbalizer Verb

Example: लेना (Take), मारना (Kill), लगाना (Set), पाना (Get), उठना (Rise)

In Indian languages the number of verbalizer verbs is quite a few. To be a conjunct verb, verbalizer verb has to be from these fixed resources. By restricting this, we can easily extract conjunct verb.

Compound Verb: Polar Verb + Vector Verb

Example: जाना (Go), देना (Give), गिराना (Drop), डालना (Cast), बैठना (Sit)

Again in Indian languages the vector verbs are quite a few and to be a compound verb, vector verb has to be from this vector verb list.

So the language resources help's in extracting the particular type of MWE.

1.6.2.2 Statistical Methods:

Statistical measures of association have been widely employed in the identification of MWEs. The idea behind their use is that they are an inexpensive language and type independent means of detecting recurrent patterns. As Firth famously said a word is characterized by the company it keeps and since we expect the component words of an MWE to occur frequently together, then these measures can give an indication of MWEness. In this way, if a group of words co-occurs with significantly high frequency when compared to the frequencies of the individual words, then they may form an MWE. Indeed, measures such as Pointwise Mutual Information (PMI), Mutual Information (MI), $\chi 2$, log-likelihood and others have been employed for this task, and some of

them seem to provide more accurate predictions of MWEness than others. In fact, in a comparison of some measures for the type-independent detection of MWEs, MI seemed to differentiate MWEs from non-MWEs, but the same was not true of $\chi 2$.

1.6.2.2.1 Pointwise Mutual Information (PMI)

Mutual Information is the amount of information shared by two random variables. [CH90] considered words in a collocation to be random variables and defined the Pointwise Mutual Information as a measure of association.

The PMI of a pair of outcomes x and y belonging to discrete random variables X and Y quantifies the discrepancy between the probability of their coincidence given their joint distribution and the probability of their coincidence given only their individual distributions, assuming independence.

For two-word expressions:

$$I_{bi} = \log_2 \frac{p(w_1 w_2)}{p(w_1) * p(w_2)}$$

Where, I_{bi} is the amount mutual information shared between w_1 and w_2,

$p(w_1 w_2)$ is the probability of words w_1 and w_2 being occurred together,

$p(w_1)$ and $p(w_2)$ is the probability of words w_1 and w_2 being occurred in the Bi-gram corpus independently.

Note that PMI is not true distributional MI, but is rather a ratio of the observed joint probability of collocation to that when the constituent words are considered to be independent. PMI is efficient to compute, once the probabilities are computed from corpus statistics. But, it is prone to highly overestimating the occurrence of rare events. This is because PMI does not incorporate the notion of support for the collocation. This allows the method to have a high recall. PMI can then serve as a good initial filter for identifying potential MWEs.

1.6.2.2.2 Pearson's Chi-Square Test

Pearson's chi-square test of independence can be used to test if the words in the collocation are independent of each other. Chi-square compares differences between the observed frequencies and the expected (under the assumption of independence) frequencies. Here, the null hypothesis

is that the words are independent of each other. From the corpus frequencies, a contingency table can be prepared as shown below for 2 words, w1 and w2.

$w_1 w_2$	$w_1 \sim w_2$
$\sim w_1 w_2$	$\sim w_1 \sim w_2$

Where, ~ denotes absence of the word. Hence, w1 ~ w2 is the frequency of collocation starting with w1, but not followed by w2.

Then, calculate the chi-square statistic as shown below:

$$X_2^2(x,y) = \sum_{i,j \in \{0,1\}} \frac{(f_{i,j} - e_{i,j})^2}{e_{i,j}}$$

$$f_{0,0=f(x,y)}, f_{0,1} = f(x,\neg y) = \sum_{v \neq y} f(x,v)$$

Where f(I,j) is the observed frequency

e(I,j) is the expected frequency in each cell when w1 and w2 occur together by chance.

Expected frequency for each cell is equal to (row total * column total) / grand total. The higher the value of the chi-square statistic, the stronger is the association between the words. A cut off according to the desired level of significance can be fixed. However, in case of low frequencies, the chi-square distribution does not hold and assumptions of normal distribution are not satisfied. Thus, the method demands large corpus to give good results.

