HCTL Open Int. J. of Technology Innovations and Research
HCTL Open IJTIR, Volume 5, September 2013
e-ISSN: 2321-1814
ISBN (Print): 978-1-62840-986-4

Hybrid Approach: A Solution for Extraction of Domain Independent Multiword Expressions

Shaishav Agrawal, Amit Jaspal, Ankit Aggarwal, Ratna Sanyal and Sudip Sanyal*
{*shaishav.engr, amitjaspal007, ankagg56*}*@gmail.com*, {*rsanyal, ssanyal*}*@iiita.ac.in*

Abstract

A hybrid technique for extraction of bigram and trigram multiword expressions is presented. This technique works in two phases as first statistical technique is applied to filter the extracted bigrams and trigrams from English text, and after it multiword expressions are extracted from this list using some linguistic rules. Two methods for threshold decision in statistical technique are also presented, first is by minimizing the error in classification, and the second is based on maximizing the recall value.

Keywords

Multiword expressions, information retrieval, statistical methods, natural language processing, collocation extraction.

*Indian Institute of Information Technology Allahabad, India

HCTL Open Int. J. of Technology Innovations and Research
HCTL Open IJTIR, Volume 5, September 2013
e-ISSN: 2321-1814
ISBN (Print): 978-1-62840-986-4

Introduction

An MWE is a single lexeme which is the association of two or more independent lexemes. The properties or semantic meaning of a multiword expression cannot be predicted by the properties or semantic meanings of the normal combination of independent lexemes [1, 2]. Consider the expression:

Kick the bucket

Here the simple combination of semantic meanings of individual lexemes is "to hit the bucket by one foot". But the actual meaning "to die" is far behind the previous meaning. Multiword expressions are widely used in text and speech [3, 4]. MWEs are also used in many Natural Language Processing applications such as Alignment of Parallel Corpora [5, 6], Information Retrieval [7, 8], Machine Translation [9, 10, 11], Speech Recognition [12, 13], Text Summarization [14, 15] etc. This is the prime reason that makes the analysis of MWEs more important.

In our work we have derived a hybrid approach which uses statistical property and linguistic property of text expressions to classify these expressions as MWEs. Statistical property deals with the statistics of text expressions. The idea behind this is, MWEs show high statistical scores than general expressions. MWEs also follow some linguistic pattern i.e. noun – noun, verb – noun, verb – particle, etc. We have used Dice's coefficient statistics and some linguistic rules to extract the MWEs. In most of the works using statistical techniques the list of n best candidates is extracted [16] but there is no proper method to decide the cut-off threshold. We have also proposed two methods for deciding the cut-off threshold by minimizing the error in classification and by maximizing the recall value. In this paper the proposed approach with experimental results is presented.

The rest of the paper is organized as follows. Next section briefly describes the related work and various MWE extraction techniques. In methodology section we describe our proposed methodology. Results and Analysis section presents the experimental results & analysis. Finally last section concludes the paper with the discussion about future scope.

Related Work

As per the literature found, the extraction techniques for MWEs can be broadly classified into four types. Statistical methods [17] in which the MWEs are extracted using statistical measures. Statistical methods are easy to apply on bigram and trigram multiword expressions but difficult to apply on more than three words MWEs. The other difficulty of this approach is that there

e-ISSN: 2321-1814
ISBN (Print): 978-1-62840-986-4

is no proper method to decide the cut-off threshold of any statistical measure. Symbolic or linguistic methods [18], which use the linguistic rules and morpho-syntactic patterns of text for extracting the MWEs. These methods give good accuracy but the main limitation is that there is a need of large annotated corpus. Hybrid methods [19, 20], which use both statistical measures and linguistic filters. In these methods first multiword expressions are classified using some statistical measure and then linguistic rules are applied to filter the MWEs for good precision. These methods give better accuracy over any single method but also have the limitations of both the methods. Word Alignment methods [21] can be used for extracting MWEs from one language to another language. If the multiword expressions are well classified in one language then with the help of parallel corpora using word alignment method MWEs for another language can be classified easily. This method gives good results on similar type of languages but lacks in performance if the languages are different in nature.

