International Journal of Engineering & Technology, 7 (2.6) (2018) 33-38

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research Paper

Hybrid method for automatic extraction of multiword expressions

Shaishav Agrawal [1] *, Ratna Sanyal [2], Sudip Sanyal [2]

[1] *Indian Institute of Information Technology Allahabad, Allahabad – 211012, India*
[2] *Computer Science Engineering, School of Engineering and Technology, BML Munjal University, Gurugram – 123413, India*
**Corresponding author E-mail: shaishav.engr@gmail.com*

Abstract

A three phase hybrid method for automatic extraction of English multiword expressions (MWEs) has been proposed. The proposed method is based on linguistic patterns, association and context similarity between constituent words of the MWEs. First, the expressions are extracted in the form of N-grams from the raw text and then filtered using well defined linguistic patterns. Next, these expressions are again filtered using association score and context similarity score between their constituent words. Two association measures, Dice's coefficient and PMI have been used for calculating the association score. The context similarity between words has been calculated using Latent Semantic Analysis (LSA) method. The problem of deciding the best value for the cut-off boundary thresholds in statistical methods is quite common. A two phase method of deciding the boundary threshold, using training dataset, has been proposed and employed in the current work. De-tailed performance analysis has been done on manually annotated dataset. The significant gain in performance has been observed for various types of multiword expressions.

Keywords: *Collocation Extraction; Information Retrieval; Latent Semantic Analysis; Multiword Expressions; Natural Language Processing.*

1. Introduction

Multiword expressions (MWEs) are an assorted class of linguistic expressions that are treated as a single lexeme which cross word boundaries [9] [34]. In other words, it is a combination of two or more words which gives a sole combined gist. MWEs possess some idiosyncratic properties that are not completely predictable from a compositional analysis. MWEs display syntactic, lexical, semantic, statistical and/or pragmatic idiosyncrasies [7] [9]. MWEs are very commonly used in all the languages. There is an estimation that in English language 30% to 45% content of the spoken language and 21% content of the academic prose is MWEs [5]. Jackendoff [19] have estimated that there are similar numbers of MWEs used in spoken language as the count of single words. Recognition of multiword expressions is very important for many NLP applications [13] like machine translation [18] [30], speech recognition [39], alignment of parallel corpora [24], text summarization [36], information retrieval [42], etc.

The objective of the present paper is to design a generalized automated method for extraction of all types of MWEs. The extraction of MWEs is a complex task for an automated machine. However, MWEs show some properties and idiosyncrasies which make them different from simple expressions. Any MWE extraction method can be designed using these properties shown by MWEs. The baseline of the current work is our previous work based on two properties of MWEs: linguistic pattern of expressions (i.e. compound noun, verb article noun, noun preposition noun, etc) and statistical idiosyncrasy (i.e. the association score between the constituent words of MWEs is higher than the normal expressions) [2]. The present work is based on the observation that often the constituent words of MWEs are used in the related context i.e. there is a context similarity between them. In other words, we can say that in any multiword expression the constituent words are correlated. Thus, we have started with a two phase method, namely Context Similarity based Filtering (CSBF) method using this concept. In the first phase, the N-grams are extracted from the raw text and then filtered using some fixed linguistic patterns similar to the previous work. Next, these N-grams are classified as MWEs using context similarity in the second phase. The extraction method of MWEs using the concept of context similarity may provide better results if we use this concept with the concept of association between the constituent words of MWEs together within a single framework. The method proposed in previous work [2] is also based on two properties of MWEs: linguistic formation and association between their constituent words. Thus, combining both of the methods may provide better MWE extraction method. So, a three phase hybrid method (Hybrid Dual Filtering method) has been proposed. In the first phase, the N-grams are extracted from the raw text using linguistic patterns similar to previous method. Then, these extracted expressions are filtered in the next two phases with the help of association scores and context similarity scores sequentially.

Two association measures (AMs), Dice's coefficient and PMI have been used for calculating the association scores. The context similarity between words has been calculated using the well known Latent Semantic Analysis (LSA) method [10]. LSA is a method for categorizing words by calculating contextual usage meanings of the words using statistical calculations [25]. It is basically used to calculate the semantic similarity between words. However, LSA also groups the words according to their contextual usage. Although, in most of the MWEs the constituent words do not have similar meaning but their contexts are different from general usage of these constituent words. Consider the example, black book. There is no similarity of meaning between black and book but they are used in a related context. The context between other combinations of book like red book, white book, etc. is the color of book. Similarly the

other combinations of black with other words like black horse, black shoes, etc. show the context as color. However, in the expression black book the context is "a book containing a list of secret contacts" which is quite different. Thus, in this example LSA categorized the different combinations of black with other words and similarly book with other words according to different contexts. The word combination of black and book will show different context compared to the general context (color) of black with other words and book with other words. This difference can be calculated using LSA and the combination, which have a different context, may be an MWE.

