RESEARCH ARTICLE OPEN ACCESS

Identification of Multiword Expressions: A Comparative Literature Study

Jisha T E [1], Thomas Monoth [2]
Department of Computer Science
Mary Matha Arts & Science College, Mananthavady,Wayanad
Kerala - India

ABSTRACT

A multiword expression (MWE) is a lexeme made up of a sequence of two or more lexemes that has properties which are not predictable from the properties of the individual lexemes or their normal mode of combination. MWEs play an inevitable role in the applications of Natural Language Processing and Computational Linguistics. This paper presents a study and analysis of types, structures and key problems related to the MWEs. Also this paper describes a comparative literature study of methodologies and associated measures to recognize MWEs, have been featured. MWEs constitute an enormous problem to unambiguous language processing due to their idiosyncratic nature and diversity of their semantic, lexical, syntactic, pragmatic and/or statistical properties.

Keywords :— Multiword Expressions, Natural language processing, PMI, Idiomaticity.

I. INTRODUCTION

In recent years Multiword Expressions have attained an abundant attention in Computational Linguistics and Natural Language processing applications like Machine translation, Named entity recognition (NER), Natural language generation, Natural language understanding, Optical character recognition (OCR), Part-of-speech tagging, Question answering, Sentence breaking or sentence boundary disambiguation, Speech recognition, Speech, topic and word segmentation etc. All these related tasks are grouped into subfields of NLP that are often considered as Information retrieval (IR), Information extraction (IE), Speech processing etc. Multi-Word expressions are those whose structure and meaning cannot be derived from their component words, as they occur independently. The term MWE has been used to refer to various types of linguistic units and expressions including idioms like *'kick the bucket'* ('to die'), noun compounds such as *'village community'*, phrasal verbs like *'find out'* ('search') and other habitual collocations (like conjunction e.g. *'as well as'* etc) [3].

They can be defined roughly as idiosyncratic interpretations that cross word boundaries [1].

The major NLP tasks relating to MWEs are: (1) identifying and extracting MWEs from corpus data, and disambiguating their internal syntax, and (2) interpreting MWEs. Increasingly, these tasks are being pipelined with parsers and applications such as machine translation. Identification is the task of determining individual occurrences of MWEs in running text. In MWE identification, a key challenge is in differentiating between MWEs and literal usages for word combinations such as make a face which can occur in both usages (Kim made a face at the policeman [MWE] vs. Kim made a face in pottery class [non-MWE]) [4]. Apart from the problem of identifying clear boundaries that distinguish MWEs from free word combinations, MWEs pose such difficult problems for computational processing so that Sag et al. (2002) call MWEs "a pain in the neck for NLP"[1] and Villavicencio et al. (2005) says MWEs " having a crack at a hard nut"[2]. The rest of the paper is organized as follows. In the next section, we describe a study on structure and classification of MWEs. Section III describe methods used for identifying MWEs, section IV highlight the contributions from the articles with a variety of proposals and approaches for handling the identification of MWEs, section V describes a comparative study of the articles and section V draws the conclusion and the future works road map.

II. STUDIES ON MWEs

The better understanding of MWEs is crucial for NLP.

A. Linguistic properties of MWEs

In languages such as English, the conventional interpretation of the requirement of decomposability into lexemes is that MWEs must in themselves be made up of multiple whitespace-delimited words. For example, marketing manager is potentially a MWE as it is made up of two lexemes (marketing and manager), while fused words such as lighthouse are conventionally not classified as MWEs. The ability to decompose an expression into multiple lexemes is still applicable, however, and leads to the conclusion, e.g. that "*compound expression*" is a MWE ("compound" and "expression" are standalone lexemes), but "*department head*" is not ("department" is a standalone lexeme, but "head" is not). The second requirement on a MWE is for it to be idiomatic.

