MULTILINGUAL MULTIWORD EXPRESSIONS

Literature Survey

by

Lahari Poddar

Under the guidance of

Prof. Pushpak Bhattacharyya

Department of Computer Science and Engineering

Indian Institute of Technology, Bombay

Mumbai

Abstract

Multiword Expressions are idiosyncratic word usages of a language which often have non-compositional meaning. The knowledge of multiword expressions is necessary for many NLP tasks like, machine translation, natural language generation, named entity recognition, sentiment analysis etc. In order for other NLP applications to benefit from the knowledge of multiword expressions, they need to be identified and stored in lexical knowledgebase. There have been many approaches towards automatic extraction of multiword expressions. In this document we present some of the definitions of multiword expressions, their classifications and different approaches towards their automatic extraction.

Table of Contents

Table Of Figures

Chapter 1

Background

In this section we will describe the formal definition Multiword Expression coined by different researchers. The different types and characteristics possessed by such expressions are elaborated.

Various researchers have defined multiword expressions differently during their research. We'll present some of the definitions here and it can be observed that all of them primarily refer to a single central concept.

- A collocation is an expression consisting of two or more words that correspond to some conventional way of saying things.[**8**]
- Idiosyncratic interpretations that cross word boundaries (or spaces) [**7**]
- Recurrent combinations of words that co-occur more frequently than chance, often with non-compositional meaning[**9**]
- A pair of words is considered to be a collocation if one of the words significantly prefers a particular lexical realization of the concept the other represents[**5**]

1.1. Features of Multiword Expressions

[**7**]

There are certain features that a group of words must have in order to be treated as collocation. The principal features are:

- **Non-Compositionality:** The meaning of a complete multiword expression can't completely be determined from the meaning of its constituent words.

  The meaning of the expression might be completely different from its constituents (the idiom *kick the bucket* means *to die*) or there might be some added element or inline meaning to it that cannot be predicted from the parts(the phrase *back to square one* means to reach back to the place from where one had started).

- **Non-Substitutability:** The components of a multiword expression cannot be substituted by one of its synonyms without distorting the meaning of the expression even though they refer to the same concept.

  For example, in the expression *bread and butter* the component words cannot be replaced by their synonym keeping the meaning(to earn one's daily living) intact.

- **Non-Modifiability:** Many collocations cannot be freely modified by grammatical transformations (like, change of tense, change in number, addition of adjective etc.). These collocations are frozen expressions, they cannot be modified in any way.

  For example, the idiom *let the cat out of the bag* cannot be modified to **let the big cat out of the bag* or something similar.

1.2. Types of MWEs

Collocations or Multiword Expressions can be classified into different classes according to their lexical and semantic characteristics. The classification as described in [**7**] is given below.

1) **Lexicalized Phrases:** This type of phrases have some form of idiosyncratic or added meaning to the structure. They are either syntactically idiosyncratic or semantically non-decomposable. Lexicalized phrases can be classified into 3 parts.

a) **Fixed Expressions:** This is the class of expressions that defy the general conventions of grammar and compositional interpretations. These expressions are completely frozen and do not undergo any modifications at all.

   Example: in short, of course, ad hoc, by and large

b) **Semi-Fixed Expressions:** This type of expressions have restrictions on word order and the structure of the phrase but they might undergo some form of lexical variations. Semi-Fixed expressions can be further classified into 3 subtypes:

(1) **Non-Decomposable Idioms:** Depending on their semantic composition, idioms can be classified into two types: Decomposable and Non-Decomposable.

For decomposable idioms each component of the idiom can be assigned a meaning related to the overall meaning of the expression. For the idiom *spill the beans*, 'spill' can be assigned the sense of 'reveal' and 'beans' can denote the sense of 'secret'. But in case of Non-Decomposable idioms no such analysis is possible.

For the idiom *kick the bucket* none of its components can be assigned a sense such that the overall idiom means 'to die'.

It is these Non-Decomposable idioms which are semi-fixed. Due to their opaque meaning they do not undergo any syntactic variations but might allow some minor lexical modification (*kick the bucket* -> *kicked the bucket*).

(2) **Compound Nominals:** Compound nominals also do not undergo syntactic modifications but allow lexical inflections for number i.e. they can be changed to their singular or plural form.

Example: car park, part of speech, railway station

(3) **Named Entities:** These are syntactically highly idiosyncratic. These entities are formed based on generally a place or a person.

Example: the cricket team names in IPL are formed based on the region. In a proper context the team names are often mentioned without the name of the place, like '(Kolkata) Knight Riders', 'Royal Challengers (Bangalore)' etc. When the team name occurs as a modifier in some compound noun a modifier is added ('the Kolkata Knight Riders player...')

c) **Syntactically-Flexible Expressions:** As opposed to the strict word order constraint of Semi-Fixed expressions, Syntactically-Flexible expressions allow a wide variety of syntactic variations. They can be classified into 3 types:

   (1) **Verb-Particle Construction:** Verb-Particle constructions or phrasal verbs consist of a main verb and a particle. Transitive verb-particle constructions are a good example of non adjacent collocations as they can take an NP argument in between (like, *call him up*).

   Example: call off, write up, eat up etc.

   (2) **Decomposable idioms:** Decomposable idioms are syntactically flexible and behave like semantically linked parts. But it's difficult to predict exactly what type of syntactic variations they undergo.

   Example: spill the beans, let the cat out of the bag

   (3) **Light-Verb Constructions:** Verbs with little semantic content (make, take, do) are called light verbs as they can form highly idiosyncratic constructions with some nouns.

   Example: *make a decision , do a favor, take a picture etc* are light-verb constructions as there is no particular reason why *do me a favor* should be preferred over **make me a favor* and so on.

2) **Institutionalized Phrases:** These phrases are completely compositional (both syntactically and semantically) but are statistically idiosyncratic. These are just fixed terms which do not have any alternate representations.

Example: traffic light, fresh air, many thanks, strong coffee etc.

Chapter 2

MWE Extraction Approaches

2.1. Approaches by various researchers

In this section we are going to present a survey of the different approaches tried out by different researchers over the years in order to extract multiword expressions from a text. The methods vary widely from one another. Some of them have taken a Linguistic approach, some have used statistical techniques and some have taken help of the open source resources available to us to solve the problem.

2.1.1 Rule Based Approaches

There have been quite a few approaches which try to detect multiwords by leveraging the rules forming them in the first place.

2.1.1.1. Identification of Reduplication in Bengali [1]

Reduplication is a subtype of Multiword Expressions and a method for identifying reduplications and then classifying them has been reported by the authors. Reduplications have been categorized into 2 levels, namely **Expression Level** and **Sense Level.** They can be further subcategorized as:

Expression Level:

a) Onomatopoeic expressions: The constituent words imitate a sound or a sound of an action. Generally in this case the words are repeated twice with the same ‘matra’.