The efficacy of the filters or classifiers based on association and context similarity scores depends critically on the value of the boundary/ cut-off threshold. A boundary/ cut-off threshold in present case is that value of association score or context similarity score which separates the MWEs from the normal expressions. Thus, determination of these threshold values is a matter of concern. A unique two phase method for calculating the thresholds of both association and context similarity based filters has been designed using annotated training dataset for the calculation of cut-off threshold. The evaluation of the proposed methods is done on manually annotated dataset and satisfactory gain in performance has been observed over the baseline methods.

The rest of the paper is organized as follows. Section 2 briefly describes related work in the field of multiword extraction. The detailed methodology is presented in Section 3. Section 4 is regarding experimental results. Here, the detailed analysis of the evaluation results obtained from all the methods is presented. Finally the concluding remarks are given in Section 5.

2. Related work

The field of multiword expressions is not very old but still many research works are available in this field. Some popular methods for extracting MWEs are statistical or association measure based methods [14], semantics based methods [3] [22], symbolic, syntactical or linguistic pattern based methods [16] [44], word alignment based methods [29] [37], hybrid methods [6] [12], etc.

Multiword expressions are extracted using association scores in statistical or association measure based methods. In these methods the candidate expressions are ranked according to their association scores using any statistical technique. The limitation of this method is that it works well only for those types of multiword expressions which show statistical idiosyncrasy such as Compound Nouns, Verb Particles, etc. while idiomatic MWEs such as Verb Noun constructions are difficult to extract with statistical methods. A domain independent MWE extraction algorithm for extracting technical MWE phrases has been proposed [20]. Here two constraints, repetition (frequency) and linguistic pattern (structure of term) have been used for identifying technical terms. Point Wise Mutual Information and T-test have also been used as association measures for the extraction of MWEs instead of simple frequency [8] [38]. Boundary threshold decision and choosing the best statistical measure are other problems with this method. Evert and Krenn [14] have proposed a method for obtaining best statistical measure from a random sample taken from the whole corpus which can be easily generalized for the whole corpus. Agrawal et al. [2] have proposed a solution for deciding boundary threshold using annotated training dataset. Piasecki et al. [32] have compared various statistical measures on a large corpus and also on a large WordNet. They have also tested the combination of various association measures and found that the combination of measures performs better than any single measure.

Working of semantics based methods is similar to the statistical methods but semantics based methods use semantics feature for classification of MWEs instead of statistical variance. These methods also use statistical measures for calculating semantic similarity. Techniques like Latent Semantic Analysis (LSA) [10] or Probabilistic Latent Semantic Analysis (PLSA) [17] have been used for measuring semantic similarity [3] [22]. Semantics based methods are basically used for identifying non-compositionality (idiomaticity) in the multiword expressions. Baldwin et al. [3] have used LSA for extracting noun-noun and verb-particle MWEs. Katz and Giesbrecht [22] have also used semantic method for classifying non-compositional German multiword expressions.

In symbolic, syntactical and linguistic pattern based methods the multiword expressions are extracted using linguistic patterns and grammatical rules of the natural language. These methods are strictly language dependent and needs large annotated corpora. Goldman et al. [16] have used parsing technique for extracting multiword expressions which works on the syntactic pattern of multiword expressions.

Word alignment methods are used to map the multiword expressions between different languages. These methods extract the MWEs in multilingual concept. If the multiword expressions are known in one language then multiword expressions for another language can be extracted with the help of parallel corpora. These methods are also useful in statistical machine translation for aligning multiword expressions in both the languages. Moirón and Tiedemann [29] have proposed a method to identify idiomatic expressions using word alignment technique. Here, the MWEs are extracted using meaning predictability and similarity between the contextual meaning of a multiword expression and the contextual meaning of combination of constituent words. Tsvetkov and Wintner [40] have proposed an alignment based approach for extraction of MWEs from the parallel corpora. They have used semantic cues and focused on misalignments in the sentences in parallel corpora.

