IDENTIFICATION AND EXTRACTION OF MULTI WORD EXPRESSION FROM INDIAN LANGUAGES: REVIEW

Kapil Dev Goyal *(Assistant Professor)*
SBAS Khalsa College, Sandaur (Sangrur).
Email: kapildevgoyal@gmail.com

ABSTRACT

Multiword Expressions (MWEs) are used frequently in natural languages, but understanding the diversity in MWEs is one of the open problems in the area of Natural Language Processing. In the context of Indian languages, MWEs play an important role. 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 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 languages, Indian Languages, language processing.

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].

Recently, various approaches have been proposed for the identification and extraction of MWEs. The quality of such approaches depends on the use of algorithms and also on the quality of resources used. Various standard MWEs datasets1 are available for languages like English, French, German, Portuguese, etc. and can be used for evaluation of MWE approaches. But for Indian languages, no such standard datasets are available publicly. Our goal is to create MWEs annotation in Indian languages (Hindi and Marathi) and make it available publicly. We have explored two types of MWEs that are compound nouns (CNs) and light verb constructions (LVCs). Since, CNs and LVCs are used very frequently in the text data in comparison to other MWEs; we have considered only these MWEs in this paper. The created resource can be useful for various natural language processing applications like information extraction, word sense disambiguation, machine translation, etc.

LITERATURE REVIEW

Multiword Expressions is a dialectal expression conveys a different meaning, that what is evident from its words. In this paper, the author has presented the technique for searching and translating English idioms into Hindi in the translation process. The rule based and statistical machine translation approaches for identification of idioms have been proposed by the author [1].

Multiword expressions are words that exhibit characteristics of a single syntactic word. In this paper, the author analyzes the challenges provided by MWE‟s in the sentences. Some collocations are used together even though they are perfectly compositional and there exist alternatives for the constituent words. This suggests that the usage of that collocation have been frozen [2].

In this paper, the author has presented the methodology to extract MWE‟s. Multiword expressions had considerably attracted researchers. However, identifying the multiword expressions properly had proven to be „A pain in the neck‟ for Natural Language Processing, due to lack of competent resources such as manually annotated corpora in languages. To analysis MWE‟s in English-Hindi Language, three corpus are used in the study. First is of agricultural domain, second is of Bharat Dharshan-Hindi Sahityik Patrika and third is of general domain [3].

Multiword detection is very difficult task in Natural Language Processing. Manual encoding of linguistic information is being challenged by automated corpus based learning methodologies for NLP. Corpus based approaches have been successful in many several areas of the Natural Language Processing. In multiword detection individual terms are analyze in the form of syntax and semantic [4].

The identification of types of the multiword expressions requires different solutions, which might be domain-related differences in the frequency and typology. The author has defined the methods for identifying noun compounds and light verb constructions can be adapted to different domains and different types of texts. The results indicate that with little effort, existing solutions for detecting multiword expressions can be successfully applied to other domains [5].

This paper presents systematic and methodology for designing the English to Khmer machine translation using Moses. Moses is an implementation of the statistical approach to machine translation. This is most used approach in the field at the moment and is employed by the online translation systems like Google and Microsoft. The author implements on very few parallel corpus [6].

In this paper, Bengali to Assamese Statistical Machine Translation Model has been created by using the Moses. Parallel corpus of 17,100 sentences in Bengali and Assamese had been built. The focus of this research, was to investigate the effectiveness of a phrase based statistical Bengali Assamese translation using the Moses. Machine translation is considered as one of the difficult task [7].

Hindi belongs to Indo-Aryan languages and Dogri also belongs to the same subgroup of the Indo-European family. For the development of Machine Translation system from Hindi to Dogri Language, there is a need to find the closeness between the languages. It is found that both the languages are closely related to each other. Dogri is written using Devanagari script and has thirty eight segmental and five supra segmental phonemes [8].

Classification of Multiword Expression

MWEs can be split-up into lexicalized phrases which have at least in part idiosyncratic syntax or pragmatics, and institutionalized phrases which are syntactically and semantically compositional. Lexicalized phrases can be further sub classified into fixed expressions, semi-fixed expressions and syntactically flexible expressions.

Fixed expressions: Fixed expressions are fully lexicalized and can neither be vitiated morph syntactically nor modificated internally. Examples for fixed expressions are: in short, by and large, and every which way. They are fixed, as you cannot say in shorter or in very short.

Semi-fixed expressions: In semi-fixed expressions word order and composition are strictly invariable, while inflection, variation in reflexive form and determiner selection is possible. In

non-decomposable idioms (i.e., idioms in which the meaning cannot be assigned to the parts of the MWE) such as kick the bucket the verb can be inflected according to a particular context: He kicks the bucket. On the other hand non-decomposable idioms do not undergo syntactic variability. For example, a passive sentence as the bucket was kicked is not possible (or at least it does not have the same meaning).