In the context of MWEs, idiomaticity refers to markedness or deviation from the basic properties of the component lexemes, and applies at the lexical, syntactic, semantic, pragmatic, and/or statistical levels. Lexical idiomaticity occurs when one or more components of an MWE are not part of the conventional English lexicon. For example, *ad hoc* is lexically marked in that neither of its components (ad and hoc) are standalone English words. Syntactic idiomaticity occurs when the syntax of the MWE is not derived directly from that of its components, for example, *by and large*. Semantic idiomaticity is the property of the meaning of a MWE not being explicitly derivable from its parts. For example, *middle of the road* usually signifies "non-extremism, especially in political views", which we could not readily predict from either middle or road. Pragmatic idiomaticity is the condition of a MWE being associated with a fixed set of situations or a particular context. *Good morning* and *all aboard* are examples of pragmatic MWEs. Statistical idiomaticity occurs when a particular combination of words occurs with markedly high frequency, relative to the component words or alternative phrasings of the same expression.

B. *Other Properties of MWEs*

Other common properties of MWE are: single-word paraphrasability, proverbiality and prosody. Unlike idiomaticity, where some form of idiomaticity is a necessary feature of MWEs, these other properties are neither necessary nor sufficient [4].

C. *Classification of MWEs*

MWEs are broadly classified into lexicalized phrases and institutionalized phrases.

1) Lexicalized phrases: Lexicalized phrases have at least partially idiosyncratic syntax or semantics, or contain 'words' which do not occur in isolation; they can be further broken down into fixed expressions, semi-fixed expressions and syntactically-flexible expressions, in roughly decreasing order of lexical rigidity. Fixed expressions are fully lexicalized and undergo neither morphosyntactic variation nor internal modification. As such, a simple words-with-spaces representation is sufficient. If we were to adopt a compositional account of fixed expressions, we would have to introduce a lexical entry for "words" such as *hoc*, resulting in overgeneration and the idiomaticity problem [1]. Semi-fixed expressions are lexically-variable MWEs that have hard restrictions on word order and composition, but undergo some degree of lexical variation such as inflection, variation in reflexive pronouns and determiner selection [4]. They can take a range of forms including non-decomposable idioms, and certain compound nominals and proper names. Syntactically-flexible expressions exhibit a much wider range of syntactic variability. They are in the form of verb-particle constructions, decomposable idioms and light verbs.

2) Institutionalized phrases: Institutionalized phrases are syntactically and semantically compositional, but occur with markedly high frequency (in a given context). Consider for example *traffic light*, in which both *traffic* and *light* retain simplex senses and combine constructionally to produce a compositional reading [1].

III. SOME METHODS FOR MWE IDENTIFICATION

The identification methods are broadly classified into statistical methods, linguistic methods and hybrid methods.

A. *Linguistic properties of MWEs*

The following are the different association measures.

1) Pointwise Mutual Information (PMI): The PMI of a pair of outcomes *x* and *y* belonging to discrete random variables quantifies the discrepancy between the probability of their coincidence given their joint distribution versus the probability of their coincidence given only their individual distributions and assuming independence. Mathematically, PMI (x y) = $log\frac{P(xy}{P(x)p(y)}$ where, P(xy) = probability of the word x and y occurring together, P(x) = probability of x occurring in the corpus, P(y) = probability of y occurring in the corpus.

2) Log-Likelihood Ratio (LLR): The LLR is the ratio of the likelihood of the observations given the null-hypothesis to that of the alternate hypothesis. Generally, it is the ratio between the probability of observing one component of a collocation given the other is present and the probability of observing the same component of a collocation in the absence of other. Here the order of the words in the candidate collocation was irrelevant.

3) Phi-Coefficient: In statistics, the Phi coefficient is a measure of association for two binary variables. The Phi coefficient is also related to the chi-square statistic as:

$$\emptyset = \sqrt{x^2}\,/\,n$$

where *n* is the total number of observations and x^2 is the chi-square distribution [3].

The Co-occurrence Measurement, Significance Function, frequency of bigrams, t-score, Dice's coefficient are the other statistical methods.

B. *Linguistic Method*

This method is depends up on the linguistic features of the language. The set of features are length of the word, acceptable prefixes and suffixes, digit features, word and surrounding word frequency, surrounding POS tag etc [11].

C. *Hybrid Method*

In this approach, the combinations of various linguistic and statistical approaches are used.