- ঝম ঝম (Bengali)

  Transliteration: jham jham

  Translation: the sound of rain

- টপ টপ (Bengali)

  Transliteration: top top

  Translation: the sound of dropping water

b) Complete Reduplication: The constituent words are meaningful and they are repeated to convey some particular sense.

- চলতে চলতে (Bengali)

  Transliteration: chalte chalte

  Gloss: walking walking

  Translation: while walking

- বার বার (Bengali)

  Transliteration: bar bar

  Gloss: time time

  Gloss: time and again/ repeatedly

c) Partial Reduplication: In partial reduplication generally three cases are possible

(i) change of the first vowel or the matra attached with first consonant

(ii) change of consonant itself in first position

(iii)change of both matra and consonant

- বোকা সোকা (Bengali)

  Transliteration: boka soka

  Translation: Foolish

- চাল চুলো (Bengali)

  Transliteration: chal chulo

  Translation: belongings

d) Semantic Reduplication: A dictionary based approach was followed to identify consecutive occurrences of synonyms and antonyms.

- দিন রাত (Bengali)

  Transliteration: *din-raat*

  Gloss: day and night

  Translation: round the clock/ all the time

- পাপ পুণ্য (Bengali)

  Transliteration: paap-punyo

  Gloss: sin and virtue

Sense Level Classification:

a) Sense of repetition:

- রোজ রোজ (Bengali)

  Transliteration: roj roj

  Gloss: day day

  Translation: everyday

- বছর বছর (Bengali)

  Transliteration: bachor bachor

  Gloss: year year

  Translation: every year

b) Sense of plurality:

- ছোটো ছোটো (Bengali)

  Transliteration: choto choto

  Gloss: small small

  Translation: small

c) Sense of Emphatic :

- সুন্দর সুন্দর (Bengali)

  Transliteration: sundor sundor

  Gloss: beautiful beautiful

  Translation: beautiful

- লাল লাল (Bengali)

  Transliteration: laal laal

  Gloss: red red

  Translation: red

d) Sense of completion :

- খেয়ে দেয়ে (Bengali)

  Transliteration: kheye deye

Translation: after finishing meal

e) Sense of incompleteness :

- বলতে বলতে (Bengali)

  Transliteration: bolte bolte
  Gloss: talking talking
  Translation: while talking

- চলতে চলতে (Bengali)

  Transliteration: cholte cholte
  Gloss: walking walking
  Translation: while walking

Some collected articles of Rabindranath Tagore have been used as a corpus. The system developed by them reportedly achieved 92% precision and a recall of 91%. There exists some combination of words which have a semantic relationship between them but are not exactly synonyms or antonyms of each other (for eg: '*slow and steady*'). The system was unable to detect such type of reduplications using only a dictionary.

2.1.1.2. Detecting noun compounds and light verb constructions [11]

The authors have described some rule based methods to detect noun compounds and light verb constructions in running texts.

Noun compounds are productive, i.e. new nominal compounds are being formed in language use all the time, which yields that they cannot be listed exhaustively in a dictionary (eg. World wide Web, Multiword Expressions). Whereas Light verb constructions are semi-productive, i.e. new light verb constructions might enter the language following some patterns (e.g. 'give a Skype call' on the basis of 'give a call').

Light Verb compounds are syntactically very flexible. They can manifest in various forms: the verb can be inflected, the noun can occur in its plural form and the noun can be modified. The nominal and the verbal component may not even be contiguous (eg. ‘He gave me a very helpful advice’).

Methods of MWE identification

1. Lowercase n-grams which occurred as links were collected from Wikipedia articles and the list was automatically filtered in order to delete non-English terms, named entities and non-nominal compounds etc.
2. Match: A noun compound is taken into consideration if it belongs to the list or it is composed of two or more noun compounds from the list.
3. POS rules: A noun compound candidate was marked if it occurred in the list and its POS-tag sequence matched one of the predefined patterns.
4. Suffix rule: The ‘Suffix’ method exploited the fact that many nominal components in light verb constructions are derived from verbs. Thus, in this case only constructions that contained nouns ending in certain derivational suffixes were allowed and for nominal compounds the last noun had to have this ending.
5. Most frequent method: This routine relied on the fact that the most common verbs function typically as light verbs (e.g. do, make, take, have etc.). Thus, the 15 most frequent verbs typical of light verb constructions were collected and constructions where the stem of the verbal component was among those of the most frequent ones were accepted.
6. Stem rule: In the case of light verb constructions, the nominal component is typically one that is derived from a verbal stem (make a decision) or coincides with a verb (have a walk).
7. Syntactic Information: Generally the syntactic relation between the verb and the nominal component in a light verb construction is verb-object.

2.1.2 Statistical Methods for Multiwords Extraction *[8], [3]*

A number of basic statistical methods can be used for extracting collocations from a given corpus. The corpus used for carrying out the experiments was a collection of The New York

Times newswire for four months that consisted of 14 million words. Let us look at these methods and their corresponding applications for extracting multiwords.

2.1.2.3. Frequency

This is the simplest method for extracting collocations as it just retrieves the most frequent bigrams in the corpora. But this naive approach produced a lot of insignificant bigrams which are very frequent (*of-the,in-the* etc.) This difficulty can be easily overcome by applying a simple heuristic - pass the candidate phrases through a POS tagger and take only those combinations into considerations that have the probability of being phrases. The POStag structures that were taken into account were: AN, NN, AAN, ANN, NAN, NNN, NPN.

As we can see in Figure 2.1 even though it is a very simple method the results produced by this method was quite impressive.

$C(w^1 w^2)$	W^1	W^2	Tag Pattern
11487	New	York	AN
7261	United	States	AN
5412	Los	Angeles	NN
3301	Last	Year	AN
3191	Saudi	Arabia	NN
2699	Last	Week	AN
2514	Vice	President	AN

Figure 2.1: Finding Collocations: Frequency Method [**8**]

2.1.2.4. Mean And Variance

The above method for frequency works only for fixed phrases but there are words which stand in a flexible or variable length relationship length from one another. These are the words that appear with each other very frequently but can take any number of words in between.

Example: *knock...door,*this is a proper collocation even though there might be any number of words between *knock* and *door* depending on the structure of the sentence but *knock* is generally the verb associated with *door*.