Hybrid methods are the combinations of simple methods. Sometimes, these methods perform better than simple methods. However, they have also the limitations of the constituent methods. Statistical method has been extended with the combination of substitution method for extraction of idiomatic MWEs [27] [31] [35]. The constituent words of the phrases are replaced with other words in substitution methods and the new expressions are mapped with original expressions. E.g. hot dog can be mapped with expressions such as cold dog, warm dog, hot cat, etc. Another hybrid method namely Maximum entropy (MaxEnt) model has been proposed for MWE extraction [43]. The relative compositionality of Hindi Noun+Verb MWEs has been measured in this work. Various features like collocation based features (i.e. statistical features), word based features, and contextual features have been mapped into a single method using MaxEnt model [43]. Karan et al. [21] have developed classification algorithms and evaluated various features for Croatian collocation extraction. A methodological framework using various features has also been developed for automatic extraction of MWEs [33]. A two phase hybrid method using association measures and linguistic patterns for MWE extraction has been developed [1]. Agrawal et al. [1] have also proposed two threshold decision methods using maximizing the recall and minimizing the classification error. Further, this work has been extended using similar properties [2]. Agrawal et al. [2] have proposed that different threshold should be calculated for separate categories of MWEs. A machine learning based method has been proposed for Russian MWE extraction [41]. They have used advanced learning to rank method with rich sets of features and two data corpuses in parallel for improving the performance of simple learning methods. Liang et al. [26] have also proposed a hybrid method. They have used correlation degree and sequence information of phrase for training the MWE extraction algorithm. This method combines weakly supervised K-means cluster, word correlation degree and Bidirectional long short-term memory (Bi-LSTM) for calculating the correlation degree and sequence information.

Although a variety of research works have been found in MWE extraction task but the majority of research is focused on noun compounds, idioms, light verb constructions, and verb-particles [4] [8] [23] [35] [38]. A good accuracy has been achieved for noun compounds and verb-particles but for light verb constructions and idioms there are still lot of challenges. Most of the previous works have been related to particular types of MWEs and very few works are available which are generalized i.e. for extraction of all types of

MWEs. The threshold decision problem has also not been solved properly in most of the previous works.
Our proposed method is a generalized hybrid method for extraction of all types of MWEs. The details of our proposed methodology, different linguistic patterns used and evaluation results are presented in the next sections.

3. Methodology

The proposed method is similar to the Multiple Threshold (MT) method proposed by Agrawal et al. [2]. However, the context similarity feature has been used here instead of association feature for MWE extraction. In the proposed Context Similarity based Filtering (CSBF) method, firstly the expressions are filtered by linguistic patterns and then these expressions are again filtered using context similarity scores. Context similarity between the words is calculated using Latent Semantic Analysis (LSA) method. Besides, a hybrid three phase method, namely Hybrid Dual Filtering (HDF) method has also been proposed. It is a combination of MT and CSBF method. In HDF method the MWEs are extracted using three properties of MWEs. These properties are: syntactical idiosyncrasy (i.e. following the structure of specific linguistic patterns), statistical idiosyncrasy (i.e. association between constituent words of MWEs is higher than normal expressions) and context correlation between their constituent words. First of all, the expressions from the raw text are extracted in the form of N-grams and then these N-grams are filtered by linguistic patterns (an ordered sequence of POS tags of words). Next, these filtered expressions are again filtered using association measure (AM) based filter and context similarity based filter. The order of applying both of these filters may affect the performance of the method. Thus, in the proposed method two different ways of applying these filters have been tested. In the first variant, the expressions obtained from the linguistic patterns are filtered using AM based filter and after this, these filtered expressions are again filtered with the help of context similarity scores between the constituent words of the expressions. The second variant is similar to the previous one but here the context similarity based filtering is applied first and then AM based filter is used for extracting the MWEs.
Association measure based filtering method (combination of linguistic pattern based filtering and AM based filtering) [2] has been used as baseline for comparing the proposed methods. CSBF method has also been considered as baseline for comparing the performance of HDF method. The detailed information about the proposed methods, threshold decision process, filtering measures and different parts of the method has been presented in the following subsections:

3.1. Obtaining expressions (N-grams) using linguistic patterns

The expressions are extracted in the form of N-grams. Since a multiword expression cannot cross sentence boundaries i.e. exist within a sentence. Thus, all the possible N-grams are extracted within the sentence boundaries to reduce the false N-grams. Next, these N-grams are filtered using the POS tag sequence of the respective N-gram satisfying the linguistic patterns. A linguistic pattern is defined as an ordered sequence of Parts of Speech tags following a specific order. The corpus is tagged in respective parts of speech using Stanford POS Tagger.
Here, three N-gram patterns and twelve bigram patterns have been used as proposed in Agrawal et al. [2]. The N-gram patterns are: Compound Noun (C-N), Noun Noun (N-N) and Verb Noun (V-N). Bigram patterns are: Adjective + Adverb (A+Adv), Adjective + Noun (A+N), Adjective + Preposition (A+P), Adverb + Adjective (Adv+A), Noun + Adjective (N+A), Noun + Preposition (N+P), Noun + Verb (N+V), Preposition + Noun (P+N), Verb + Adverb (V+Adv), Verb + Particle (V+Par), Verb + Preposition (V+P) and Verb + Verb (V+V).