IV. THE ARTICLES HIGHLIGHTED FOR THE IDENTIFICATION OF MWES

After an extensive review process we have selected a few papers.

Katuscak and Genci introduced several identification methods and achieved results by applying them on Slovak Language. They have applied statistical and linguistic methods. Four statistical methods were applied in the work:

frequency of bigrams, pointwise mutual information (PMI), t-score and Dice’s coefficient. At first, the frequency of bigrams was applied. They stated that this method was not designated for identification but, it has helped to eliminate numbers of candidates. The second applied method was PMI. They confirmed that of statistical methods, PMI method is the most suitable for identifying MWE in the text (85%precision). The third applied method was t-score. The prediction of this method concerning the identification of MWE was significantly lower than in PMI and Dice’s coefficient (56 % precision). As the last statistical method, Dice’s coefficient, was applied. In this method they observed almost the same prediction as in PMI (78 % precision). Regarding the weak database they are limited in applying linguistic methods for Slovak language. They stated that the usage of linguistic method is not sufficient enough to differentiate between MWE and free collocations. They concluded that set goals were accomplished and multiword expressions in Slovak language were successfully identified. The best results were obtained by the usage of statistical methods. But there are some disadvantages. They focused only on bigrams, so testing presented methods on longer n-grams are needed. The automatic identification of multiword expression is still unresolved topic therefore, the future research is required. The main contribution of this work is the list of approximately ninety thousand candidates for multiword expressions from the original list of almost one hundred million bigrams [5].

Gayen and Sarkar presents a machine learning based approach for identifying noun-noun compound MWEs from the Bengali corpus. They have used a variety of association measures, a set of WordNet-based similarity features and syntactic and linguistic clues which are combined by random forest learning algorithm for recognizing noun-noun compound MWEs. They consider the association measures namely phi, PMI, salience, log likelihood, Poisson stirling, chi, t-score, co-occurrence and significance. The F-measure value in their proposed system achieved is 86.9% [6].

Chakraborty et al. presents an approach of identifying bigram noun-noun MWEs from a medium-size Bengali corpus by clustering the semantically related nouns and incorporating a vector space model for similarity measurement. They measure the semantic similarity using cosine-similarity measurement, Euclidean distance and English WordNet. They have used the standard IR matrices like Precision, Recall and F-score for evaluating the results.They observe that English WordNet becomes a very helpful tool to identify Bengali MWEs. WordNet detects maximum MWEs correctly at the cut-off of 0.5(precision 80.90%) [7].

Chakraborty presents another paper deals with the investigation of Noun-Noun bigram collocations from the medium-size untagged Bengali corpus of the articles of Rabindranath Tagore using simple unsupervised approach with various statistical evidences to show the affinity of the constituents of each bigram candidate as a proof of the MWE and build a weighted measurement to get a distinction between MWE or non-MWE. The experimental results show that functions based on the co-occurrence distribution has given more accurate results than the frequency based measurement approaches. This paper concluded that the complete identification of MWEs in Bengali is still far apart from the present work due to the lack of lexical resources[3].

Attia et al. presents three basic approaches to identify and extract MWEs that are Crosslingual Correspondence Asymmetries, Translation-Based Approach, Corpus-Based Approach. In these approaches they use the Arabic Wikipedia (AWK), PMI and Chi-square test. They stated that MWEs encompass a set of diverse and related phenomena and also any degree of compositionality, idiosyncrasy, lexical and semantic flexibility. This complicates the task of MWE identification [8].

Kunchukuttan and Damani describe a system for extracting Hindi compound noun MWEs from a given corpus. They use various statistical co-occurrence measures to exploit the statistical idiosyncrasy of MWEs. They stated that Log-Likelihood ratio performs best among the statistical co-occurrence tests. PMI proves to be a bad measure due to the very small size corpus. The serious limitations of their approach are the use of a very small corpus and the absence of a Name-Entity recognizer [9].

Nagy T. et al. focus on the identification of two types of MWEs, namely noun compounds and light verb constructions in different domains namely, Wikipedia articles and general texts of miscellaneous topics. The applied the methods “Match” and “POS rules” and also implemented a new method “Merge” for identifying larger noun compounds. For noun compounds using POS-tagging leads to acceptable results. In the case of light verb constructions they applied different rule-based methods. Out of these, MFV (Most frequent verb) is most useful. They stated that their methods can be further improved [10].