In this method we calculate the mean and variance of the distance between two words. The variance is defined as:

$$s^2 = \frac{\sum_{i=1}^{n}(d_i - \bar{d})^2}{n-1}$$

Where 'n' is the number of times the two words co-occur, d_i is the offset for co-occurrence 'i' , and $\bar{d}$ is the sample mean of the offsets. If the offsets are same for most occurences the variance will be low and if the offsets differ highly for the occurences then the variance will be very high.

s	d	Count	Word1	Word2
0.43	0.97	11657	New	York
0.48	1.83	24	Previous	Games
0.15	2.98	46	Minus	Points
4.03	0.44	36	Editorial	Atlanta
4.03	0.00	78	Ring	New
3.96	0.19	119	Point	Hundredth
1.07	1.45	80	Strong	Support
1.13	2.57	7	Powerful	Organizations
1.01	2.00	112	Rechard	Nixon

Figure 2.2: Finding Collocations: Mean and Variance[**8**]

2.1.2.5. Hypothesis Testing

The basic problem that we want to solve for collocation extraction is determining whether two words occur together more often than chance. Hypothesis testing is a classic approach in statistics for this type of problems. A null hypothesis H_0 is formed for this stating that the two words occur merely by chance. Now the probability of occurence of the two words given that H_0 is true is calculated,and then depending on this value of probability the null hypothesis is accepted or rejected.

2.1.2.5.1. The t-test

The t-test looks at the mean and variance of a sample, where the null hypothesis is that the sample is drawn from a distribution with mean μ . The test computes the difference between the observed and expected means, scaled by the variance of the data, and tells us how likely it is to get a sample of that mean and variance (or a more extreme mean and variance) assuming that the sample follows normal distribution.

$$t = \frac{\bar{x} - \mu}{\sqrt{s^2/n}}$$

Where s^2 is the sample variance, N is the sample size, μ is the mean of the distribution. If the t statistic is large enough we can reject the null hypothesis stating that the words are associated. For example,in the corpus, *new* occurs 15,828 times, *companies* 4,675 times, and there are 14,307,668 tokens overall.

new companies occurs 8 times among the 14,307,668 bigrams

$$H_0 : P(newcompanies) = P(new)P(companies)$$

$$= \frac{15828}{14307668} * \frac{4675}{14307668}$$

$$\approx 3.675 * 10^{-7}$$

The observed frequency of occurence of *new companies* is 8 in the corpus.

$$\bar{x} = \frac{8}{14307668}$$

Now applying the t-test:

$$t = \frac{\bar{x} - \mu}{\sqrt{s^2/n}}$$

$$\approx \frac{5.591 * 10^{-7} - 3.675 * 10^{-7}}{\sqrt{\frac{5.591 * 10^{-7}}{14307668}}}$$

$$\approx .999932$$

This t value of 0.999932 is not larger than 2.576, the critical value for $\alpha = 0.005$. So we cannot reject the null hypothesis that new and companies occur independently and do not form a collocation.

2.1.2.5.2. *Hypothesis Testing of Differences*

A variation of the basic t-test can be used to find words whose co-occurences best distinguish the subtle difference between two near synonyms. Figure 2.3 shows the words that occur significantly more often with *powerful* (the first ten words) and *strong* (the last ten words).

The formula of the basic t-test is modified as

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

The application for this form of the t test is lexicography. Such data is useful to a lexicographer wanting to write precise dictionary entries that bring out the difference between *strong* and *powerful*.

t	C(w)	C(strong w)	C(powerful w)	Word
3.1622	933	0	10	Computers
2.8284	2337	0	8	Computer
2.4494	289	0	6	Symbol
7.0710	3685	50	0	Support
6.3257	3616	58	7	enough
4.6904	986	22	0	Safety

Figure 2.3: Hypothesis Testing Of Differences [**8**]

2.1.2.5.3. *Pearson's Chi-Square Test*

The t-test assumes that the probabilities of occurence are approximately normally distributed, which is not true in general. It is an alternative test that doesnot depend on the normality assumption. The essence of the test is to compare the observed frequencies with the frequencies expected for independence. If the difference between observed and expected frequencies is large, then we can reject the null hypothesis of independence.

	$W_1 =$ new	$W_1 \neq$ new
$W_2 =$ companies	8 (new companies)	4667 (eg: old companies)
$W_2 \neq$ companies	15820 (eg: new machines)	14287181 (eg: old machines)

Figure 2.4: Pearson's Chi-Square Test [**8**]

Figure 2.4 shows the observed frquency values for *new* and *companies*. On these values the test is applied. If the difference between observed and expected frequencies is large, then we can reject the null hypothesis of independence.

The χ^2 statistic sums the differences between observed and expected frequencies,scaled by the magnitude of the expected values:

$$\chi^2 = \sum_{i,j} \frac{O_{ij} - E_{ij}{}^{2}}{E_{ij}}$$

Where i ranges over rows of the table, j ranges over columns, O_{ij} is the observed value for cell and E_{ij} is the expected value.

2.1.2.5.4. Likelihood Ratio

This test produces simply a number that tells us how much more likely one hypothesis is than the other. So it more interpretable than any other forms of hypothesis testing. Moreover, likelihood ratios are more appropriate for sparse data than the Chi-Square test.

For applying likelihood testing, let us consider the following two hypothesis:

$$Hypothesis1: P(w^2 \mid w^1) = p = P(w^2 \mid \neg w^1)$$

$$Hypothesis2: P(w^2 \mid w^1) = p_1 \neq p_2 = P(w^2 \mid \neg w^1)$$

Hypothesis1 is a formalization of independence whereas Hypothesis2 is a formalization of dependence. We calculate the log likelihood ratio as:

$$log_2(\lambda) = log_2 \frac{L(H_1)}{L(H_2)}$$

-2log(λ)	C(w^1)	C(w^2)	C(w^1w^2)	W^1	W^2
-1291.42	12593	932	150	Most	Powerful
99.31	379	932	10	Politically	Powerful
82.96	932	934	10	Powerful	Computers
80.39	932	3424	13	Powerful	Force

-2log(λ)	C(w^1)	C(w^2)	C(w^1w^2)	W^1	W^2
57.27	932	291	6	Powerful	Symbol
51.66	932	40	4	Powerful	Lobbies
51.52	171	932	43	Economically	Powerful
51.05	932	43	4	Powerful	Magnet
50.83	4458	932	10	Less	powerful

Figure 2.5: Likelihood Ratio[**8**]

The Figure 2.5 shows the top bigrams consisting of *powerful* when ranked according to likelihood ratio.

2.1.2.5.5. Relative Frequency Ratio

Ratios of Relative Frequencies between different corpora can be used to discover collocations that are characteristic of a corpus when compared to the other.