For three words, the Chi-square formula can be extended as:

$$X_3^2(x,y,z) = \sum_{i,j,k \in \{0,1\}} \frac{(f_{i,j,k} - e_{i,j,k})^2}{e_{i,j,k}}$$

1.6.2.2.3 Likelihood Ratio Test

A measure suggested by [Dun93] that indicates how much more likely the co-occurrence is than mere coincidence. More general test of significance compared to the $\chi 2$ test and makes no assumptions of approximation to the normal distribution. Hence, the test performs better for smaller corpus and ones with different distributional characteristics. It expresses how many times more likely the data are under one model than the other. The likelihood ratio is the ratio of the likelihood of observations given the null-hypothesis defining subspace to that of the entire parameter space. The likelihood ratio is given as:

$$\lambda = \frac{\max_{\omega \in \Omega_o} H(\omega; k)}{\max_{\omega \in \Omega} H(\omega; k)}$$

Where, ω is considered to be a point in the parameter space Ω, and k a point in the space of observations K. The quantity −2 log λ is used as the measure for the test since it is asymptotically $\chi 2$ distributed with degrees of freedom equal to the difference in dimension between Ω and Ω0.

[Dun93] applied the likelihood-ratio test to the problem of detecting statistically significant collocations. For this, model the corpus frequencies of the word in the collocation as binomial distribution. For a collocation w1w2, the null hypothesis is that P(w2|w1) = P(w2| w1). Now the likelihood of the parameters given the corpus frequencies of w2 and assuming a binomial distribution of the text is:

$$H(p_1, p_2; k_1, n_1, k_2, n_2) = p_1^{k_1}(1-p_1)^{n_1-k_1}\, p_2^{k_2}(1-p_2)^{n_2-k_2}$$

Where,

$p_1 = P(w_2 \mid w_1), p_2 = P(w_2 \mid\sim w_1)$

$n_1 = c_1, k_1 = c_{12}$

$n_2 = n - c_1, k_2 = c_2 - c_{12}$

$c_1, c_2 =$ Corpus frequencies of w_1, w_2

$c_{12} =$ Corpus frequencies of $w_1 w_2$

N = total number of words in the corpus

For the alternate hypothesis, the MLE estimates of p1, p2 are,

$$p_1 = \frac{k_1}{n_1}$$

and

$$p_2 = \frac{k2}{n_1}$$

$$p = \frac{k_1 + k_2}{n_1 + n_2}$$

For the null hypothesis, we have p1 = p2 = p.

1.6.3 Extraction method mentioned by different researchers

1.6.3.1 Automatic Extraction of Arabic MWE

It used three approaches to the extraction of Arabic MWE's [Att10]

1. First make use of the cross-lingual correspondence asymmetry, or many-to-one relations between the titles in the Arabic Wikipedia (AWK) and the corresponding titles in other languages to harvest MWEs
2. Second assumed that automatic translation of MWEs collected from Prinston wordnet (PWN) into Arabic are high likelihood MWE candidates that need to be automatically checked and validated.
3. Third, try to detect MWEs in a large raw corpus relying on statistical measures and POS-annotation filtering.

It considers many-to-one correspondence relationships (an MWE in one language has a single-word translation in another language) as empirical evidence for the detection of MWEs. Here our candidate MWE's are the AWK titles that are made up of more than one word. For each of them we check whether there exists a many-to-one correspondence relation for this title in other languages.

1.6.3.2 Domain-Specific Multi-Word Terms from Different Types of Corpora

Firstly, a shortlist of well formed and relevant candidate MWTs is extracted from a given target corpus and secondly a contrastive method is applied against the selected MWTs only. In fact, in the first stage, candidate MWTs are searched for in an automatically POS-tagged and lemmatized text and they are then weighted with the C-NC Value method. In the second stage, the list of MWTs extracted is revised and re-ranked with a contrastive score, based on the distribution of terms across corpora of different domains; the Contrastive Selection of multi-word terms [Bon11].

1.6.3.3 Mining Complex Predicates In Hindi Using a Parallel Hindi-English Corpus

A CP is hypothesized by detecting absence of the conventional meaning of the light verb in the aligned English sentence [Sin09]. This simple strategy exploits the fact that CP is a multiword expression with a meaning that is distinct from the meaning of the light verb.

