ISSN (online): 2250-141X
Vol. 5 Issue 3, October 2015

AUTOMATICALLY MINING OF MULTIWORDS IN PARALLEL ENGLISH HINDI SENTENCES

Vivek Dubey[1],Pankaj Raghuwanshi[2],Sapna Vyas[3]
Alpine Institute of Technology, Dewas Road, Ujjain (MP)
vivekdubey22@gmail.com

Abstract:- *Now a day, getting touch with friends and relatives, people are usually online. Many are using social site through facebook, whatsapp, gtalk, etc. However, they are need of assistance to get proper words and sentences. Many times, they are also using online translator to get correct and quick translation of English sentence into Hindi sentence and vice-versa. For simple sentences, online translators are perfect as they are translated word-to-word translation. But when two-words verbs like draw_back – पीछे_हटना /मुकर_जाना, has_read – पढ़_लिया are occurred in sentence, online translators are helpless. In this paper, a simple method for identifying Multiwords Verbal Chunk of all kinds by means of python is presented for an English-Hindi parallel corpus and said system yields mining English-Hindi MWVC with an average precision is 90%-83% and a average recall is 93%-98%. The English-Hindi MVWC dictionary will be improved Natural Language Processing like Parts of Speech Tagging, Information Retrieval, Summarization, Word Alignment, Machine Translation etc.*

Keywords: *Multiwords, Verbal Chunk, Parts of Speech, Word Alignment, Python.*

I. Introduction

In Natural Language Processing (NLP), one of the most challenging jobs is the proper treatment of multiword chunk (MWC). They are lexical items [1] that are composed of a word i.e. boy, dog, go, etc, a part of word i.e. *non*sense, top*most* or a group of words i.e. ask for, smart card, again and again, all of a sudden. Ambiguities [2,3,11] in NLP are many times mainly due to not catching multiword chunk in a sentence during analysis i.e. parsing and during generation. For example, in the English sentence: *The policemen made after the thief very fast* and in its Hindi translation: *पुलिसकर्मियों ने बहुत तेजी से चोर के बाद किया*, the multiword verbal chunk *made_after* is not meant as *के_बाद_किया* but it is as *के_पीछे_दौड़े*. In sentences, multiword may be formed in subject, object and verb. The identification of Multiword Verbal Chunk (MWVC) is the initial task in mapping parallel English-Hindi sentences for extracting words and multiwords. It is observed as simple problem but practically it is complex task. Hindi verbal multiword chunk has been identified by light verb construction. This construction [4] is also called Complex Predicate (CPs) where part of speech (POS) likes a noun, a verb, an adjective are followed by a light verb, for example HinMWVC: *परेशान_करना* – EngV: *bother*. Language industry is the sector of activity dedicated to facilitating multilingual communication, both oral and written. These industries are growing exponentially. It also requires parallel English-Hindi multiwords to trained many application of NLP like Part of Speech Tagging, Information Retrieval, Summarization, Word Alignment, Machine Translation etc. Manually, identification and mapping of MWVC are time consuming, tedious, expensive, and error-prone and it also requires intelligence and knowledgeable person. Proper automated processing system will be impact and reduced manual processing costs, while also improved processing speed and accuracy.

ISSN (online): 2250-141X
Vol. 5 Issue 3, October 2015

The formation of the paper is as follows. Section-1 describes related work of MWVC in parallel English-Hindi corpus, section-2 discusses the analysis of parallel MWVC, section-3 explains the automatic identification and extraction system and section-4 briefs the experiments and results.

II. Related Work

Bannard identified verb and noun construction in English on the basic of syntactic fixedness [5]. He examined whether the noun could had a determiner or not, whether the noun could be modified and whether the construction could had a passive form, which features are exploited in the identification of the construction. Gurrutxaga and Alegria extracted idioms and light verb constructions from Basque text by employing statistical methods [6]. Since Basque is a free word-order language, they hypothesized that a wider window would yield more significant co-occurrence statistics; however, their initial experiments did not confirm this.

Tu and Roth classified verb+noun object pairs as being light verb construction or not [7]. They operated with both contextual and statistical features and conclude that on ambiguous examples, local contextual features perform better. Vincze et al. exploited shallow morphological features in identifying English light verb constructions [8] and domain specificity of the problem was emphasized in [9].