Ratio	1990	1989	W^1	W^2
0.0241	2	68	Karim	Obeid
0.0372	2	44	East	Berliners
0.0372	2	44	Miss	Manners
0.0399	2	41	17	Earthquake
0.0409	2	40	HUD	officials

Figure 2.6: Relative Frequency Ratio [**8**]

This approach is most useful for the discovery of subject-specific collocations. It can be used to compare a general text with a domain-specific text.

2.1.2.6. Mutual Information

This is a method derived from information theory measures where we can find out how much information does the presence of one word gives about another word in the context. Informally, it is a measure of the company that a word keeps.

Mutual information (for two words, x and y) can be defined as:

$$I(x,y) = log_2 \frac{P(x^{'}y^{'})}{P(x^{'})P(y^{'})}$$

$$= log_2 \frac{P(x^{'} \mid y^{'})}{P(x^{'})}$$

$$= log_2 \frac{P(y^{'} \mid x^{'})}{P(y^{'})}$$

None of the statistical methods work very well for sparse data but Mutual Information works particularly badly in sparse environments because of the structure of the equation.

For perfect dependence (i.e. whenever they occur,they occur together):

$$I(x,y) = log_2 \frac{P(x^{'}y^{'})}{P(x^{'})P(y^{'})}$$

$$= log_2 \frac{P(x^{'})}{P(x^{'})P(y^{'})}$$

$$= log_2 \frac{1}{P(y^{'})}$$

The value of mutual information score gets inversely proportional to the frequency value of the bigram. So the bigrams that are rare in the corpus gets an artificially inflated mutual information score.

For perfect independence (i.e. their occurence together is completely by chance):

$$I(x,y) = log_2 \frac{P(x^{'}y^{'})}{P(x^{'})P(y^{'})}$$

$$= log_2 \frac{P(x^{'})P(y^{'})}{P(x^{'})P(y^{'})}$$

$$= log_2 1$$

$$= 0$$

It can be inferred that Mutual Information is a good measure of independence between two words but it is a bad measure for deciding the dependence between a bigram.

2.1.2.7. *Comparative Analysis*

We would like to present a comparative analysis in this section highlighting which method will be useful for what type of collocation.

- Frequency based method is simple and easy to implement hence it will be very useful for lightweight computations (Eg: Information Retrieval through search engines).
- Mean and Variance method can be used for terminological extraction and Natural Language Generation as it works well for variable length phrases.
- t-Test is most useful for ranking collocations and not so much for classifying whether a bigram is a collocation or not.
- Hypothesis Testing Of Differences is most useful for choosing between alternatives while generating text.
- Pearson's χ^2 test is useful for identification of translation pairs among aligned corpora and measuring corpus similarity.
- Likelihood Ratios are more appropriate for sparse data than any other statistical method.

2.1.3 Word Association Measures

[2]

This is one of the very early attempts at collocation extraction by Kenneth Church and Pattrick Hanks (1990). They have generalized the idea of collocation to include **co-occurrence**. Two words are said to co-occur if they appear in the same documents very frequently.

For example: *doctor* and *nurse* or *doctor* and *hospital* are highly associated with each other as they occur together very frequently in a text.

The information theoretic measure, mutual information was used for measuring the word association norms from a corpus and then the collocations were produced.

2.1.3.1. Word Association And Psycholinguistics

Word association norms are an important factor in psycholinguistic research. Informally speaking, a person responds quicker to a word hospital when he has encountered a highly associated word doctor before. In a psycholinguistic experiment a few thousand people were asked to write down a word that comes to their mind after each of the 200 words that were given to them. This was an empirical way of measuring word associations.

2.1.3.2. Information Theoretic Measure

Mutual Information: If two words(x, y) have their probability of occurrence as P(x) and P(y) then their mutual information is defined as:

$$I(x,y) = log_2 \frac{P(x,y)}{P(x)P(y)}$$

Informally, mutual information compares the probability of x and y appearing together to, the probability of x and y occuring independent of each other. If there is some association between x and y then the mutual probability P(x,y) will be much greater than their independent probability P(x).P(y) and hence I(x,y)>>0. On the other hand,if there is no association between x and y then $P(x,y) \approx P(x).P(y)$, hence $I(x,y) \approx 0$.

The word probabilities P(x) and P(y) are estimated by counting the number of observations of x and y in a corpus (normalized by N,the size of the corpus).

Mutual probabilities, P(x,y) is estimated by counting the number of times x is followed by y in a window of w words, $f_w(x,y)$ (normalized by N,the size of the corpus). The window size allows us to look for different kinds of associations. Smaller window size identifies the fixed expressions whereas larger window size enables us to understand semantic concepts.

The association ratio is technically different from mutual information since in case of mutual information $f(x,y) = f(y,x)$ but that is not the case for association ratio because here linear precedence is taken into account.

2.1.3.3. Lexico-Syntactic Regularities

The association ratio is also useful to find out important lexico-syntactic relationships between verbs and their arguments or adjuncts. For example, consider the phrasal verb *set off*.

Using Sinclair's estimates

$$P(set) \approx 250 * 10^{-6}, P(off) \approx 556 * 10^{-6}$$
$$P(set, off) \approx 70/(7.3 * 10^{6})$$

The mutual information for *set off* is:

$$I(set; off) = log_2 \frac{P(set, off)}{P(set)P(off)} \approx 6.1$$

From the above value we can infer that the association between *set* and *off* is quite large (2^6 i.e. 64 times larger than chance).

2.1.3.4. *Importance Of Word Association*

This was a pioneering approach towards extracting word associations. It extended the psycholinguistic notion of word association norm towards an information theoritic measure of mutual information. Informally,it helped us predict what word to look for if we have encountered some word. A lot can be predicted about a word by looking at the company that it keeps.

2.1.4 Retrieving Collocations From Text : XTRACT *[9]*

Frank Smadja has implemented a set of statistical techniques and developed a lexicographic tool, Xtract to retrieve collocations from text. As already stated, the definiton of collocation varies from one author to another.

According to the author, collocations have the following features:

- **Arbitrary** : They cannot be directly translated from one language to another as they are difficult to produce from a logical perspective.
- **Domain-dependent** : There are expressions which make sense only in a specific domain. These collocations will be unknown to someone not familiar with the domain.
- **Recurrent** : Collocations are not exceptional or chance co-occurences of words, rather they occur very frequently in a given context
- **Cohesive lexical clusters** : Encountering one word or one part of a collocation often suggests the probability of encountering the rest of the collocation as well.