Syntactically-flexible expressions: Syntactically flexible expressions have a wider range of syntactic variability than semi-fixed expressions. They occur in the form of decomposable idioms, verb-particle constructions and light verbs. Decomposable idioms are likely to be syntactically flexible to some degree. Examples are let the cat out of the bag and sweep under the rug. Yet, it is hard to predict which kind of syntactic variation a given idiom can undergo.

Verb-particle constructions, such as write up and look up are made up of a verb and one or more particles. Either they are semantically idiosyncratic as brush up on or compositional as break up in the meteorite broke up in the earth's atmosphere. In some transitive verb-particle constructions as call up, an NP argument can occur either between or following the verb and particle(s): call Kim up or call up Kim, respectively. In addition adverbs can often be inserted between the verb and particle as in fight bravely on. For light verb constructions, as make a mistake, give a demo it is difficult to predict which light verb combines with a given noun. Though they are highly idiosyncratic they have to be distinguished from idioms: "The noun is used in a normal sense, and the verb meaning appears to be bleached, rather than idiomatic."

Institutionalized phrases: Institutionalized phrases are conventionalized phrases, such as salt and pepper, traffic light and to kindle excitement. They are semantically and syntactically compositional, but statistically idiosyncratic. Regarding the phrase traffic light, traffic and light both retain simplex senses but produce a compositional reading by combining constructionally.

System Architecture

The multiword expression need to design as a set of Python scripts that handle intermediary XML (eXtended Markup Language) representing the corpus, the list of MWE patterns, the list of

MWE candidates and the reference dictionary. Each script performs a specific task in the pipeline of MWE extraction, from the raw corpus to the filtered list of MWE candidates including their automatic evaluation if a reference gold standard is given. Fig. 1 summarizes the architecture of multiword expression preprocessing using external tools should include (a) consistent tokenization, (b) lemmatization and (c) part- of speech tagging. Steps (b) and (c) are optional, but lemma and POS (Part-of-Speech) [3] information can be crucial for determining the quality of the extracted MWEs.

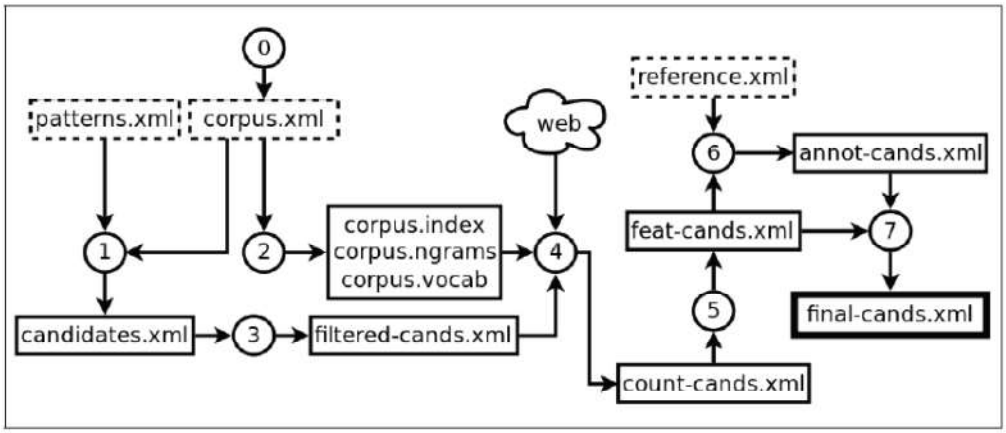

Fig. 1 System architecture of MWE identification (0) Corpus Preprocessing , (1) Extraction of the candidates that match the patterns, (2) using suffix arrays, index the corpora (3) filter the candidates list, (4) count n-grams and words in the corpora, (5) calculate AMs and descriptive features, (6) automatically annotate (part of) the candidates and (7) train/apply a machine learning model. Inputs are boxed with dashed lines, output with a thick line. [17]

Algorithm Required

For each candidate in multiword expression, a set of features is generated in order to allow the application of machine learning models. Two kinds of features are included in the multiword expression identification package: statistical Association Measures (AMs) and descriptive

features. The former measure the degree of independence between the number of occurrences of the MWT candidate and the number of occurrences of the individual words that compose it.

$$mle = \frac{c(w_1.....w_n)}{N}$$

$$dice = \frac{n \times c(w_1.....w_n)}{\sum_{i=1}^{n} c(w_i)}$$

$$pmi = log_2 \frac{c(w_1.....w_n)}{E(w_1.....w_n)}$$

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

AMs are calculated as follows:

A corpus containing N word tokens is indexed using a suffix array, a memory-efficient data structure that allows for n-grams of arbitrary size to be searched efficiently in very large corpora.