The steps involved are as follows:

1. Align the sentences of Hindi-English corpus.

2. Create a list of Hindi light verbs and their common English meanings as a simple verb.

3. For each Hindi light verb, generate all the morphological forms.

4. For each English meaning of the light verb generate all the morphological forms.

5. For each Hindi-English aligned sentence, execute the following steps:

a. For each light verb of Hindi execute the following steps:

i) Search for a Hindi light verb (LV) and its morphological derivatives in the Hindi sentence and mark its position in the sentence (K)

ii) If the LV or its morphological derivative is found, then search for the equivalent English meanings for any of the morphological forms in the corresponding aligned English sentence

iii) If no match is found, then scan the words in the Hindi sentence to the left of the Kth position (as identified in step (i)); else if a match is found, then exit {i.e. go to step (a)}

iv) If the scanned word is a 'stop word' then ignores it and continues scanning

v) Stop the scan when it is not a 'stopword' and collect the Hindi word (W);

vi) If W is an 'exit word' then exit {i.e. go to step (a)}, else the identified CP is W+LV.

Observation: - It is obvious that this empirical method of mining CPs will fail whenever the Hindi light verb maps on to its core meaning in English

1.6.3.4 Statistically-Driven Alignment-Based Multiword Expression Identification for Technical Domains

In this the word alignment is the basis for the MWE extraction process. Given this context, it uses the alignment techniques for identifying MWEs, looking at sequences detected by the aligner as containing more than one word, which form the MWE candidates [Cas09]. As a result, sequences of two or more consecutive source words are treated as MWE candidates regardless of whether they are translated as one or more target words.

Observation: The alignment-based method generates a targeted precision-oriented list of MWE candidates, while the statistical methods produce recall-oriented results at the expense of precision. Therefore, the combination of these methods can produce a set of MWE candidates.

1.6.3.5 Stepwise Mining of Multi-Word Expressions in Hindi

Here the process of identification is semi-automatic. The automatic process generates the probable MWEs and then filtered manually [Sin11].

The process starts with sentence boundary identification followed by POS tagging. Then morphological analysis to perform the stemming is done. Acronym and abbreviation with dots are identified afterwards.

1. Hindi chunker and verb-phrase form separation
2. identification of replicating class
3. identification of doublet class
4. complex predicates and compound verb identification
5. identification of acronym (with no dots)
6. identification of named-entities

In this many of these characteristics are generic in nature in the sense that it is not based on any statistical inference but it is the linguistic property that helps in MWE extraction. For example, all replicating words irrespective of their POS, all doublets with plural-singular form combinations etc are all strong candidates for MWEs in Hindi irrespective of whether these have earlier been encountered in the corpus or not. This means that even the low frequency MWEs can be captured.

The statistical approach will anyway be needed to mine other types of MWEs and discover new and institutionalized MWEs (mostly domain specific) that keep getting added.

1.6.3.6 Identification and treatment of MWE in IR

It used some standard statistical measures such as mutual information, point-wise mutual information, chi-square to extract MWEs from a collection of documents [Cos11] (i.e. we consider the collection of documents indexed by the IR system as our corpus).

In order to determine the impact of the quality of the dictionary used in the performance of the IR system, we examined several different sources of MWE of varying quality. The dictionaries containing the MWEs to be inserted into the corpus as a single term are created by a number of techniques involving automatic and manual extraction. Below we de-scribe how these MWE dictionaries were created.

1. **Compound Nouns (CN)**: - For the creation of this dictionary, we extracted all bigrams contained in the corpus. Since the number of available bigrams was very large (99,744,811 bi-grams) the next step was the selection of bigrams that had the highest frequency in the text, so we chose candidates occurring at least ten times in the whole

corpus. As a result, the first list of MWEs was composed by 15,001 bigrams, called D1.

2. **Best Compound Nouns**: - After D1, we refined the list with the use of statistical methods. The methods used were the mutual information and chi-square. The first 7,500 entries composed the second dictionary, called D2.

3. **Worst Compound Nouns**: - This dictionary was created from bigrams that have between five and nine occurrences and are more likely to co-occur by chance. It was created in order to evaluate whether the choice of the potentially more noisy MWEs entailed a negative effect in the results of IR, compared to the previous dictionaries. The third dictionary, with 17,328 bi-grams, is called D3.

4. **Gold Standard:** - This was created from a sub-list of the Cambridge International Dictionary of English, containing MWEs. Since this list contains all types of MWEs, it was necessary to further filter these to obtain compound nouns only, using Postag information and formed 568 MWEs D4 dictionary will be called D4.

5. **Manual**: - For comparative purposes, we also created two dictionaries by manually evaluating the text of the 310 query topics. The first dictionary contained all bigrams which would achieve a different meaning if the words were concatenated (e.g. space shuttle). This dictionary was called D5 and contains 254 expressions. The other one was created by a specialist (linguist) who classified as true or false a list of MWE candidates from the query topics. The linguist selection of MWEs formed D6 with 178 bigrams.