The author has also classified collocations into three types:

- **Predicative Relations :** Two words are said to form a predicative relation if they occur very frequently in a similar syntactic structure (like, Adjective-Noun, Noun-Verb etc)
  For example : *make-decision* , *hostile-takeover*
- **Rigid Noun Phrases :** This involves uninterrupted, fixed sequences of words
  For example : *stock exchange,railway station*
- **Phrasal Templates :** Phrasal templates consist of idiomatic phrases consisting of one or more or no empty slots. These are generally used for language generation.
  For example : *Temperatures indicate yesterday's highest and lowest readings* is how generally a weather report begins.

2.1.4.1. Xtract: The lexicographic tool for collocation extraction

Xtract does a three stage analysis to locate interesting word associations in the context and make statistical observation to identify collocations. The three stages of analysis are:

- First Stage: statistical measures are used to retrieve from a corpus pair wise lexical relations.
- Second Stage: uses the output bigrams (of 1st stage) to produce collocations of n-grams.
- Third Stage: adds syntactic information to collocations retrieved at the first stage and filters out inappropriate ones.

The experiments were carried out on a 10million word corpus of stock market news reports.

2.1.4.1.1. Xtract: Stage One

Two words are said to co-occur if they are in a single sentence and if there are fewer than five words between them.

The words form a collocation if:

- They appear together significantly more often than expected by chance.
- Because of syntactic constraints they appear in a rigid way.

The algorithm used for extracting the bigrams forming collocations is:

1. Given a tagged corpus output all sentences containing a word w
2. Produce a list of words w_i with frequency information on how w and w_i co-occur.

$Freq_i$ (the frequency of appearance of w_i with w in the corpus), POStag of w_i, $P_j^i(-5 \geq j \leq 5, j \neq 0)$ (frequency of occuring w_i with w such that they are j words apart).

3. Analyze the statistical distribution and select interesting word pairs.

$$\text{Strength }(\text{w}, w_i) = k_i = \frac{freq_i - \bar{f}}{\sigma}$$

$\bar{f}$ and σ are the average frequency and standard deviation of all the collocates of a word w

$$\text{Spread }(U_i) = \frac{\sum_{j=1}^{10} (p_i^j - \overline{p}_i)^2}{10}$$

If U_i is small then the histogram will be flat implying that w_i can be used at any position around w. Whereas if U_i is large then the histogram will have sharp peaks implying that w_i can be used only in some specific positions around w.

At the end of this stage a lexical relation corresponding to w is produced as output. It is of the form of a tuple (w_i ,distance,strength,spread,j) verifying the following inequalities:

$$\text{Strength} = \frac{freq_i - \bar{f}}{\sigma} \geq k_0$$

$$\text{Spread } \geq U_0$$

$$p_j^i \geq \overline{p}_i + (k_1 * \sqrt{U_i})$$

Where k_0, k_1, U_0 are thresholds set manually.

2.1.4.1.2. *Xtract: Stage Two*

The second stage of Xtract produces collocations consisting of more than two words and also filters out some pairwise relations. The algorithm followed in stage two is given below.

1. Produce Concordances : Given a pair of words and the distance of the two words, produce all the sentences containing them in the specific position.
2. Compile and Sort : compute the frequency of appearance of each of the collocates of w
3. Analyze and Filter : a word or a POS is kept in the final n-gram at position if and only if

$$p(word[i] = w_0) \geq T$$

where T is a threshold set manually while performing the experiment

Some of the results after stage two are shown below:

Tuesday	**the Dow Jones industrial average**	**rose 26.28 points to 2304.69**
	The Dow Jones industrial average	went up 11.36 points today.
…that sent	the Dow Jones industrial average	down sharply..
Monday	the Dow Jones industrial average	was down 17.33 points to 2287.36…
…in	the Dow Jones industrial average	was the biggest since…

Figure 2.7 : Producing concordances for “the Dow Jones Industrial Average”[**9**]

The NYSE composite index of all its listed common stocks	**fell 1.76 to 164.13**
The NYSE composite index of all its listed common stocks	fell 0.98 to 164.97
The NYSE composite index of all its listed common stocks	fell 0.91 to 164.98
The NYSE composite index of all its listed common stocks	rose 0.76
The NYSE composite index of all its listed common stocks	fell 0.33 to 170.63

Figure 2.8: Producing the “NYSE's composite index of all its listed common stocks “ [**9**]

In stage two of Xtract:

- Phrasal templates are also produced in addition to rigid noun phrases
- Produces the biggest possible n-gram
- Relatively simpler way of producing n-grams

2.1.4.1.3. Xtract: Stage Three

In stage three of Xtract the collocations produced in stage one are analyzed and the syntactic relationship between them is established otherwise they are rejected.

1. Produce Concordances : Given a pair of words and the distance of the two words, produce all the sentences containing them in the specific position.
2. Parse : For each sentence produce set of syntactic labels

3. Label and Filter : count the frequencies of each possible label identified for the bigram (w,wi) and accept if and only if

$$p(label[i] = t) \geq T$$

Where T is a threshold defined manually while performing the experiment

For example: If after the first two stages of Xtract the collocation *make-decision* is produced then in the third stage it is identified as a *verb-object* collocation. If no such relationship can be established then such collocations are rejected.

2.1.4.2. Analysis Of Xtract

The precision and recall value of Xtract are 80% and 94% respectively. An observation that can be made from the results of Xtract is that the extracted collocations are domain dependent. Hence the domain and size of the corpus has heavy influence on the type of collocations extracted from it. This work showed a nice method of extracting 'n-grams' and by adding syntax to the collocations it could explain the syntactic relationships between the colloactes as well.

2.1.5 Collcation Extraction By Conceptual Similarity *[*5*]*

This is a method suggetsed in [**5**] where the author uses Wordnet to find out the conceptual similarity between different words. It is observed that in spite of the similarity between words due to the arbitrary nature of collocations only one of the many possible synonyms of a word a candidate phrase prefers one word over another. From this point of view collocation can be redefined as:

A pair of words is considered to be a collocation if one of the words significantly prefers a particular lexical realization of the concept the other represents. Consider the following examples:

Correct Expression	Incorrect Expression
many thanks	several thanks
emotional	emotional luggage

baggage	
strong coffee	powerful coffee
tap water	pipe water

Table 2.1: Collocation Preference

For example, *coffee* significantly prefers *strong* over *powerful* and similarly the other examples. In this new outlook there's an inherent directionality as each candidate phrase prefers one synonym over another. So this is termed as **collocation preference**.

The authors studied the usages of two similar words, *baggage* and *luggage:*

1. 2 million parsed sentences of BNC were searched for occurrences of the synonyms *baggage* and *luggage*. If the difference of their occurrence for a particular word was greater than 2 then that bigram was taken into account.
2. For each such bigram obtained in step1, Alta Vista search was used to find occurrences of it in the world wide web.
3. Details of collocation according to CIDE(Cambridge International Dictionary Of English) was used as standard of judgment.