Rasooli proposed a bootstrapping approach for identification of compound verb and light verb construction [10]. Their consist corpus considered for MWE was annotated with POS tags and some morpho-syntactic features. Parallel corpus is highly importance in the automatic identification of multiword chunk. It is usually one-to-many correspondences that are exploited when designing methods is for detecting multiword expressions. On the other hand, aligned parallel corpus can also enhance the identification of multiword expressions in different language. Caseli et al. (2010) developed an alignment-based method for extracting multiword expressions from parallel corpora. The first step was to align the corpus on the sentence level, which was followed by POS-tagging. After this, sentence alignment units were word aligned. Candidates for multiword expression were produced by the word aligner and the POS-tagger as well, then they were filtered according to some empirically defined pattern or frequency data.

ZarrieB and Kuhn argued [12] that multiword expression could be reliably detected to parallel corpora by using dependency-parsed, word- aligned sentences. For one-to-many translation pairs, they applied a generate-and-filter strategy. First, aligned syntactic configurations were generated, which were then filtered and post-edited.

Sinha detected Hindi complex predicates [13] (i.e. a combination of light verb and a noun ,a verb, or an adjective) in a Hindi-English parallel corpus by identifying a mismatch of the Hindi light verb meaning in the aligned English sentence. Although the method required the generation of all possible light verbs, it seemed to be applicable to languages of the indo Aryan family.

Many-to-one correspondence was also exploited in Attia et al. when identifying Arabic multiword expressions relying asymmetries between entry titles of Wikipedia [14]. After packet has been reached to the destination, destination will wait for time δt and collects all the packets. Tsvetkov and Wintner identified Hebrew multiword expressions by searching for misalignments in an English-Hebrew parallel corpus [15]. MWE candidates were then ranked and filtered based on monolingual frequency data.

III. Analysis of Multiword Verbal Chunk

Light verb constructions may occur in various forms due to their syntactic flexibility. Besides, the prototypical *noun+verb*

ISSN (online): 2250-141X
Vol. 5 Issue 3, October 2015

Combination in Hindi and the *verb+noun* combination in English, light verb constructions may be declared in different syntactic structure, that is, PARTICIPLES (e.g. *give up*) and may be also undergo nominalization, yielding a NOMINAL COMPOUND (e.g. *service provider*). Some common verbal components are use in Hindi and English language likes give/देना, go/जाना, take/लेना, be/होना, do/करना, keep/रखना, make/बनाना etc. Using main verb in English-Hindi, mapping of MWVCs are usually:- 2:1, 1:2, 2:2, 3:2, 2:3, 3:3, 4:3 in ratio of English-Hindi words as described in table-1.

Table-1

Words Ratio	E-H MWVC Mapped Words
2:1	shall_go – जाऊंगा
1:2	go - जाता_हूँ ; went - जाता_था
2:2	had_gone - गया_था
3:2	shall_be_going - जाता_रहूँगा
2:3	am_going - जा_रहा_हूँ; have_ gone - जा_चुका_हूँ
3:3	have_been_going - जाता_रहा_हूँ
4:3	shall_have_been_going - जाता_रहा_होऊंगा

In Hindi sentence, some constructions of main verb are also followed by a noun, a verb, an adjective, an adverb as explained in table-2.

Table-2

English Main Verb	Hindi Main Verb	Example
Harassed	Noun+verb	परेशान_किया
Blamed	Adjective+verb	दोषी_ठहराया
Has read	Verb+ verb	पढ़_लिया
completely ignore	Adverb+verb	पूरी_तरह_से_उपेक्षा

English sentence are followed grammar rules as SUBJECT + VERB+ OBJECT (SVO) and Hindi sentence as SUBJECT + OBJECT+ VERB (SOV). For research work, initially, English-Hindi parallel sentence are identified from [17] and English and Hindi sentences are saved in a separate file for preprocessing. Then all sentences are cleaned from unwanted characters like space, unrecognized characters and segmented properly. Then also, in sentences, short words are replaced with proper multiword, like I'm - I am, I've – I have. Later, Parts-of-Speech (POS) of English sentences and Hindi sentences have been identified and English-Hindi Sentences are tagged from [18,19] and tagged English-Hindi sentences are saved in a Tagged file separately. The complete process is described in figure-1.