3.2. Association measure based filtering (AMBF) method

Here, the expressions (N-grams) satisfying the syntactical structure of MWEs are obtained using linguistic patterns and then these N-grams are filtered using association measure (AM) scores. Here the assumption is that the higher the AM score of an expression the chance will be more for the expression to be an MWE. The expressions with higher AM scores are extracted as MWEs while the expressions with lower association scores are filtered. The boundary threshold for the association score is decided using annotated training dataset. The threshold is set on the highest f-score obtained using the manually annotated training dataset. The system calculates the f-score by taking association score of each expression as temporary boundary threshold. Finally, it sets the boundary threshold on that value of association score which provides the highest f-score on the training dataset. This final boundary threshold is used to extract MWEs further on the test dataset. The N-grams obtained by distinct linguistic patterns show different properties and frequencies. Thus, distinct thresholds are obtained for distinct types of N-grams. Two AMs, Dice's coefficient and Point wise Mutual Information (PMI) have been used. The details of these association measures are mentioned in following subsections.

3.2.1. Dice's coefficient (DC)

Dice's Coefficient has named after Lee Raymond Dice [11]. In case of multiword expressions DC is the ratio of simple frequency of any bigram in the corpus with the frequency of its constituent words. Dice's Coefficient of any bigram (w_1w_2) can be measured as:

$$DC(w_1w_2) = \frac{2f(w_1w_2)}{f(w_1)+f(w_2)} \quad (1)$$

In the above equation, $f(w_1w_2)$ is the frequency of observing the bigram (w_1 w_2). $f(w_1)$ and $f(w_2)$ are the frequencies of words w_1 and w_2 respectively. Similarly the Dice's Coefficient for N-gram (w_1 w_2 .... w_n) can be expressed as:

$$DC(w_1w_2...w_n) = \frac{n*f(w_1w_2...w_n)}{f(w_1)+f(w_2)+.....+f(w_n)} \quad (2)$$

3.2.2. Point wise/ specific mutual information (PMI)

PMI has been firstly proposed by Fano [15]. Later, Church and Hanks [8] have used it as lexical association measure. PMI is the logarithmic ratio of the probabilities of any expression and its component words. The PMI of any bigram (w_1w_2) can be calculated as:

$$PMI(w_1w_2) = \log_2 \frac{P(w_1w_2)}{P(w_1)P(w_2)} \quad (3)$$

Here, $P(w_1w_2)$ is the probability of the bigram having words w_1 and w_2. The probabilities of the single words, w_1 and w_2 are represented as $P(w_1)$ and $P(w_2)$ respectively. PMI for N-gram (w_1 w_2 .... w_n) has been calculated using the formula proposed by McInnes [28]:

$$PMI(w_1w_2...w_n) = \log_2 \frac{P(w_1w_2...w_n)}{P(w_1)P(w_2).....P(w_n)} \quad (4)$$

3.3. Context similarity based filtering (CSBF) method

The structure of CSBF method is similar to the AMBF method as described in the previous section. However, in the proposed method, the expressions obtained from linguistic patterns are filtered using context similarity scores instead of association scores. Here, the assumption is that there is a correlation between the contexts of the constituent words of MWEs i.e. the context of constituent words of MWEs are more correlated than the contexts of the component words of the normal expressions. Latent semantic anal-

ysis (LSA) method has been used for calculating the context similarity. The context similarity score is calculated using correlation coefficient. The step by step details of applying this method is as follows:

Step 1: Creating Term document matrix and reduction of term dimensions: The term document matrix is created by taking tf–idf (term frequency–inverse document frequency) of the i^{th} term in the j^{th} document as the (i, j) element of the matrix. The dimensions of this term document matrix are reduced by singular value decomposition in three matrices U, S and V according to the following expression.

$$TD = U \times S \times V^{T} \tag{5}$$

Here, S is the diagonal matrix containing the singular values of TD. According to the summation of singular values the dimensions of matrix TD are reduced up to 30%.

Step 2: Calculating context similarity score: The context similarity between two terms is calculated using correlation coefficient. Correlation calculates the strength of linear association between two terms and its value is always between –1.0 to +1.0. Positive value shows the positive relationship i.e. similarity between the contexts of two terms while the negative value shows the negative relationship i.e. dissimilarity between the contexts of two terms. The Correlation (R) between the vectors of two terms x and y can be calculated as:

$$R = \frac{N\sum_{j=1}^{N}\left(f_{x_j} f_{y_j}\right) - \sum_{j=1}^{N} f_{x_j} \sum_{j=1}^{N} f_{y_j}}{\sqrt{\left(\left(N\sum_{j=1}^{N} f_{x_j}{}^{2} - \left(\sum_{j=1}^{N} f_{x_j}\right)^{2}\right)\left(N\sum_{j=1}^{N} f_{y_j}{}^{2} - \left(\sum_{j=1}^{N} f_{y_j}\right)^{2}\right)\right)}} \tag{6}$$

Where:

N = Total number of documents.

f_{x_j} = Tf–idf of term x in j^{th} document i.e. value of (x, j) element in the TD matrix.

f_{y_j} = Tf–idf of term y in j^{th} document i.e. value of (y, j) element in the TD matrix.