Boukobza and Rappoport and presents a supervised learning method for identification that uses sentence surface features based on expressions’ canonical form. The base line methods Canonical Form (CF) and Distance Order (DO) and the supervised methods (using surface and syntactic features) were run on the development and training/test sets. They stated that unlike previous research, their method is not tailored to specific MWE types and they did not ignore non-expressions uses in their experiments. They concluded that the baseline accuracy, (for DO) 82.7% on the development set and 87.2% on the test set, is probably insufficient for many NLP applications [11].

Green et al. shows the effectiveness of statistical parsers

TABLE 1
COMPERATIVE LITERATURE STUDY OF 10 ARTICLES

Authors	Language	Identification method used	Corpus	Accuracy	Merits/Demerits
Katuscak and Genci	Slovak	Statistical and Linguistic methods	Approximately ninety thousand candidates for MWEs from one hundred million bigrams	85%(using PMI method)	Focused only on bigrams
Gayen and Sarkar	Bengali	Random forest learning algorithm	Online version of Bengali news paper ANANDABAZAR PATRIKA consists of 233430 tokens	F-measure value – 86.9%	Identify only noun-noun MWEs
Chakraborty et al.	Bengali	Semantic Clustering	Medium-size Bengali Corpus	80.90%	Only focused on bigram noun-noun MWEs
Chakraborty	Bengali	Simple unsupervised approach with various statistical evidences	Medium-size untagged Bengali corpus of the articles of Rabindranath Tagore	Precision-39.64% and Recall - 91.29% at top 1000 candidates	Developed a list of bigram noun-noun candidates, annotated them and ranked them. The complete identification of MWEs in Bengali is still apart from the present work
Attia et al.	Arabic	Crosslingual Correspondence Asymmetries, Translation-Based Approach, Corpus-Based Approach	Used three data resources- Wikipedia, Princeton WordNet, Arabic Gigaword Fourth Edition	Intersection between the corpus-based and the other approaches is very low	They stated that the identification of MWEs is too complex
Kunchukuttan and Damani	Hindi	Statistical co-occurrence measures	160,000 words	Recall 80% and precision 23% at rank 1000	Limitation-use of a very small corpus and the absence of a Name-Entity recognizer
Nagy T. et.al.	English	Match,POS rules and Merge	List constructed from Wikipedia articles and general texts of miscellaneous topics	Achieved better results than the original ones	They stated that their methods can be further improved
Boukobza and Rappoport	English	Sentence surface features based on expressions' canonical form	Collins COBUILD Advanced Learner's Dictionary	Baseline accuracy 82.7%o the development set and 87.2 on the test set	Not tailored to specific MWE types
Green et al.	French	Tree substitution grammers	French Treebank version from June 2010	Better baseline for parsing raw French text	Stated that MWEs are hard to identify
Nongmeikapam and Bandyopadhyay	Manipuri	CRF based Genetic Algorithm in Feature Selection	45,000 tokens	Recall 64.08% Precision 86.84%, F-measure - 73.74%	Burden of manual feature selection is reduced

for MWE identification. Specifically Tree Substitution Grammars (TSG) can achieve the best results over the surface statistics method. The choose French which has pervasive MWEs for their experiments. The experimental results provide a better baseline for parsing raw French text. They stated that a dilemma thus exists:

MWE knowledge is useful, but MWEs are hard to identify [12].

Nongmeikapam and Bandyopadhyay adopted an integrated model, which can perform CRF based MWE identification but changes are made with the feature list and feature selection. The feature selection is applied with the concept of Genetic algorithm. This model has come up with the successful implementation of GA in feature selection of CRF for the first time in Manipuri language. Using GA for feature selection they are able to find the optimal features to run the CRF. They have tried with fifty generations in feature selection along with three fold cross validation as fitness function. This model demonstrated the Recall of 64.8%, precision of 86.84% and F-measure of 73.4%, showing an improvement over the CRF based Manipuri MWE identification without GA application [13].