Figure 2.9 shows the difference in usage for the two synonyms *baggage* and *luggage*.

Word	BNC	Alta Vista	CIDE	Collocation
allowance	B 5 0	B 3279 502	B	baggage allowance
area	B 3 1	B 1814 1434		? baggage area ?
car	B 4 0	B 3324 357	B	baggage car
compartment	L 1 3	L 2890 5144	L	luggage compartment
label	L 0 6	L 103 333	L	luggage label
rack	L 0 8	L 164 14773	L	luggage rack

Figure 2.9: Collocational Information for 'baggage' and 'luggage'[**5**]

2.1.5.1. Collocation Graph

Collocation graphs are diagrammatic representation of the different senses represented by a word and the arcs are used to denote colloactional preferences described as follows.

2.1.5.1.1. Concept Set

A collocation graph consists of two or more **concept nodes** that represent the senses that a word has according to the Wordnet. For a word *w* the concept set *C(w)* is defined as:

$$C(w) = \{S_i : w \in S_i\}$$

For example, the word *information* has five meanings according to the Wordnet. So its concept node will have five entries, one for each of the meanings.

2.1.5.1.2. Intersection Of Concept Sets

If two words are synonyms in some sense i.e. they share a sense in common then their concept nodes will have an intersection and that sense (common to both of them will be present in the intersection).

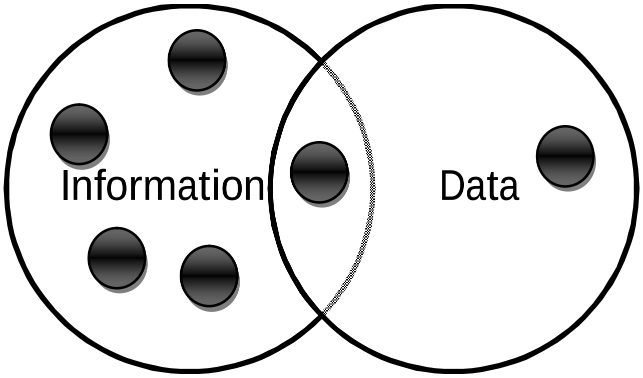


Figure 2.10: Intersection Of Concept Sets for information and data

Figure 2.10 shows the intersection of the concept sets of *information* and *data*.

2.1.5.1.3. Collocation Preference

Concept nodes in a concept graph are connected by collocation arcs to show the preference that is being exhibited due to the property of collocations. The direction of the arc represents which word is expressing preference for which word.

2.1.5.1.4. Intersection Graphs

While trying to determine significant collocations the concept nodes for the synonyms are drawn. They have one or more senses in common. A candidate phrase is said to exhibit collocational preference if it is expressing more preference for one word than the other for representing the same sense. This is denoted by a directed preference arc in the collocation graph and the arc passes through the preffered word first. This is shown as an example for *emotional baggage* and *emotional luggage* in Figure 2.11.

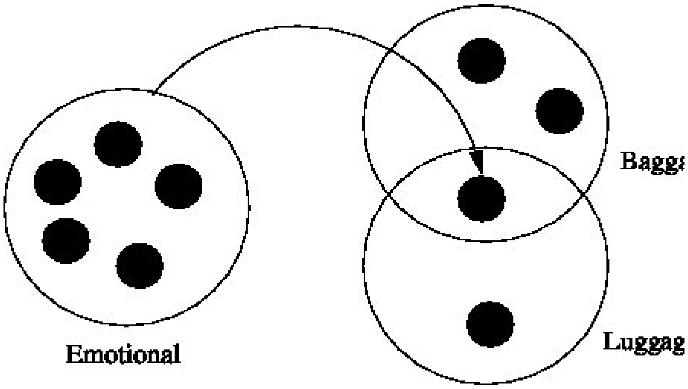


Figure 2.11: Collocational Preference

2.1.5.2. *Poly-Collocations*

It might be possible for a word to express preference for another word in more than one of its synsets. These are termed as **poly-collocations**. Depending on whether sense information is available or not a variety of configurations are possible for the collocation graph.

2.1.5.3. *Anti-collocations*

A synonym set with respect to a particular target phrase can be classified into three disjoint sets :

- The words which are frequently used with the target word (**Collocations**).
- The words which are generally not used with the target word but do not lead to unnatural reading.
- The words which are never used with the target word (**Anti Collocation**).

The knowledge of anti-collocations will be very much helpful for natural language generation and foreign language learners.

Example: *strong drugs, *powerful coffee

2.1.5.4. *Formalization*

The algorithm takes a sequence of bigrams $p^1, p^2 ... p^N$ as input.

- The occurence count for each such pair is defined as :

$$C(a,b) = \sum_{i=1}^{n}(\delta(p^i = \prec a,b \succ)$$

where, δ (x) = 1,if x is true and is 0, if x is false.

- The co-occurence set of a word w is defined as:

  $$cs(w) = \{v : c(w,v) > 0\}$$

- Wordnet is defined as a set of synsets,W. Candidate collocation synset of a word w is defined as:

  $$CCS(w) = \{S \in W : | S \cap cs(w) | > 2\}$$

  So each candidate collocation synset S,(for a word w) consistes of atleast two elements whose co-occurence count is non-zero.

- Most frequently co-occurring element of a synset and its frequency are defined as:

  $$w^{'} = argmaxc(w,v) f^{'} = argmaxc(w,v)$$

- Collocation strength is defined as $f^{'} - f^{''}$ where $f^{''}$ is the second highest frequency in the synset.

2.1.5.5. Analysis

The idea presented in the paper looks promising and since the work is at the semantic level it is more intuitive and easy to connect to how a human mind works in reality.

The future work needs to focus on improving the basic algorithm in particular aspects :

- The idea of synonym set can be extended to concept set.
- Experiments need to be conducted for synsets other than Nouns.
- Morphological processing need to be done.
- Some thesaurus can be used along with Wordnet.

2.1.6 Verb Phrase Idiomatic Expressions *[4]*

An idiom can be defined as a speech form or an expression of a given language that is peculiar to itself grammatically or cannot be understood from the individual meanings of its elements.

For example: by and large, spill the beans, shoot the breeze, break the ice etc.

These are very typical to a language and evolve over time. Even within a language they vary from one dialect to another.

Idioms don't follow some general conventions among its class. Like,some of them might allow some form of verbal inflection (*shot the breeze*) whereas some might be completely fixed (*now and then*). The idioms that are perfectly grammatical are difficult to be identified as an idiom having idiosyncratic meaning as opposed to its similar structures (*shoot the breeze* and *shoot the bird*).