Step 3: Filtering expressions (N-grams): The expressions are filtered using the correlation scores between the constituent words of any expression as calculated in step 2. Here the boundary threshold is also decided using the manually annotated training dataset for highest f-score similar to AMBF method. Different thresholds are calculated for each type of N-gram expression categorized by different linguistic patterns. Finally, these thresholds are used to extract the MWEs on the test dataset.

3.4. Hybrid dual filtering (HDF) method

The MWEs are extracted using both types of filters, AM based filter and context similarity based filter in HDF method. Two different ways of applying these filters have been proposed. The extracted expressions (N-grams) from linguistic patterns are initially filtered using AM based filter and then context similarity based filter in the first variant of HDF method. Here, the AM based filtering is applied similar to the AMBF method but there is a difference in threshold decision method. In AMBF method, the boundary threshold is calculated using highest f-score but here, AM based filter is an intermediate filter. Thus, the threshold cannot be decided on highest f-score because in this way most of the expressions would be filtered by AM based filter and the final filter, context similarity based filter would no longer be effective. The threshold should be in such a manner that some expressions should be filtered in intermediate phase for an intermediate filter and remaining expressions should be passed for the next filter. Thus, the intermediate boundary threshold is calculated on highest recall instead of highest f-score with similar process. In this way, some N-grams are filtered using AM based filter and the remaining N-grams are filtered using context similarity based filter with boundary threshold obtained using highest f-score on training dataset similar to CSBF method. Distinct thresholds have been calculated for distinct types of N-grams similar to previous methods.

The process is same for the second variant but the order of filtering has been changed. In this variant, the expressions obtained from linguistic patterns are firstly filtered using context similarity based filter with the intermediate threshold calculated on highest recall. Next, these filtered N-grams are again filtered using AM based filter and the boundary threshold is calculated on highest f-score on training dataset.

4. Results and analysis

The proposed methods have been evaluated on the same dataset proposed by Agrawal et al. [2]. Four experiments have been performed for evaluating the performance of each method. The performance has been evaluated by directly comparing the extracted list of MWEs using all the methods with the manually annotated test dataset. The performance of each method has been measured for each type of MWE categorized by different linguistic patterns in all the experiments. The overall f-score for all MWEs has also been calculated. The overall f-score in all the methods has been calculated using the total of correct MWEs extracted, total of incorrect MWEs extracted and the total of MWEs not recognized in all the categories. First two experiments have been performed for baseline methods and the other two experiments have measured the performance of both variants of proposed HDF method. In the first experiment, the performance of AMBF method has been measured in terms of f-score. Here, the training has been performed using both association measures, DC and PMI. The same AM has been used for extracting MWEs in test phase which has performed better in training phase for particular type of MWEs.

The performance of CSBF method has been evaluated in second experiment. The next set of experiments has been performed to evaluate the performance of both variants of hybrid dual filtering (HDF) methods. Here, the training has also been performed using both association measures, DC and PMI in AM based filter. The same AM has been used for extracting MWEs in test phase which has performed better in training phase for particular type of MWEs similar to AMBF method. The evaluation results of the proposed methods are presented in

Table **1**. The results are also shown as comparative chart in Fig. 1 for more clarification.

Table 1: Evaluation Results of Proposed Methods

MWE Type	AMBF	CSBF	HDF 1	HDF 2
A+Adv	0.40020	0.17647	0.40385	0.38889
A+N	0.49333	0.61789	0.63500	0.49322
A+P	0.13333	0.19192	0.21429	0.15789
Adv+A	0.02740	0.05556	0.07407	0.02667
C-N	0.77608	0.78218	0.80047	0.77794
N-N	0.29167	0.24242	0.35417	0.29231
N+A	0.52459	0.38095	0.39130	0.53571
N+P	0.08219	0.00000	0.09655	0.08642
N+V	0.16667	0.00000	0.20438	0.16667
P+N	0.17021	0.03448	0.03448	0.17647
V-N	0.23704	0.32479	0.32479	0.26087
V+Adv	0.37647	0.28571	0.39316	0.38356
V+Par	0.94241	0.94681	0.95187	0.93684
V+P	0.33898	0.14815	0.15038	0.34694
V+V	0.79310	0.72289	0.85185	0.81250
All MWEs	0.54083	0.56024	0.58835	0.50045

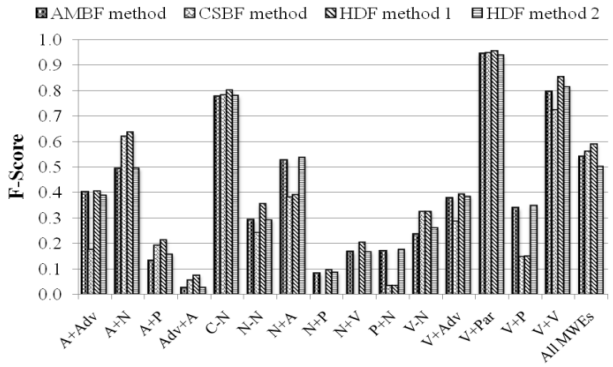

Fig. 1: Comparative Performance Chart of Proposed Methods.