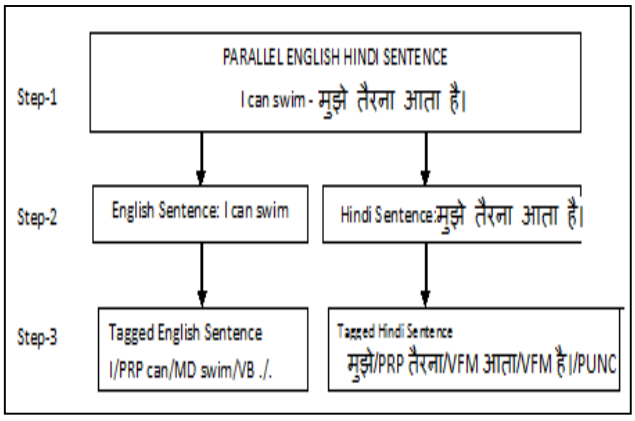

Figure-1

After getting tagged sentence English and Hindi sentence, English and Hindi sentence are read from file till end of file (EOF) and Multiword Verbal Chunk (VMWC) are found using Rule Based methodology in both English and Hindi Tagged Sentence and Lastly Eng-MWVC and Hin-MWVC are saved in file separately. The program is written in Python Language. The complete process of identification and extraction of MWVC in English Hindi sentence are briefed in flow chart as figure-2 and program coding for Identifying and Extracting English

IV. MWVC Identification System

ISSN (online): 2250-141X
Vol. 5 Issue 3, October 2015

MWVCs and Hindi MWVCs are shown in figure-3 and figure-4 respectively.

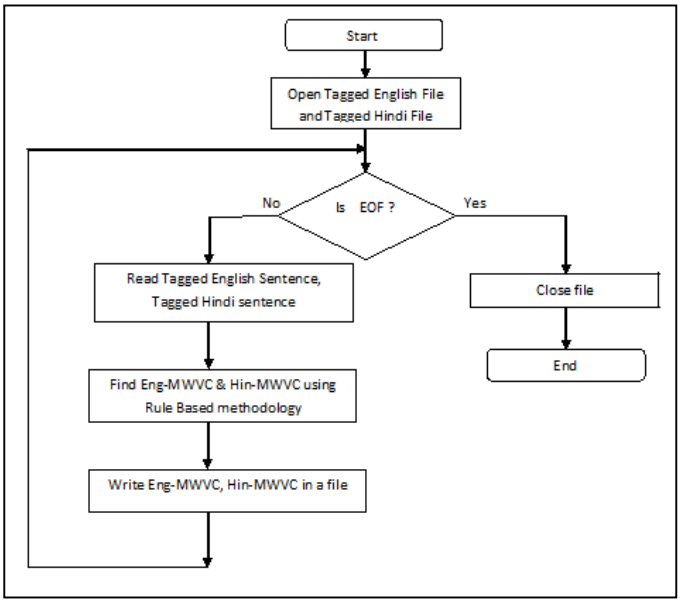

Figure-2

```
import nltk
import re
import nltk.data
g1=open("TagME05.txt")
g2=open("VMWME15.txt", mode='w')
rf=g1.read()
SubStr=''
for token in rf:
    if token == '\n':
        SubStr = SubStr + token
        #print(SubStr, end='')
        tk = nltk.word_tokenize(SubStr)
        tagtk = [nltk.tag.str2tuple(q) for q in SubStr.split()]
        Ev1=[w for w,pos in tagtk if(pos=="VBZ" or pos=="VB" or
        pos=="VBN" or pos=="VBP" or pos=="MD" or pos=="VBD"
        or pos=="VBG" or pos=="VNN" or pos=="VJJ" or
        pos=="RP")]
        print(Ev1,end='\n',file=g2)
        SubStr=''
        token=''
    SubStr = SubStr + token
g1.close()
g2.close()
```