The authors have looked into two closely related problems confronting the appropriate treatment of Verb-Noun Idiomatic Combinations(where the noun is the direct object of the verb):

- The problem of determining their degree of flexibility
- The problem of determining their level of idiomaticity

2.1.6.1. Recognizing VNICs

Even though VNICs vary in their degree of flexibility on the whole, they contrast with compositional phrases (which are more lexically productive and appear in a wider range of syntactic forms). Hence the degree of lexical and syntactic flexibility of a given verb+noun combination can be used to determine the level of idiomaticity of the expression. The authors have tried to measure the lexical and syntactic fixedness of an expression by a statistical approach to determine whther it is an idiom or not.

2.1.6.2. Analysis

Idioms form a very interesting part of natural language but due to its pecularity and arbitrary nature it has been side-stepped by the NLP researchers for long. The authors have tried to provide an effective mechanism for the treatment of a broadly documented and crosslinguistically frequent class of idioms, i.e., VNICs. They have done a deep examination of several linguistic properties of VNICs that distinguish them from similar literal expressions. Novel techniques for translating such characteristics into measures that predict the idiomaticity level of verb+noun combinations have also been proposed.

2.1.7 Extraction of Multi-word Expressions from Small Parallel Corpora

[**10**]

The authors present an approach towards detecting multiword expression using bilingual parallel corpora using word alignment. Their methodology for extracting multiword expressions is:

- Use GIZA++ to automatically align bilingual parallel corpora
- Extract misalignments as MWE candidates
- Use large monolingual corpus to filter and rank the candidates
- Extract the translations of the MWEs from parallel corpus and use in Machine Translation system

Unlike many other approaches which trust on word alignment for detecting multiword expressions, this method focuses on 'misalignments', i.e. the word combinations which the automatic word aligner failed to align. Idiomatic expressions are often translated to a single word (or to an expression with a different meaning) in another language. However, due to the non compositional nature of these expressions, word alignment algorithms have difficulties aligning MWEs, hence 1:n and n:m alignments are often noisy.

2.1.7.1. Identifying MWE candidates

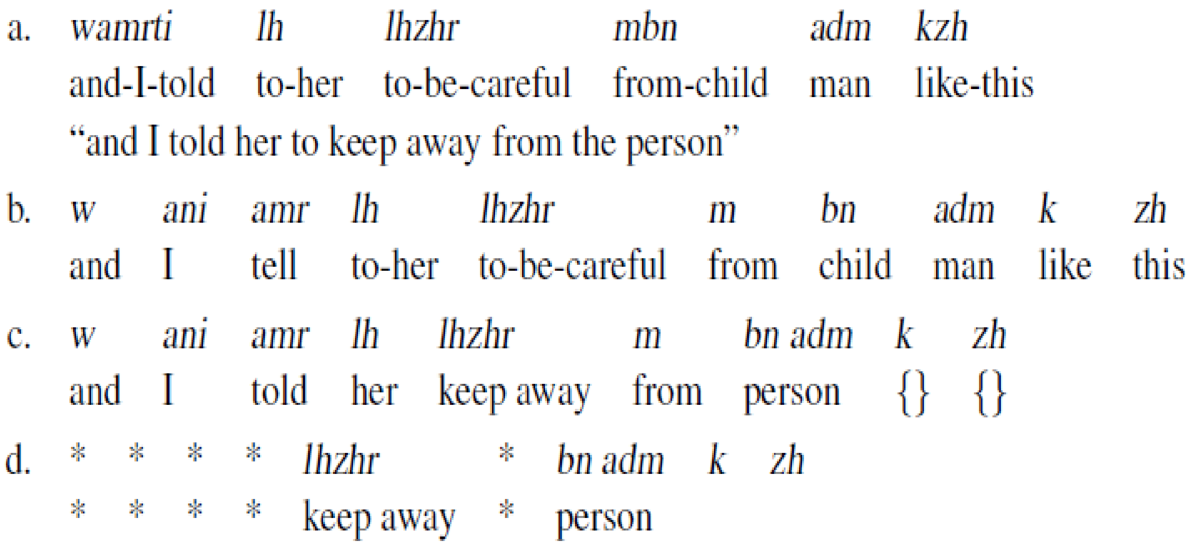
The different resources that have been used in this approach are:

- **A small bilingual, sentence aligned parallel corpus**: A Hebrew-English corpora consisting of 19,626 sentences, mostly from newspapers.
- **Large monolingual corpora**: For Hebrew they have used Morph analyzed MILA corpus which contains 46,239,285 tokens and for English they have used Google's Web 1T corpus
- **Morphological processors** for both languages
- A **bilingual dictionary** consisting of 78,313 translation pairs

To reduce data sparsity and language specific differences both the corpora are preprocessed. Tokenization and removal of punctuation is done. Stop words from English corpus are also removed. The Hebrew corpus is analyzed morphologically, each word is reduced to its base form, and bound morphemes are split to generate stand-alone “words”. English side of the corpus is also tokenized and lemmatized using NLTK package. Frequent function words are removed at this stage.

GIZA++ have been used to word align the bilingual parallel corpora. The quality of alignments is checked against the bilingual dictionary. If a 1:1 alignment is present in the dictionary, that implies it is a valid translation pair and hence not an MWE. If a 1:1 alignment is not present in the dictionary but appears very frequently in the corpus and have been aligned with a high score, they are added to the dictionary and also retained as a multiword candidate. All the misalignments (i.e. not 1:1 alignments) are taken into consideration as multiword expressions. The following assumptions are made for misalignments in a parallel corpora; either they are due to language specific differences (morphological or syntactical) or due to noise (from translation source or word alignment algorithm) else they are multiword expressions since they can trigger 1:n or m:n alignments.

Figure 2.12 shows an example of mwe extraction using this approach.



a.	*wamrti*	*lh*	*lhzhr*	*mbn*	*adm*	*kzh*
	and-I-told	to-her	to-be-careful	from-child	man	like-this

“and I told her to keep away from the person”

b.	*w*	*ani*	*amr*	*lh*	*lhzhr*	*m*	*bn*	*adm*	*k*	*zh*
	and	I	tell	to-her	to-be-careful	from	child	man	like	this

c.	*w*	*ani*	*amr*	*lh*	*lhzhr*	*m*	*bn adm*	*k*	*zh*
	and	I	told	her	keep away	from	person	{}	{}

d.	*	*	*	*	*lhzhr*	*	*bn adm*	*k*	*zh*
	*	*	*	*	keep away	*	person		

Figure 2.12: identifying MWE by alignment[**10**]

In order to rank the extracted mwe candidates statistics from a large monolingual corpus is used. PMI score is calculated from the monolingual corpus for the bigrams identified due to misalignments. At this stage the noise due to poor translation and erroneous word alignment is also eliminated as, about 20,000 candidate MWEs are removed in this stage because they do not occur at all in the monolingual corpus.