It has been observed from the evaluation results that CSBF method has performed better than AMBF method for some categories. Although, for some categories AMBF method is better but the overall performance of CSBF method for all MWEs is better than the AMBF method. Hybrid Dual Filtering methods have performed better than both AMBF and CSBF methods. The first variant has performed better than the second variant in Hybrid Dual Filtering method. The final filtering has been performed by context similarity based filter in HDF1. It has been observed that context similarity based filter has extracted MWEs more accurately than the association measure based filter i.e. precision has been increased. The boundary threshold of context similarity based filter has been decided based on best recall in HDF method 2. Hence, it cannot enhance the precision in this method. Thus, the first variant of HDF method has performed better than the second variant of HDF method.

The improvement in performance using HDF method is very small for some types of multiword expressions but the performance has been improved satisfactorily for other types of multiword expressions. The performance gain for Adjective + Noun, Adjective + Preposition, Adverb + Adjective, Noun Noun, Noun + Verb and Verb + Verb MWEs is higher than other type of MWEs. The increment in performance is small for other types of MWEs but still satisfactory. Overall, HDF method has performed better than CSBF and AMBF method.

Earlier, various association measures have also been used for extracting the MWEs [14]. In their work the precision has been calculated for different sample sizes. The results have been shown for n-top listed candidates according to the association scores of the expressions. The precision has been observed between 0.30 to 0.40 for 1000 top listed expressions and about 0.20 to 0.30 for 2000 top listed expressions. In this work there is no proper method for deciding the value of n to get best results. In present work, the threshold has been decided using annotated training dataset. Agrawal et al. [2] have also worked in similar manner. They have improved the performance using Multiple Threshold method in which MWEs have been extracted in different categories. In present work it has been used as baseline. The proposed CSBF method and HDF method have performed better than these baseline methods.

5. Conclusion and future perspectives

A two phase CSBF method has been proposed for extraction of MWEs using the idea of context correlation between constituent words of MWEs. Here, the MWEs are extracted using linguistic patterns and context similarity between constituent words of the expressions. It has been analyzed that adding more filters in CSBF method may improve the performance. Thus, a three phase hybrid method has been proposed in which the MWEs are extracted using three properties of multiword expressions. These are following the linguistic pattern of expressions i.e. syntactical idiosyncrasy, association between constituent words i.e. statistical idiosyncrasy and context similarity between constituent words i.e. semantic idiosyncrasy. Besides, a two phase AMBF method as proposed in previous work [2] has also been evaluated. The comparative study of the performance of all the methods has also been presented.

Overall, CSBF method has performed better than the AMBF method but it has failed in few cases. Conversely, Hybrid Dual Filtering method which uses both, association based and context similarity based filters, has overcome the limitation of both methods and has performed best for all types of MWEs. In Hybrid Dual Filtering method, two cases have been experimented in which the case when context similarity based filtering is applied after the statistical filtering is better than the case where the order of filters is reversed. The method for choosing the cut-off boundary threshold for association measure based and context similarity based filtering has also been proposed using training dataset. Due to unavailability of standard dataset the performance of proposed methods has been measured on a manually annotated dataset. The improvements in the results, compared to the two phase methods, show the potency of proposed three phase HDF method.

The proposed methods are simple and generalized. These methods can be used for extracting all types of MWEs. Although, the performance of any generalized method is lower than the specialized method designed for any particular type of MWE. However, generalized methods are also desirable in some problems of NLP. The proposed methods are evaluated on limited manually annotated dataset but it can also give similar types of results on any corpus. The only limitation of proposed methods is that the manually annotated dataset is needed for training purpose of same domain as the whole corpus.

The solution of many problems in statistical and context similarity based filtering process has been proposed but some issues have been left to be resolved. One such issue is choosing the correct value for dimension reduction of term vector. Here, we have tested only on original term document matrix and the 30% dimension reduced term document matrix. Better performance may be obtained with the other values of dimension reduction which may also be varied for different datasets and different types of MWEs. Further, the research may be extended by solving this problem.