Generally, multiword expressions are identified on the basis of idiosyncrasies exhibited by these words. A novel approach for finding the compound noun MWEs has been derived by Kunchukuttan and Damani [22]. The authors have used the candidate extraction and ranking phenomena for MWEs. Other research works [17, 23] have used lexical substitution to calculate the difference between the distributional characteristics of one collocation and other similar collocations for identifying MWE. They proposed, if in some collocations one word is similar then these collocations can be called similar collocations and if one collocation is identified as an MWE then others can be extracted as MWEs *e.g. traffic signal, traffic sign and traffic light.* Latent semantic analysis is also used to compare the similarity between an MWE and its component words [24, 25]. The automatic word alignment technique for identifying idiomatic expressions is proposed by Moirón and Tiedemann [21]. In this method they have used two criteria: first is meaning predictability measured as semantic entropy and second is the overlap between the contextual meaning of a phrasal expression and the combination of the semantic meaning of its component words. A good work for extracting Noun – Verb MWEs in Hindi is found in Venkatapathy et al. [26]. They proposed an approach to measure the relative compositionality of Noun – Verb expressions automatically using maximum entropy model (MaxEnt). Some researchers have designed lexical recourses which are very useful in extracting MWEs. Chakrabarti et al. [27] has been designed the Indo Wordnet to extract words having reduplication in Hindi corpora. The author has also proposed the various methods for detecting MWEs especially in Hindi. Similarly the approach for finding reduplication in Bengali is proposed [28].

e-ISSN: 2321-1814
ISBN (Print): 978-1-62840-986-4

Proposed approach filters the MWEs extracted by statistical method with the help of linguistic rules. For different languages linguistic rules may be different. Even for a single language there are no such strict rules that can define the multiword expressions completely. This is due to the idiosyncrasy inhibited with the MWEs. Some researchers have given some rules in their work, for example, Collocation Extraction using a Syntactic Parser [29] follows the rules as N-N, N-V, N-A, N-P-N, V-P-N, V-N, V-P; Collocation extraction system "Xtract" [30] follows the rules as N-D, N-A, V-N, N-V, V-Adv, N-P, V-V, V-P; the BBI Dictionary of English Word Combinations [31] follows the rules as V-Adv, N-A, N-P-N, V-N, N-V, A-Adv; Word Sketch system [32] follows the rules as N-Conj-N, V-N, N-A, N-P-N, N-N, N-V, V-P, V-A, A-P; and A-Adv, N-V, N-A, V-N, V-Adv, N-[P]-N rules are given in Le dictionnaire de collocations [33]. The rules used in our proposed work are influenced with these rules.

Methodology

We have used the hybrid approach. The stepwise approach is described below and the flow diagram is shown in Figure 1.

Step 1: In first step the document is fragmented into bigrams and trigrams.

Step 2:A list of patterns is identified which could never be a part of an MWE. These include e-mail, url of website, dates, abbreviations and posts. Consequently, for each one of them, a regular expression is formed which could satisfactorily be searched in the corpus and hence these patterns are removed from the input text in the preprocessing stage only.

Step 3: Term frequency and the frequency of each bigram and trigram is calculated from the corpus.

Step 4: The statistical measure of each bigram is calculated with the help of the frequencies of previous step.

Step 5: The MWEs are extracted from the list with the cut-off boundary threshold. Here boundary threshold is the minimum acceptable statistical score of any expression to be an MWE.

Step 6: MWEs extracted from the statistical measure are filtered by linguistic rules.

Statistical Filtering and Threshold Decision

The bigrams and trigrams are filtered using statistical technique. Dice's co-efficient is used to calculate the statistical scores. It measures the ratio of the frequency of the combined occurrence of two constituent words with the

e-ISSN: 2321-1814
ISBN (Print): 978-1-62840-986-4

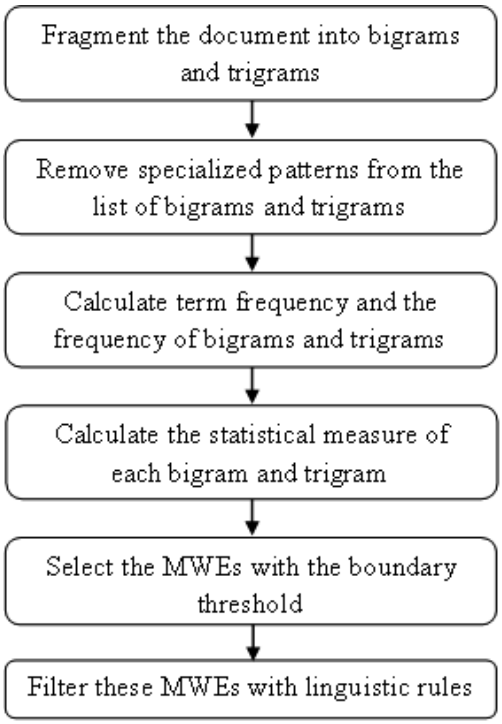

Figure 1: *Flow diagram of proposed approach*

frequency of individual words. The value of Dice's coefficient for a bigram (w_1w_2) can be calculated as:

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

Here, $f(w_1w_2)$ is the frequency of observing the bigram consisting of the words w_1 and w_2. The frequency of observing a single word w_1 is given by $f(w_1)$. Similarly Dice's Coefficient for trigram $(w_1w_2w_3)$ can be calculated as:

$$DC(w_1w_2w_3) = \frac{3f(w_1w_2w_3)}{f(w_1) + f(w_2) + f(w_3)} \tag{2}$$

Two methods are used to decide the value of threshold based on minimum error and maximum recall respectively:

e-ISSN: 2321-1814
ISBN (Print): 978-1-62840-986-4

Minimizing Error in Classification

In this method the value of threshold is decided on the basis of Error in Classification which is defined as:

$$\textit{Error in Classification} = \textit{False Positive } (F_P) + \textit{False Negative } (F_N)$$

False Positive (F_P) is also known as "Type-I Error", "α Error" or "Error of first kind". It is basically an error of rejecting a null hypothesis which is actually true. It occurs when any system observes a difference but in reality there is none. Here, False Positive (F_P) means that the N-gram is classified as MWE but actually it is not.

False Negative (F_N) is also known as "Type-II Error", "β Error" or "Error of second kind". It is basically an error of failing to reject a null hypothesis which must be rejected actually. It occurs when any system fails to observe a difference but in reality there is one. Here, False Negative (F_N) means that the N-gram is not classified as MWE but it is actually an MWE.

Minimizing the Type-I Error will increase the precision while minimizing the Type-II Error will increase the recall. Thus the value of threshold should be decided on the value which minimizes the sum of both i.e. "Error in Classification". It gives the balanced result of precision and recall.

Maximizing Recall value

Here the threshold is set on the maximum value of recall. The idea is, as our approach is hybrid approach in which the MWEs are filtered in two phases. So in first phase precision can be compromised but the recall should be highest. In this way all the MWEs are extracted and the precision will be gained by filtering irrelevant expressions using linguistic filtering technique.

Linguistic Filtering

For filtering the bigrams and trigrams extracted from statistical technique, we use some linguistic rules. A linguistic rule is a pattern of POS tags occurring in an order that have a high chance of being an expression as MWE. For this, we have analyzed the pattern of N-grams formed by our system to derive the linguistic rules. We have also used the rules found in lexicographic dictionaries and derived by other researchers. This led us to take a decision on the set of rules involving POS (Part of Speech) for both bigrams and trigrams which are most probable to be an MWE. Out of the various rules identified, the rules for bigrams used by our system are: Noun – Noun, Verb – Noun, Adjective –

e-ISSN: 2321-1814
ISBN (Print): 978-1-62840-986-4

Noun, Verb – Preposition, Verb – Adverb, and Noun – Preposition. While the rules for trigrams consist of: Adjective – Adjective – Noun, Adjective – Noun – Noun, Noun – Adjective – Noun, Noun – Noun – Noun, and Noun – Preposition – Noun.

For filtering bigrams and trigrams using these rules, Rita Wordnet[1] is used for annotating the corpus. It has an inbuilt WordNet and provides various utility functions to make calls on the WordNet. However, any English WorldNet is able to classify a token as only one of the following: noun, adjective, adverb, and verb. But in our rules, there is a need to identify the tokens as prepositions also. So a manually prepared list of prepositions is used and before searching a token for finding its POS in WordNet, system searches the token in the list of prepositions to check whether it is a preposition or not. If it is not, then only the WordNet is used. Secondly, a word or token can be used in many different contexts. Consequently, it can have variety of meanings and hence variety of POS. Thus, for each token, WordNet produce a list of POS. However, the POS of synset having highest frequency in WordNet can be assumed the most generic sense of that token. Hence only the POS of highest frequency synset returned by WordNet is considered for any single token to be its POS. Once the POS of all the tokens have been found, the list of rules is used to find a match between the rules and the order of POS of the tokens of the N-gram. If there is a successful match, system classifies the bigram or trigram as an MWE otherwise filter it.