- For each candidate sequence of n contiguous words w_1 through w_n, multiword expression identification gets the individual word counts $c(w_1)……c(w_n)$ and the overall n-gram count $c(w_1…....w_n)$ from the index.
- We calculate the expected n-gram count E if words co-occurred by chance, i.e., if we suppose that word occurrences are independent events, an n-gram would occur $E(w_1…….w_n) \approx c(w_1)$ ….$c(w_n)$/Nn-1 times.
- That information is used to calculate four statistical AMs for each MWE candidate in each corpus, namely: mle stands for maximum likelihood estimator.

We are able to calculate these measures for arbitrary-size n-grams because none of them uses contingency tables. Since candidate extraction and counting are two separate steps, an arbitrary number of corpus frequencies can be calculated.

MWE Annotation Guidelines

In this section, we describe guidelines given to human annotators to annotate MWEs from the possible candidates. Annotators has been told to check whether the candidate (word-pair) satisfy the following criteria of MWEs formation.

• **Reduplication:** Here, a root or stem of a word, or part of it is repeated. Reduplication can further be subdivided into:

– **Onomatopoeic Expression:** In this case, the constituent words imitate a sound or a sound of an action. Generally, in this case, the words are repeated twice with the same 'matra'. E.g. ਟਿਕ-ਟਿਕ (tick tick, the ticking sound of watch's needle).

– **Non-Onomatopoeic Expression:** Here, the constituent words have meaning but they are repeated to convey a particular meaning. E.g. ਚਲਦੇ ਚਲਦੇ (chalde chalde, while walking).

– **Partial Reduplication:** In this case, one of the constituent word is meaningful while the other is constructed by partially repeating the first word. E.g: ਪਾਣੀ ਵਾਣੀ (paani vaani, water).

– **Semantic Reduplication:** Here, the constituent words have some semantic relationship among them. E.g. ਧਨ ਦੋਲਤ (dhana daulat, Wealth) (Synonymy), ਦਿਨ ਰਾਤ (din raat always) (Antonymy).

• **Fixed Expression:** Fixed Expressions are immutable expressions, which do not undergo any transformation or morphological inflections or possibility of insertion between two words. E.g. ਥੋੜੇ ਤੋਂ ਥੋੜੇ (thode ton thode, atleast), ਜਿਆਦਾ ਤੋਂ ਜਿਆਦਾ (jiyada ton jiyada, maximal).

• **Semi-fixed Expression:** Semi-fixed expressions obey constraints on word order and composition. They might show some degree of lexical variation. E.g. ਕਾਰ ਪਾਰਕ (car park, It can be car park(s)).

• **Non-Compositional:** The meaning of a complete multiword expression cannot be completely determined from the meaning of its constituent words. E.g. ਅਕਸ਼ਯਾ ਤ੍ਰਿਤੀਆ (akshaya Tritiyaa, a festival in India)

• **Decomposable Idioms:** Decomposable idioms are syntactically flexible and behave like semantically linked parts. But it is difficult to predict exactly what type of syntactic expression they are. E.g. ਆਟੇ ਦਾਲ ਦਾ ਭਾਅ ਮਾਲੂਮ ਹੋਣਾ (aate daal da bhah maalam honaa, to creat a knowledge). Here in this example, we can replace the phrase ਆਟੇ ਦਾਲ ਦਾ ਭਾਅ ਮਾਲੂਮ ਹੋਣਾ to ਆਟੇ ਦਾਲ ਦੀ ਕੀਮਤ ਮਾਲੂਮ ਹੋਣਾ.

• **Non-Decomposable Idioms:** Non-Decomposable idioms are those idioms, which do not undergo any syntactic variations but might allow some minor lexical modification. E.g. ਨੋ ਦੋ ਗਿਆਰਾਂ ਹੋਣਾ (Nau do gyaraaha honaa, to run off).

• **Name Entity Recognition (NER):** Named entities are phrases that contain the names of persons, organizations, locations, times, and quantities. NERs are syntactically highly idiosyncratic. These entities are formed based on generally a place or a person. E.g. ਭਾਰਤੀਆ ਪਰੋਦਯੋਗੀਕੀ ਸੰਸਥਾਨ (Bhartiya Prodyogiki Sansthan, Indian Institute of Technology) (Organization), ਸਚਿਨ ਤੇਂਦੁਲਕਰ (Sachin Tendulkar, Sachin Tendulkar) (Proper noun), ਤਾਜ ਮਹਲ (Taj Mahal) (Location), etc.