References

[1] Agrawal S, Jaspal A, Aggarwal A, Sanyal R & Sanyal S. (2013). Hybrid Approach: A Solution for Extraction of Domain Independent Multiword Expressions. *International Journal of Technology Innovations and Research (IJTIR)*, Vol. 5, pp. 1–16.

[2] Agrawal S, Sanyal R & Sanyal S. (2014). Statistics and linguistic rules in multiword extraction: A comparative analysis. *International Journal of Reasoning-based Intelligent Systems.* Vol. 6, No. 1/2, pp. 59–70. https://doi.org/10.1504/IJRIS.2014.063954.

[3] Baldwin T, Bannard C, Tanaka T & Widdows D. (2003). An empirical model of multiword expressions decomposability. In *Proceedings of the ACL–2003 workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pp. 89–96, Sapporo, Japan. https://doi.org/10.3115/1119282.1119294.

[4] Baldwin T. (2005). The deep lexical acquisition of English verb-particles. *Computer Speech and Language, Special Issue on Multiword Expressions*, Vol. 19, pp. 398–414.

[5] Biber D, Johansson S, Leech G, Conrad S & Finegan E. (1999). *Grammar of Spoken and Written English*, Longman, Harlow, United Kingdom.

[6] Boulaknadel S, Daille B & Aboutajdine D. (2008). A multi-word term extraction program for Arabic language. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, pp. 1485–1488, Marrakech, Morocco.

[7] Calzolari N, Fillmore CJ, Grishman R, Ide N, Lenci A, Macleod C & Zampolli A. (2002). Towards best practice for multiword expressions in computational lexicons. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC)*, pp. 1934–1940, Las Palmas, Canary Islands.

[8] Church KW & Hanks P. (1990). Word association norms, mutual information & lexicography. *Computational Linguistics*, Vol. 16, No. 1, pp. 22–29.

[9] Dahlmann I & Adolphs S. (2007). Pauses as an indicator of psycholinguistically valid multi-word expressions (mwes)? In *Proceedings of the ACL–2007 Workshop on A Broader Perspective on Multiword*

Expressions, pp. 49–56, Prague, Czech Republic. https://doi.org/10.3115/1613704.1613711.

[10] Deerwester SC, Dumais ST, Landauer TK, Furnas GW & Harshman RA. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science (JASIS)*, Vol. 41, No. 6, pp. 391-407. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9.

[11] Dice LR. (1945). Measures of the Amount of Ecologic Association between Species. *Ecology*, Vol. 26, No. 3, pp. 297–302. https://doi.org/10.2307/1932409.

[12] Duan J, Zhang M, Tong L & Guo F. (2009). A hybrid approach to improve bilingual multiword expression extraction. In *Proceedings of the 13th Pacific-Asia Conference on Knowledge Discovery and Data (PAKDD 2009)*, pp. 541–547, Bangkok, Thailand.

[13] Dubey V, Raghuwanshi P & Vyas S. (2015). Impact of Multiword Expression in English-Hindi Language. *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)*, Vol. 4, No. 3, pp. 101–105.

[14] Evert S & Krenn B. (2005). Using small random samples for the manual evaluation of statistical association measures. *Computer Speech and Language*, Vol. 19, No. 4, pp. 450–466. https://doi.org/10.1016/j.csl.2005.02.005.

[15] Fano RM. (1961). *Transmission of Information: A Statistical Theory of Communications*. MIT Press, Cambridge, Massachusetts, United States.

[16] Goldman JP, Nerima L & Wehrli E. (2001). Collocation extraction using a syntactic parser. In *Proceedings of the ACL Workshop on Collocations*, pp. 61–66, Toulouse, France.

[17] Hofmann T. (1999). Probabilistic Latent Semantic Analysis. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI'99)*, pp. 289–296, San Francisco, CA.

[18] Hurskainen A. (2008). Multiword expressions and machine translation. *Technical Report 1, Technical Reports in Language Technology*.

[19] Jackendoff R. (1997). Twistin' the night away. *Language*, Vol. 73, No. 3, pp. 534–559. https://doi.org/10.2307/415883.

[20] Justeson JS & Katz SM. (1995). Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, Vol. 1, No. 1, pp. 9–27. https://doi.org/10.1017/S1351324900000048.

[21] Karan M, Šnajder J & Bašić BD. (2012). Evaluation of classification algorithms and features for collocation extraction in Croatian. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pp. 657–662, Istanbul, Turkey.

[22] Katz G & Giesbrecht E. (2006). Automatic identification of noncompositional multi-word expressions using latent semantic analysis. In *Proceedings of the ACL–2006 workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pp. 12–19, Sydney, Australia.

[23] Kunchukuttan A & Damani OP. (2008). A system for compound nouns multiword expression extraction for Hindi. In *Proceedings of the 6th International conference on Natural Language Processing (ICON 2008)*, Pune, India.