Results and Analysis

The proposed approach is evaluated on the manually created corpus containing articles from four different News Papers. The articles are not domain specific and based on daily News. From each Newspaper the news of 6 consequent days has taken for corpus building.

On applying our base model to the corpus, we obtained a list of about 0.425 million N-grams (bigrams and trigrams) for which the score of statistical measure (i.e. the Dice's coefficient in our case) comes in a very wide range. It is very difficult to check error on each value of threshold. Thus to obtain the threshold following procedure has performed:

The mean value of all the Dice's coefficient scores for all the bigrams and trigrams is calculated, which is a constant value. This is taken as the starting point of our threshold analysis. For this we begin our threshold analysis by taking mean score as first threshold value and decrease the threshold progressively by

[1] Available at, http://www.rednoise.org/rita/wordnet/RiTaWN.zip

e-ISSN: 2321-1814
ISBN (Print): 978-1-62840-986-4

a factor of 10. In this way the analysis on the entire range of statistical measure is divided into 9 iterations, which is shown in Table 1. For each threshold value, F_P (No. of false positives), F_N (No. of false negatives), the Error in Classification, T_P (No. of true positives), T_N (No. of true negatives), Precision, Recall, and F-Score is calculated. For this cause, we used a list of MWEs obtained from Princeton Wordnet[2] which listed around 68,000 MWEs as the Gold Standard.

Although the Princeton Wordnet list of MWEs is not a complete list but for the purpose of analysis, decision making on the value of threshold, and to check the feasibility and correctness of the threshold algorithm and statistical measure, this list is used. In the later part of our final analysis, manually annotated list of multiword expressions is used for analysis instead of Princeton Wordnet list, as it leaves out many of the standard MWEs like "brook no truck", "master blaster", etc.

Table 1: *Analysis performed for obtaining threshold value*

Threshold	F_P	F_N	Error in Classification	T_P	T_N	Precision	Recall	F-Score
1.695E–01	3191	699	3890	454	152150	0.1246	0.3938	0.1892
1.695E–02	13345	622	13967	531	141996	0.0383	0.4605	0.0707
1.695E–03	33896	551	34447	602	121445	0.0175	0.5221	0.0338
1.695E–04	67783	462	68245	691	87558	0.0101	0.5993	0.0198
1.695E–05	120459	129	120588	1024	34882	0.0084	0.8881	0.0167
1.695E–06	151456	14	151470	1139	3885	0.0075	0.9879	0.0148
1.695E–07	155166	0	155166	1153	175	0.0074	1.0000	0.0146
1.695E–08	155337	0	155337	1153	4	0.0074	1.0000	0.0146
1.695E–09	155341	0	155341	1153	0	0.0074	1.0000	0.0146

The Result obtained is as follows:

Mean Value = 1.695E-01
Minimum Error = 3890
Maximum Recall = 1.0
Optimal Threshold by minimizing Error = 1.695E-01
Optimal Threshold by maximizing Recall = 1.695E-07

Consequently, the minimum error is found at the first threshold value i.e. mean value. However, the value of recall and precision at this value is quite low (Table 1). And the maximum recall occurs on three values (1.695E-07,

[2]WordNet 3.0, http://wordnetcode.princeton.edu/3.0/WordNet-3.0.tar.gz

e-ISSN: 2321-1814
ISBN (Print): 978-1-62840-986-4

1.695E-08, and 1.695E-09), so we consider the first value where the error is minimal among the three.

Thus in order to increase precision, it is required to use some linguistic filtering techniques which will somehow identify and separate only the strongest possible MWEs from among the list of all possible candidate MWEs.

For evaluating the results of the filtering technique, we used the methodology described in Evert and Krenn [34]. For this we randomly took a sample of 10,000 N-grams (bigrams and trigrams) that are obtained from our base model i.e. 10,000 out of a list of 0.425 million candidate MWEs. We manually tagged these 10,000 N-grams to decide amongst them which are genuine MWEs and which are not. Consequently, we got 252 N-grams as MWEs. The result of our proposed approaches is shown in Table 2.