Figure-3

```
import nltk
import re
from nltk import word_tokenize
import codecs
hg1=codecs.open("tagMh05.txt","r","utf-8")
hg2=codecs.open("VMWMH15.txt", "w","utf-8")
hrf=hg1.read()
SubStr=''
for token in hrf:
    if token == '\n':
        SubStr = SubStr + token
        tk = nltk.word_tokenize(SubStr)
        tagtk = [nltk.tag.str2tuple(q) for q in SubStr.split()]
        Hv1=[w for w,pos in tagtk if
            (pos=="VFM" or pos=="VAUX" or pos=="VJJ" or pos=="VRB"
            or pos=="VNN" or pos=="NVB" or pos=="RP" or
            pos=="NO")]
        print(Hv1,end='\r\n',file=hg2)
        SubStr=''
        token=''
    SubStr = SubStr + token
hg2.close()
hg1.close()
```

Figure-4

V. Experiment and Results

The MWVC mining methodology outlined has been implemented and tested over English-Hindi parallel Sets. A summary of the results obtained are given in table-3. As can be seen from table-3, the precision obtained is in English as 82% - 100% and in Hindi as 77% - 86% and the recall is in English and in Hindi between as 94% - 100%. The F-measure of English is 87% to 100% and Hindi is 85% to 92%. Without much of linguistic or statistical approach, it is an amazing and to some extent unforeseen result.

ISSN (online): 2250-141X
Vol. 5 Issue 3, October 2015

Table-3

	Set-01		Set-02		Set-03		Set-04		Set-05	
	English	Hindi	English	Hindi	English	Hindi	English	Hindi	English	Hindi
No of Sentences	32	32	26	26	35	35	36	36	36	36
Total No of MWVC (N)	39	58	34	52	48	78	50	67	49	76
Correctly identified MWVC (TP)	30	27	25	23	33	30	31	31	35	32
Parallel E-H MWVC	24		21		28		29		31	
Incorrectly identified MWVC (FP)	02	08	01	04	04	05	05	05	04	05
Unidentified MWVC(FN)	02	01	01	01	02	00	05	01	02	00
Accuracy %	76%	46%	73%	44%	68%	38%	62%	46%	71%	42%
Precision % (TP/ (TP+FP))	93%	77%	96%	85%	89%	85%	86%	86%	89%	86%
Recall % (TP / (TP+FN))	93%	96%	96%	95%	94%	100%	86%	96%	94%	100%
F-measure % (2PR / (P+R))	93%	85%	96%	89%	91%	91%	86%	90%	91%	92%

In fact due to hard in English-Hindi tagging, the result are lacking. For example, the one online tagger is tagged English sentence of *I am bored* as *I/PRP am/VBP bored/VBN* and other is tagged as *I/PRP am/VBP bored/JJ*. Similarly in another example, the one online tagger is tagged English sentence of *Have fun* as *Have/NNP fun/VBP* and other is tagged as *Have/VBP fun/NN*. Likewise in the online tagger is tagged Hindi sentence of *पंछी गाते हैं* as *पंछी/PREP गाते/PREP हैं/VFM*. However, after preprocessing, the results will be quite improved.

VI. Conclusion

In this paper, an in-depth case study on Verbal Chunk recognition as multiword and extraction is proposed with rule-based methods, in which a supervised learning system using tagging is built. Accuracy is found in English – 61% and in Hindi - 43%. It is also observed that still quality tagger for both English and Hindi have to design and develop to improve made system.

VII. Acknowledgement

Authors are grateful to Prof. Vineet Chataniya, IIIT Hyderabad and Prof. (Dr.) Amba Kulkarni, University of Hyderabad, for their valuable suggestions and Madhya Pradesh Council of Science and Technology Bhopal for the sanction of research project no. A/RD/RP-2/2014-15/234.

VIII. References

I. https://en.wikipedia.org/wiki/Lexical_item.

II. Anjali M. K. and Babu Anto P., (2014), International Journal of Innovative Research in Computer and Communication Engineering, ISSN(Online): 2320-9801, Vol.2, Special Issue 5, pp 392-394.

III. Green, Spence, Marie-Catherine de Marneffe, John Bauer, and Christopher D. Manning. Multiword Expression Identification with Tree Substitution Grammars: A Parsing tour de force with French, in proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 725–735, Edinburgh, Scotland, UK., July 2011, Association for Computational Linguistics.