The quality of machine translation has been checked after incorporating the knowledge of multiword expressions. For each MWE in the source-language sentence, all the words in the target language that are aligned to the word constituents of the MWE , are considered as translation. 2,955 MWE translation pairs and 355 translation pairs produced by high-quality word alignments are augmented to the dictionary.

Dictionary	**BLEU**	**Meteor**
Original	13.69	33.38
Augmented	13.79	33.99

Figure 2.13: Performance of machine translation with MWE knowledge

The algorithm proposed by the authors capitalizes on semantic cues provided by ignoring 1:1 word alignments, and viewing all other material in the parallel sentence as potential MWE. It also emphasizes the importance of properly handling the morphology and orthography of the languages involved, reducing wherever possible the differences between them in order to improve the quality of the alignment.

2.2. Study of an Ongoing Project: MWEToolkit *[6]*

Multiword Expression Toolkit (mwetoolkit) is developed for type and language-independent MWE identification. It is a hybrid system for detecting multiwords from a corpus using rule based as well statistical association measures. The toolkit is an open source software can be downloaded from sf.net/projects/.

2.2.1 MWEToolkit System Architecture

Given a text corpora the toolkit filters out the MWE candidates from the corpora. The different phases present in the toolkit to achieve this goal are:

1. Preprocessing the corpus: Preprocess the corpus for lowercase conversion, lemmatization and POS tagging (using Tree tagger).
2. Extract 'ngrams' depending on the predefined POS patterns.
3. For each of these bigrams take into account their corpus count as well as the web count (number of pages in which the particular bigram is present) using Google and Yahoo
4. Apply some Association Measures (statistical) to filter out the candidates.

    i. The corpus containing the N word tokens is indexed and from that index the counts of the tokens are estimated. Using the index, individual word counts, $c(w_1)$, $c(w_2)$……$c(w_n)$ and the overall ngram count $c(w_1w_2…w_n)$ is computed.

    ii. The expected N gram is computed if words occurred just by chance

$$E \approx \frac{c(w_1).c(w_2).c(w_3)\dots\dots c(w_n)}{N^{n-1}}$$

    iii. Using the above information four Association Measures are computed

- Maximum Likelihood Estimator

$$mle = \frac{c(w_1w_2\dots\dots\ w_n)}{N}$$

- Dice's coefficient

$$dice = \frac{n * c(w_1w_2\dots\dots\ w_n)}{\sum_{i=1}^{n} c(w_i)}$$

- Pointwise Mutual Information

$$pmi = \log_2 \frac{c(w_1w_2\dots\dots\ w_n)}{E(w_1w_2\dots\dots\ w_n)}$$

- Students' t-score

$$t - score = \frac{c(w_1 w_2 \dots \dots\ w_n) - E(w_1 w_2 \dots \dots\ w_n)}{\sqrt{c(w_1 w_2 \dots \dots\ w_n)}}$$

5. Once each candidate has a set of associated features, an existing machine learning model can be applied to distinguish true and false positives or a new model can be designed by assigning a class to the new candidate set.

2.2.2 Using Web as corpora

Another novel aspect of the system is, it uses web count of MWEs as a feature for their Machine Learning model. Let us look a bit more closely and analyze the advantages and disadvantages of using web as a corpus.

Issues:

- Web counts are "estimated" or "approximated" as page counts, whereas standard corpus counts are the exact number of occurrences of the n-gram.
- In the web count, the occurrences of an n-gram are not precisely calculated in relation to the occurrences of the (n − 1)-grams composing it.

  For instance, the n-gram "the man" may appear in 200,000 pages, while the words "the " and "man" appear in respectively 1,000,000 and 200,000 pages, implying that the word "man" occurs with no other word than "the".
- Unlike the size of a standard corpus, which can be easily computed, it is very difficult to estimate how many pages exist on the web and especially because this number is always increasing.

Advantage:

- In spite of the issues, the biggest advantage of the web is its **availability**, even for resource-poor languages and domains. It is a free, expanding and easily accessible resource that is representative of language use, in the sense that it contains a great variability of writing styles, text genres, language levels and knowledge domains.

- The web can minimize the problem of sparse data. Most of the statistical methods suffer due to the sparsely distributed data in the corpus. Web can lend a hand for dealing with this problem. Due to the sheer volume of data present on the web, it can assist us to distinguish rare occurrences from invalid cases.

Conclusion

We have presented the notion of multiword expressions through various definitions and numerous examples. We have also presented a literature survey on the extraction approaches of multiword expressions. We can observe that that there have been very different approaches towards detection of multiword expressions. Researchers have formed rules, applied statistical association measures and used alignment from parallel corpora to detect multiword expressions. Multiword expressions are of diverse nature and not one best method exists to extract mwes of all types.

Bibliography

[1] Tanmoy Chakraborty and Sivaji Bandyopadhyay, "Identification of Reduplication in Bengali Corpus and their Semantic Analysis : A Rule Based Approach," in *Proceedings of the Multiword Expressions: From Theory to Applications*, 2010.

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

[3] Ted Dunning, "Accurate Methods for the Statistics of Surprise and Coincidence," in *Computational Linguistics*, 1993.

[4] Afsaneh Fazly and Suzanne Stevenson, "Automatically Constructing a Lexicon of Verb Phrase Idiomatic Combinations," in *EACL*, 2006.

[5] Darren Pearce, "Using conceptual similarity for collocation extraction.," in *Proceedings of the Fourth annual CLUK colloquium*, 2001.

[6] Carlos Ramischy, Aline Villavicencio, and Christian Boitet, "Multiword Expressions in the wild? mwetoolkit comes in handy," in *COLING*, 2010.

[7] Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger, "Multiword Expressions:A pain in the neck for NLP," in *CICLing. Springer*, 2002.

[8] Christopher Manning and Henry Schutze., *Foundations Of Statistical Natural Language*.: MIT Press, 1999.

[9] Frank Smadja, "Retrieving collocations from text: Xtract," in *Computational Linguistics*, 1993.

[10] Yulia Tsvetkov and Shuly Wintner, "Extraction of multi-word expressions from small parallel corpora," in *23rd International Conference on Computational Linguistics*, 2010.

[11] Veronika Vincze, Istvan Nagy T, and Gabor Berend, "Detecting noun compounds and light verb constructions: a contrastive study," in *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE 2011)*, 2011.