Table 2: *Results obtained for proposed approaches*

		Predicted MWEs by system	Rejected MWEs by system	Precision	Recall	F-Score
Minimizing Error in Classification Method	MWEs	$T_P = 94$	$F_N = 158$	0.2938	0.3730	0.3287
	Non MWEs	$F_P = 226$	$T_N = 9522$			
Maximizing Recall Method	MWEs	$T_P = 198$	$F_N = 72$	0.2672	0.7333	0.3917
	Non MWEs	$F_P = 543$	$T_N = 9187$			

The overall accuracy is calculated in terms of precision, recall, and f-score. The f-score for "Minimizing Error in Classification Method" is 0.3287 and for "Maximizing Recall Method" is 0.3917. But before linguistic filtering the f-score of "Minimizing Error in Classification Method" is higher than "Maximizing Recall Method". This is because in "Minimizing Error in Classification Method" most of the positive MWEs are not extracted for next phase while in second method all the positive MWEs are taken for the next phase, although by linguistic filtering false positives are increased but true positives are also increased so better precision, and recall is occurred. Thus "Maximizing Recall Method" for threshold decision is performing better than "Minimizing Error in Classification Method" in two phase hybrid approach.

The concept behind using any statistical technique is that N-grams having high frequencies are the most probable candidates for being MWEs i.e. the words occurring in a recurring pattern can be an MWE. Thus the combination of words occurring together very frequently is the claim to be an MWE. In

e-ISSN: 2321-1814
ISBN (Print): 978-1-62840-986-4

Table 3: *Sample output of simple frequency count*

N-gram	Frequency	N-gram	Frequency
the spicy	19943	to suppress	12190
the roots	18975	the spacecraft	11987
the urban	16906	of college	11876
the fetus	16108	to real	11654
the spine	15643	and robbery	10987
the incentive	15390	in crime	10872
of loving	14562	a technician	9821
the senses	14356	in training	9432
and tomatoes	13345	a fact	9123
to mountains	13224	a further	8761
to elaborate	12340	was well	8650

this way, if any N-gram co-occurs relatively high frequency than others can be classified as multiword expression. But the flaw with this statistic can be observed in Table 3. N-grams formed with stop words like articles (the, a, an), prepositions (of, up, off etc), conjunctions, etc. tend to have high frequency which suppresses the actual MWEs. But Dice's coefficient is the ratio of the frequency of the combined occurrence of two constituent words with the frequency of individual words. Thus the effect of increased frequency of stop words is now mitigated due to the introduction of the denominator part which acts as normalization operator, producing the desired result. So with this measure, the flaw of simple frequency measure is corrected. Hence we used Dice's coefficient measure instead of simple frequency.

Conclusion

This paper basically presents our technique, experimental results and analysis on improving the performance in extraction of multiword expressions. The best f-score of our system is 39.17% by "Maximizing Recall Method" for threshold decision, which is not a very good accuracy but the proposed approach is an unsupervised approach and can be used in any domain. If multiword expressions in particular domain are extracted like VNC (Verb Noun Constructions), Compound Noun MWEs, Verb Particle MWEs, Phrasal Verb MWEs etc, then there are many approaches have identified with better results but in very general domain there is no much work present with such a good accuracy. Despite a low recall value, the result of linguistic filtering technique indicates the fact

e-ISSN: 2321-1814
ISBN (Print): 978-1-62840-986-4

that MWEs definitely have some linguistic (POS, etc.) relationship.

The main reason of low recall value in first method is the decision of threshold value used in the statistical technique. As if threshold value is decided for high recall value then it decreases the precision value as it identifies many wrong MWEs. But if the linguistic rules are very strong then threshold value can be set for high recall value and the precision will be gained by linguistic technique. Another factor is, the corpus used is not domain specific i.e. it is a mixture of various topics and hence, the frequency of same type multiword expressions in it is not very high which is an essential part of any statistical approach. So on any other ideal domain specific corpus this approach can show dramatic increase in recall and precision values.