IV. R. Mahesh K. Sinha, 2009, Mining Complex Predicates In Hindi Using A Parallel Hindi-English Corpus, in proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications, pages 40–46, Singapore,August, Association for Computational Linguistics.

V. Colin Bannard, 2007, A measure of syntactic flexibility for automatically identifying multiword expressions in corpora, in proceedings of the Workshop on a Broader Perspective on Multiword Expressions, MWE '07, pages 1–8, Morristown, NJ, USA, ACL.

ISSN (online): 2250-141X
Vol. 5 Issue 3, October 2015

a. 23–30, Singapore, August. Association for Computational Linguistics.

VI. A. Gurrutxaga and I. Alegria, 2011, Automatic extraction of NV expressions in Basque: basic issues on cooccurrence techniques, ACL HLT 2011, page 2.

VII. Y. Tu and D. Roth, Learning English Light Verb Constructions: Contextual or Statistical , ACL-HLT workshop: Multiword Expressions: from Parsing and Generation to the Real World (MWE 2011), Portland, Oregon, 2011.

VIII. Veronica Vincze, Istvan Nagy, and G ´ abor Berend, 2011a, Detecting noun compounds and light verb constructions: a contrastive study, in proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE'11), pages 116–121.

IX. Nagy T., István; Berend, Gábor; Vincze, Veronika, 2011, Noun compound and named entity recognition and their usability in keyphrase extraction, in proceedings of RANLP 2011, Hissar, Bulgaria.

X. Mohammad Sadegh Rasooli, Heshaam Faili, and Behrouz Minaei-Bidgoli, 2011a, Unsupervised identification of Persian compound verbs, in proceedings of the Mexican international conference on artificial intelligence (MICAI), pages 394–406, Puebla, Mexico.

XI. Vivek Dubey, Pankaj Raghuwanshi, Sapna Vyas, Impact of Multiword Expression in English-Hindi Language, in proceedings of the International Journal of Emerging Trends & Technology in Computer Science (IJETTCS),Volume 4, Issue 3, May-June 2015,ISSN 2278-6856,pp. 101-105.

XII. Sina ZarrieB and Jonas Kuhn, 2009, Exploiting Translational Correspondences for Pattern-Independent MWE Identification, in proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications, pages

XIII. Sinha R. Mahesh K., 2009, Mining Complex Predicates In Hindi Using A Parallel HindiEnglish Corpus, Multiword Expression Workshop, Association of Computational Linguistics, International Joint Conference on Natural Language Processing-2009, pp. 40-46, Singapore.

XIV. Attia, Mohammed, Antonio Toral, Lamia Tounsi, Pavel Pecina and Josef van Genabith, 2010, Automatic extraction of Arabic multiword expressions. In Proceedings of the 7th Conference on Language Resources and Evaluation, LREC-2010, Valletta, Malta.

XV. Yulia Tsvetkov and Shuly Wintner, 2010, Extraction of multi-word expressions from small parallel corpora, Coling 2010: Poster Volume, pages 1256–1264, Beijing, August 2010.

XVI. https://en.wikipedia.org/wiki/Precision_and_recall.

XVII. http://www.manythings.org/bilingual/hin/

XVIII. http://nlpdotnet.com/Services/Tagger.aspx

XIX. http://text-processing.com/demo/tag/.

Vivek Dubey is the Principal of Alpine Institute of Technology Ujjain, MP, India. He is the incharge of NLP Laboratory at the Institute. He did BE (CSE), M.Tech. (CT) and Ph.D. in Computer Science & Engg. He has 15 years of engineering teaching experience, 3 years industry experience and 7 years in other. He has published around 45 papers in various national and international journals/conferences. He is also Editor and Reviewer in various journals.

ISSN (online): 2250-141X
Vol. 5 Issue 3, October 2015

Pankaj Raghuwanshi is working in Alpine Institute of Technology as Project Assistant in the department of Computer Science Engineering. He received BE degree in Mahakal Institute of Technology.

Sapna Vyas is a Ph. D. Scholar of Pacific University, Udaipur, Rajasthan. She completed her MCA in 2013 from RGPV, Bhopal, M.P. She has participated in college projects- HR Summit, Indore and CSI Votting. Her interest is in Artificial Intelligence, Data Mining, and Text Processing.