[24] Lambert P & Castell N. (2004). Alignment of parallel corpora exploiting asymmetrically aligned phrases. In *Proceedings of the LREC 2004 Workshop on the Amazing Utility of Parallel and Comparable Corpora*, pp. 26–29.

[25] Landauer TK & Dumais ST. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, Vol. 104, No. 2, pp. 211–240. https://doi.org/10.1037/0033-295X.104.2.211.

[26] Liang Y, Tan H, Li H, Wang Z & Gui W. (2017). A language-independent hybrid approach for multi-word expression extraction. In *proceedings of International Joint Conference on Neural Networks (IJCNN)*, pp. 3273-3279, Anchorage, AK, USA.

[27] Lin D. (1999). Automatic identification of non-compositional phrases. In *Proceedings of the 37th Association of Computational Linguistics (ACL-1999)*, pp. 317–324, College Park, Maryland, USA. https://doi.org/10.3115/1034678.1034730.

[28] McInnes BT. (2004). Extending the loglikelihood measure to improve collocation identification. *Master thesis*, University of Minnesota, USA.

[29] Moirón BV & Tiedemann J. (2006). Identifying idiomatic expressions using automatic word alignment. In *Proceedings of the EACL–2006 workshop on Multiword Expressions in a multilingual context*, pp. 33–40, Trento, Italy.

[30] Monti J, Barreiro A, Elia A, Marano F & Napoli A. (2011). Taking on new challenges in multiword unit processing for machine translation. In *Proceedings of the Second International Workshop on Free/Open-Source Rule-Based Machine Translation*, pp. 11–19, Barcelona, Spain.

[31] Pearce D. (2001). Synonymy in collocation extraction. In *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, pp. 41–46, Pittsburgh, Pennsylvania, USA.

[32] Piasecki M, Wendelberger M & Maziarz M. (2015). Extraction of the Multiword Lexical Units in the Perspective of the Wordnet Expansion. In *Proceedings of Recent Advances in Natural Language Processing*, pp. 512–520, Hissar, Bulgaria.

[33] Ramisch C. (2012). A Generic Framework for Multiword Expressions Treatment: from Acquisition to Applications. In *Proceedings of the ACL 2012 Student Research Workshop*, pp. 61–66, Jeju Island, Korea.

[34] Sag IA, Baldwin T, Bond F, Copestake A & Flickinger D. (2002). Multi-word expressions: A pain in the neck for nlp. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002), Vol. 2276 of Lecture Notes in Computer Science*, pp. 1–15, London, UK. https://doi.org/10.1007/3-540-45715-1_1.

[35] Schone P & Jurafsky D. (2001). Is knowledge-free induction of multiword unit dictionary headwords a solved problem? In *Proceedings of the 6th conference on Empirical Methods in Natural Language Processing (EMNLP-2001)*, pp. 100–108, Hong Kong.

[36] Seretan V. (2011). A collocation-driven approach to text summarization. In *Proceedings of the Traitement Automatique des Langues Naturelles (TALN 2011)*, pp. 9–14, Montpellier, France.

[37] Singh A & Jamwal SS. (2016). Identification, Extraction and Translation of Multiword Expressions. *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 6, No. 7, pp. 445–449.

[38] Smadja F. (1993). Retrieving collocations form text: Xtract. *Computational Linguistics*, Vol. 19, No. 1, pp. 143–177.

[39] Strik H, Binnenpoorte D & Cucchiarini C. (2005). Multiword expressions in spontaneous speech: Do we really speak like that? In *Proceedings of the Interspeech-2005 (IS–2005)*, pp. 1161–1164, Lisbon, Portugal.

[40] Tsvetkov Y & Wintner S. (2012). Extraction of multi-word expressions from small parallel corpora. *Natural Language Engineering*, Vol. 18, No. 4, pp. 549–573. https://doi.org/10.1017/S1351324912000101.

[41] Tutubalina E & Braslavski P. (2016). Multiple Features for Multiword Extraction: A Learning to Rank Approach. In *Proceedings of the International Conference on Computational Linguistics and Intellectual Technologies: "Dialogue 2016"*, pp. 782–793, Moscow, Russia.

[42] Vechtomova O. (2005). The role of multi-word units in interactive information retrieval. In *Proceedings of the Advances in Information Retrieval, 27th European Conference on IR Research (ECIR–2005)*, pp. 403–420, Santiago de Compostela, Spain.

[43] Venkatapathy S, Agrawal P & Joshi AK. (2005). Relative compositionality of Noun + Verb multiword expressions in Hindi. In *Proceedings of the International Conference on Natural Language (ICON)*, pp. 37–44, Kanpur, India.

[44] Vintar S & Fiser D. (2008). Harvesting multi-word expressions from parallel corpora. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, pp. 1091–1096, Marrakech, Morocco, 2008.