The IR system evaluation is based on recall and precision.

Precision is the portion of the retrieved documents which is actually relevant to the query.

Recall is the fraction of the relevant documents which is retrieved by the IR System.

$$Precision(P) = \frac{\#Relevant \cap \#Retrieved}{\#Retrieved}$$

$$Recall(R) = \frac{\#Relevant \cap \#Retrieved}{\#Relevant}$$

It was proved that manual extraction of MWE improves the IR as Precision is increased.

1.7 Summary

In this chapter, we presented a literature survey on MWEs. In particular, we started with the necessary and sufficient condition for a group of words to be a MWE. In succession, we presented the characteristics of a MWE. Then, we presented different types of MWEs that we are concentrating on, with examples from different Indian languages.

Then we move to the different measures to extract MWEs from the corpus, concentrating on linguistic and statistical measures. Finally we end with the extracting methods undertaken by different researchers to extract MWEs for their particular language.

References

[Bal03] T. Baldwin, C. Bannard, T. Tanaka, and D. Widdow. *"An empirical model of multiword expressions decomposability"* In Proc. of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment, 2003

[Sag02] Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger *"Multiword Expressions: A Pain in the Neck for NLP"* @ CICLing 2002

[Ren09] Zhixiang Ren, Yajuan Lu, Jie Cao, Qun Liu1 and Yun Huang *"Improving Statistical Machine Translation UsingDomain Bilingual Multiword Expressions"* Proceedings of the 2009 Workshop on Multiword Expressions, ACL-IJCNLP 2009; Suntec, Singapore, 6 August 2009. @ 2009 ACL and AFNLP

[Ano08] Anoop, *"Multiword Expressions",* Master of Technology, Thesis

[Kmw11] Multiword Expression Workshop, Kashmir 2011

[Deb08] Debashi, *"A study of comples predicates of Hindi"*, Phd, CSE, IIT Bombay 2008

[PB12] Dr. Pushpak Bhattacharyya, *"Multiwords Processing in Indian Languages"*, WILDRE workshop, LREC, Istanbul, 2012

[Deb07] Debasri Chakrabarty, Vaijayanthi Sarma and Pushpak Bhattacharyya, *"Complex Predicates in Indian Language Wordnets,"* Lexical Resources and Evaluation Journal, 40 (3-4), 2007

[Deb06] Debasri Chakrabarti, Vaijayanthi Sarma and Pushpak Bhattacharyya, *"Hindi Verb Knowledge Base and Noun Incorporation in Hindi"*, 3rd Global Wordnet Conference (GWC 06), Jeju Island, Korea, January, 2006

[Sin09] R. Mahesh K. Sinha *"Mining Complex Predicates In Hindi Using A Parallel Hindi-English Corpus"* Proceedings of the 2009 Workshop on Multiword Expressions, ACL-IJCNLP 2009, Suntec, Singapore, 6 August 2009. @ 2009 ACL and AFNLP

[Sin11] R. Mahesh K. Sinha *"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), Portland, Oregon, USA, 23 June 2011. @ 2011 Association for Computational Linguistics

[Att10] Mohammed Attia, Antonio Toral, Lamia Tounsi, Pavel Pecina and Josef van Genabith *"Automatic Extraction of Arabic Multiword Expressions"* Proceedings of the Workshop on Multiword Expressions: from Theory to Applications (MWE 2010),Beijing, August 2010

[Bon11] Petter Haugereid and Francis *Bond "Extracting Transfer Rules for Multiword Expressions from Parallel Corpora"* Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE 2011), Portland, Oregon, USA, 23 June 2011, @2011 Association for Computational Linguistics

[Cos11] Otavio Costa Acosta, Aline Villavicencio and Viviane P. Moreira *"Identification and Treatment of Multiword Expressions applied toInformation Retrieval"* Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE 2011), Portland, Oregon, USA, 23 June 2011, @2011 Association for Computational Linguistics

[Cas09] MedHelena de Medeiros Caseli, Aline Villavicencio, Andr´e Machado and Maria Jos´ e Finatto *"Statistically-Driven Alignment-Based Multiword ExpressionIdenti?cation for Technical Domains"* Proceedings of the 2009 Workshop on Multiword Expressions, ACL-IJCNLP 2009, Suntec, Singapore, 6 August 2009. @ 2009 ACL and AFNLP