The base foundation for the development of MWE classifier has been presented in this paper. In future, the efficiency of this approach can be improved by using more effective linguistic filtering techniques besides the one mentioned in the paper. Moreover, if other modules of Natural Language Processing like Named Entity Recognition (NER), Word Sense Disambiguation (WSD) are added to this then the performance and accuracy of the system will definitely be increased. For identifying the POS of the words good POS Tagger can be applied instead of Rita Wordnet. Also, clausal analysis can be done to identify phrases separated by comma and other punctuation marks to delineate the boundary for N-gram formation. This will lead to formation of better N-grams from the corpus. We have also observed that there is a great variance in the statistical scores of different types of MWEs such as Noun – Noun MWEs, Verb – Noun MWEs, Adjective – Noun MWEs, Verb – Preposition MWEs, Verb – Adverb MWEs, Noun – Preposition MWEs, etc. In the proposed technique a common single threshold is used for filtering N-grams. In future separate thresholds can be used for filtering different types of MWEs to achieve better accuracy.

Acknowledgement

The authors gratefully acknowledge support from the Indian Institute of Information Technology, Allahabad, India; and the Technology Development for Indian Languages, Ministry of Communications and Information Technology, Government of India.

References

[1] I. A. Sag, T. Baldwin, F. Bond, A. Copestake, and D. Flickinger, "Multiword 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*, (London, UK), pp. 1–15, 2002.

[2] I. Dahlmann and S. Adolphs, "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*, (Prague, Czech Republic), pp. 49–56, June 2007.

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

[4] R. Jackendoff, "Twistin' the night away," *Language*, vol. 73, pp. 534–559, September 1997.

[5] S. S. Piao and T. McEnery, "Multi-word unit alignment in english-chinese parallel corpora," in *Proceedings of the corpus linguistics*, (Lancaster, UK), pp. 466–475, 2001.

[6] P. Lambert and N. Castell, "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, 2004.

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

[8] W. Zhang, T. Yoshida, and X. Tang, "Tf–idf, lsi and multi-word in information retrieval and text categorization," in *Proceedings of IEEE International Conference on Systems, Man and Cybernetics (SMC 2008)*, (Singapore), pp. 108–113, 2008.

[9] P. Lambert and R. Banchs, "Data inferred multi-word expressions for statistical machine translation," in *Proceedings of Machine Translation Summit X*, (Phuket, Thailand), pp. 396–403, 2005.

[10] A. Hurskainen, "Multiword expressions and machine translation," Tech. Rep. 1, Technical Reports in Language Technology, 2008.

e-ISSN: 2321-1814
ISBN (Print): 978-1-62840-986-4

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

[12] D. Binnenpoorte, C. Cucchiarini, L. Boves, and H. Strik, "Multiword expressions in spoken language: An exploratory study on pronunciation variation," *Computer Speech and Language*, vol. 19, pp. 433–449, October 2005.

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

[14] K. Zechner, "Automatic summarization of open-domain multiparty dialogues in diverse genres," *Computational Linguistics*, vol. 28, pp. 447–485, December 2002.

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

[16] S. Evert, "A lexicographic evaluation of german adjective-noun collocations," in *Proceedings of the LREC workshop towards a Shared Task for Multiword Expressions (MWE 2008)*, (Marrakech, Morocco), pp. 3–6, 2008.

[17] T. V. d. Cruys and B. V. Moirón, "Lexico-semantic multiword expression extraction," in *Proceedings of the 17th Meeting of Computational Linguistics in the Netherlands (CLIN)* (P. Dirix *et al.*, eds.), (Leuven, Belgium), pp. 175–190, 2007.

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

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

e-ISSN: 2321-1814
ISBN (Print): 978-1-62840-986-4

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

[21] B. V. Moirón and J. Tiedemann, "Identifying idiomatic expressions using automatic word alignment," in *Proceedings of the EACL-2006 workshop on Multiword Expressions in a multilingual context*, (Trento, Italy), pp. 33–40, April 2006.

[22] A. Kunchukuttan and O. P. Damani, "A system for compound nouns multiword expression extraction for hindi," in *Proceedings of 6th International conference on Natural Language Processing (ICON 2008)*, (Pune, India), 2008.

[23] T. V. d. Cruys and B. V. Moirón, "Semantics-based multiword expression extraction," in *Proceedings of the ACL-2007 Workshop on A Broader Perspective on Multiword Expressions*, (Prague, Czech Republic), pp. 25–32, 2007.

[24] T. Baldwin, C. Bannard, T. Tanaka, and D. Widdows, "An empirical model of multiword expressions decomposability," in *Proceedings of the ACL-2003 workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, (Sapporo, Japan), pp. 89–96, 2003.