• **Collocations:** A collocation is an expression consisting of two or more words that correspond to some conventional way of saying things. E.g. ਕੜਕ ਚਾਹ (kadak chah, strong tea), ਪੋਸਟ ਆਫ਼ਿਸ (Post office, post office), etc.

• **Foreign Words:** A set of words borrowed from another languages are called as Foreign words. They can be treated as valid MWEs in the context of Indian languages. E.g. ਰੇਲਵੇ ਸਟੇਸ਼ਨ (Railway station, Railway Station), ਪੋਸਟ ਆਫਿਸ (Post office, post office), etc.

Conclusions

The multiword expression identification can be used not only to speed up the work of lexicographers and terminographers in the creation of terminological resources for new domains and languages, but also to contribute to the porting of NLP systems such as Machine Translation and Information Extraction across domains. This methodology employed in the multiword expression identification is not based on symbolic knowledge or dictionaries, and the techniques implemented in it are language independent. Therefore, it can straightforwardly be applied to any language and domain for which a corpus is available, with the execution of simple corpus preprocessing steps and the definition and tuning of POS patterns, for improved performance. We expect, to improve that integrate a higher number of features about the MWE candidates into the classifiers, in order to provide more accurate results. Among possible improvements are new descriptive features, contingency-table association measures and information coming from peripheral sources such as parallel corpora (word alignments) and general-purpose or domain-specific dictionaries. Moreover, we would like to provide better integration between the candidate extraction step and the classifier construction step. We would like to investigate the use of a plethora of preprocessing alternatives such as language independent (de)capitalization and tokenization tools with customizable parameters and incorporate those to the MWE identification, along with the integration with a number of external language-dependent tools like lemmatisers and POS taggers (e.g., for English or any other Indian languages).

REFERENCES

[1] Monika Gaule and Dr. Gurpreet Singh Josan, (2012) “Machine Translation of Idioms from English to Hindi”, International Journal of Computational Engineering Research, Vol. 2 Issue 6,

[2] Anoop Kunchukuttan, Munish Minia and Pushpak Bhattacharyya, (2010) "Multiword Expressions in the CLIA Project".

[3] Vivek Dubey, Pankaj Raghuwanshi, Sapna Vyas, (2015) "Impact of Multiword Expression in English-Hindi Language", International Journal of Emerging Trends & Technology in Computer Science, Volume 4, Issue 3, ISSN 2278- 6856,

[4] Md J. Abedin, B. S. Purkayastha, (2013) "Detection Of Multiword From A Wordnet Is Complex", International Journal of Research in Engineering and Technology, Volume: 02 ,Special Issue: 02, ISSN: 2319-1163 | ISSN: 2321- 7308.

[5] Istvan Nagy T., Vernoika Vincze and Gabor Berend, (2014) "Domain-Dependent Identification of Multiword Expressions".

[6] Suraiya Jabin, Suos Samak and Kim Sokphyrum, (2013) "How to Translate from English to Khmer using Moses", International Journal of Engineering Inventions, e-ISSN: 2278-7461 |pISSN: 2319-6491, Volume 3, Issue 2 pp: 71-81.

[7] Nayan Jyoti Kalita, "Baharul Islam, (2012) Bengali to Assamese Statistical Machine Translation using Moses (Corpus Based)".

[8] Preeti Dubey, Shashi Pathania and Devanand, (2015) "Comparative Study of Hindi and Dogri Languages with Regard to Machine Translation".

[9] Lahari Poddar and Puhshpak Bhattacharyya, (2014) "Multilingual Multiword Expressions".

[10] Rakesh Chandra Balanbantaray, Deepak Sahoo, (2014) "Odia Transliteration engine using Moses".

[11] Jeremy D. Brightbill and Scott D. Turner, (2015) "A Sociolinguistic Survey of the Dogri Language", Jammu and Kashmir.

[12] Murali Nandi and Ramasree R.J., (2013) "Rule-based Extraction of Multi-Word Expressions for Elementary Sanskrit Texts", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 11, ISSN: 2277 128X.

[13] Vishal Goyal (2012) "Development Of A Hindi To Punjabi Machine Translation System".

[14] Yulia Tsvetkov and Shuly Winter, (2010) "Extraction of Multi-word Expressions from small parallel corpora".

[15] Pallavi and Dr. Anitha S Pillai, (2013) "Named Entity Recognition For Indian Languages: A Survey", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 11, ISSN 2277 128x.

[16] Shubhnandan S Jamwal and Sunil Dutt, (2015) "Tuning of Moses Decoder for Dogri SMT", International Journal of Computer Science & Communication, Vol- 6 • Issue-1 pp.145-147.

[17] Ramisch, C., Villavicencio, A., & Boitet, C. (2010, May). Mwetoolkit: a framework for multiword expression identification. In *LREC* (Vol. 10, pp. 662-669).