V. COMPARATIVE STUDY

We compared ten articles with language, identification method used, corpus, accuracy and highlighted the merits/demerits of their method. Table 1 shows the result of the comparative literature study. Most of the work is focussed only on bigrams. From these results we concluded that identification of MWEs is too complex.

VI. CONCLUSIONS AND FUTURE WORK

In this paper we focused on the classifications and features of MWEs and the different identification techniques. We discussed various articles of different authors in this field and made a comparative study of their contributions. But the identification of MWEs is still a key challenge in NLP especially in Indian languages due to its lack of capitalization information, confusion between named entities and normal words, lack available corpus etc. However, for our future study we are trying to develop a new identification method.

REFERENCES

[1] Ivan A.Sag, Timothy Baldwin, Francis Bond, Ann Copestake and Dan Flickinger " Multiword Expressions: A Pain in the Neck for NLP", *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, Springer-Verlag London, UK, February 17 - 23 2002, pp. 1-15.

[2] Aline Villavicencio, Francis Bond, Anna Korhonen, Diana McCarthy, "Introduction to the Special Issue on Multiword Expressions: having a crack at a hard nut", *Computer Speech & Language*, October 2005, 19(4):365-377.

[3] Chakraborty, T., "Towards Identification of Nominal Multiword Expressions in Bengali Language", *Open Access LibraryJournal*,2014,**1**:e582. http://dx.doi.org/10.4236/oalib.1100582.

[4] Nitin Indurkhya (Editor), Fred J. Damerau, "Handbook of Natural Language Processing", Second Edition,*Chapman & Hall/CRC Machine Learning & Pattern Recognition*, pp.267-281.

[5] Matej Katuscak, Jan Genci, "Identification of Multiword Expressions for Slovak Language", *Proceedings of the Faculty of Electrical Engineering and Informatics of the Technical University of Kosice*, 2015.

[6] Vivekananda Gayen, Kamal Sarkar, "A Machine Learning Approach for the Identification of Bengali Noun-Noun Compound Multiword Expressions", *Proceedings of ICON-2013: 10th International Conference on Natural Language Processing*, http: //ltrc.iiit.ac.in/proceedings /ICON-2013.

[7] Tanmoy Chakraborty, Dipankar Das ,Sivaji Bandyopadhyay, "Semantic Clustering: an Attempt to Identify Multiword Expressions in Bengali", *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE 2011)*, Portland, Oregon, USA, 23 June 2011, pp. 8–13.

[8] Mohammed Attia, Antonio Toral, Lamia Tounsi, Pavel Pecina, Josef van Genabith, "Automatic Extraction of Arabic Multiword Expressions", *Proceedings of the Multiword Expressions: From Theory to Applications (MWE 2010)*, Beijing, August 2010, pp 19–27.

[9] Anoop Kunchukuttan, Om P. Damani, " A System for Compound Noun Multiword Expression Extraction for Hindi", *Proceedings of ICON-2008: 6th International Conference on Natural Language Processing* ,Macmillan Publishers, India.

[10] Istvan Nagy T, Veronika Vincze, Gabor Berend, "Domain-dependent identification of multiword expressions", *proceedings of the Conference: Recent Advances in Natural Language Processing, RANLP 2011*, Hissar, Bulgaria, 12-14 September, 2011.

[11] Ram Boukobza, Ari Rappoport, "Multi-Word Expression Identification Using Sentence Surface Features", *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Singapore, 6-7 August 2009, pp. 468–477.

[12] Spence Green, Marie-Catherine de Marneffe, John Bauer, Christoper D. Manning, "Multiword expression identification with tree substitution grammars: A

parsing TOUR DE FORCE with French", *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011*, , John McIntyre Conference Centre, Edinburgh, UK, 27-31 July 2011, pp.725-735.

[13] Kishorjit Nongmeikapam, Sivaji Bandyopadhyay, "Genetic Algorithm (GA) in Feature Selection for CRF Based Manipuri Multiword Expression (MWE) Identification", *International Journal of Computer Science & Information Technology (IJCSIT) Vol 3, No 5*, Oct 2011, pp 53-66.