[25] G. Katz and E. Giesbrecht, "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*, (Sydney, Australia), pp. 12–19, 2006.

[26] S. Venkatapathy, P. Agrawal, and A. K. Joshi, "Relative compositionality of noun+verb multi-word expressions in hindi," in *Proceedings of the International Conference on Natural Language (ICON 2005)*, 2005.

[27] D. Chakrabarti, D. K. Narayan, P. Pandey, and P. Bhattacharyya, "Experiences in building the indo wordnet – a wordnet for hindi," in *Proceedings of International Conference on Global WordNet (GWC 02)*, (Mysore, India), January 2002.

[28] T. Chakraborty and S. Bandyopadhyay, "Identification of reduplication in bengali corpus and their semantics analysis: A rule based approach," in *Proceedings of the Multiword Expressions: From Theory to Applications (MWE2010)*, (Beijing, China), pp. 73–76, August 2010.

[29] J. P. Goldman, L. Nerima, and E. Wehrli, "Collocation extraction using a syntactic parser," in *Proceedings of the ACL Workshop on Collocations*, (Toulouse, France), pp. 61–66, 2001.

[30] F. Smadja, "Retrieving collocations form text: Xtract," *Computational Linguistics*, vol. 19, pp. 143–177, March 1993.

[31] M. Benson, E. Benson, and R. Ilson, *The BBI Dictionary of English Word Combinations*. Amsterdam, Philadelphia: John Benjamins Publishing Company, 1986.

[32] A. Kilgarriff, P. Rychly, P. Smrz, and D. Tugwell, "The sketch engine," in *Proceedings of the Eleventh EURALEX International Congress*, (Lorient, France), pp. 105–116, 2004.

[33] F. J. Hausmann, "Le dictionnaire de collocations," in *Wörterbücher: ein Internationals Handbuch zur Lexikographie*, (de Gruyter, Berlin), pp. 1010–1019, 1989.

[34] S. Evert and B. Krenn, "Using small random samples for the manual evaluation of statistical association measures," *Computer Speech and Language*, vol. 19, pp. 450–466, October 2005.

Authors

Shaishav Agrawal has received Bachelors of Technology in computer science from Uttar Pradesh Technical University, Lucknow, India in 2007. He has completed Masters in Technology from Indian Institute of Information Technology Allahabad, India in 2009. He has served Lovely Professional University, Jalandhar, India as assistant professor from 2009 to 2010. Presently he is pursuing his doctoral degree from Indian Institute of Information Technology Allahabad. His research areas are: Information Retrieval, Natural Language Processing, Image Processing, and Speech Emotion Recognition.

Amit Jaspal has received Bachelor of Technology in Information Technology from Indian Institute of Information Technology, Allahabad, India in 2011. Currently he is pursuing Masters in Computer Science in University of Illinois at Urbana

e-ISSN: 2321-1814
ISBN (Print): 978-1-62840-986-4

Champaign. His area of expertise includes Artificial Intelligence, Data Mining and Information Retrieval. He has worked as a member of Technical Staff in D.E.Shaw & Co. from Sep 2011 to Jul 2013.

Ankit Aggarwal has received Bachelor of Technology in Information Technology from Indian Institute of Information Technology, Allahabad, India in 2011. His area of expertise includes Artificial Intelligence, Data Mining and Information Retrieval. Currently he is serving as a member of Technical Staff in Adobe Systems Software Ltd. under the team Adobe Illustrator.

Dr. Ratna Sanyal has received her PhD degree in 1989 from Banaras Hindu University and she has been doing research in the areas of Machine Translation, Document Summarization, Software Requirement Engineering and issues related to Digital Library. She has more than 23 years of teaching experience. Dr. Ratna has published/presented more than 50 papers in international journals and conferences and given 20 invited talks. She is the member of editorial board and reviewer of international journal. She is also the member of program committee and reviewer of many international conferences.

Prof. Sudip Sanyal has received his PhD degree in 1987 from Banaras Hindu University. He has more than 25 years of experience in teaching and research. His research interests are in the areas of Statistical Machine Translation, Pattern Recognition including Optical Character Recognition, Bio-medical Engineering, etc. Prof. Sanyal is also the member of program committee and reviewer of many international conferences. He has published/presented more than 60 papers in international journals and conferences.