Volume 6, Issue 7, July 2016 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Identification, Extraction and Translation of Multiword Expressions

Avinash Singh, Dr. Shubhnandan S Jamwal
Department of Computer Science & IT, University of Jammu,
Jammu and Kashmir, India

Abstract— The main objective behind this research work is to work on English-Dogri parallel corpus and translate the multiword expressions from the English to Dogri language and vice versa. In machine translation when different types of multiword expressions exist in the sentences they pose challenges for the machine translation systems. Multiword expressions are distinctive word usages of a language which have non-compositional meaning. The technique used in the translation of the MWE's is based on the statistical approach in this paper. This paper represents study and analysis of different types of multiword expressions and how multiword expressions are translated with the help of parallel corpora using the Moses. The English-Dogri parallel corpus consists of 83,618 sentences and there are approximately 80,000 multiword present in it.

Keywords— Natural Language Processing, Machine Translation, Multiword Expressions, Parallel Corpus, Moses, BLEU

I. INTRODUCTION

Machine translation is the translation from one language into another language [1]. The idea of translation came into exist in 1947. Newer machine translation techniques are being developed used for yielding better results. Machine translation is the biggest applications of the natural language processing (NLP). NLP is the ability of a computer program to understand human speech as it is spoken. Multiword expressions are distinctive word usages of a language which have non-compositional meaning. The knowledge of multiword expressions is necessary for the NLP tasks like natural language generation, named entity recognition, machine translation etc. It has been described as 'A pain in the neck for the NLP' [2]. Multiword expressions are lexical items that consist of more than one word. MWE's are recurrent word combinations. MWE's are expressions consisting of two or more words that correspond to some conventional way of saying things [3]. A pair of words is considered to be a collocation if one of the words significantly prefers a particular lexical realization of the concept.

Example of MWE's:

- The *bus driver* accidentally hit the *garbage bin.*
- Kim *took* her pen *out.*
- उस समां उनके बब्ब *मैडिकल कालेज* च विद्यार्थी गै हे।
- *बक्ख- बक्ख* खाध पदार्थ।

India is a multilingual country having 22 official languages and 12 scripts, a lot of research have already been done on some Indian languages such as Hindi, Bengali, Manipuri, Punjabi etc. Jammu and Kashmir has a diversity of languages. The languages which are majorly spoken in it are: Dogri, Urdu and Kashmiri. Very few research have been done in Dogri language. Dogri language uses the script of Devanagari. Dogri is spoken in the Jammu region of J&K state and adjoining areas of Punjab, Himachal Pradesh and in the borders of Sialkot & Shakar Ghar tehsils in Pakistan.

II. TYPES OF MWE'S

Multiword expressions are classified into two types [1] [2]. 1) Lexicalized phrases have some form of added meaning to the structure. Lexicalized phrases are classified into three types. Fixed expressions are completely frozen and do not go any modifications at all e.g. in short, of course. Semi fixed expressions have restrictions on the structure of the phrase but might have some form of lexical variations. These are further classified into 3 types. Non-Decomposable idioms which are semi-fixed and do not undergo any syntactic variations but might allow some lexical modification e.g. kick the bucket to kicked the bucket. Compound nominal do not have syntactic modifications but allow lexical inflections for number e.g. car park. Named entities are syntactically highly idiosyncratic. These entities are formed based on generally a place or a person e.g. Gujarat lions. Syntactically flexible expressions allow huge variety of syntactic variations. Verb-Particle construction consists of a main verb and a particle e.g. Look up. Decomposable idioms are syntactically flexible and behave like semantically linked parts e.g. spill the beans. Light-Verb constructions have less semantic content are called light verbs as they can form highly idiosyncratic constructions with some nouns e.g. take a picture. 2) Institutionalized phrases are just fixed terms which do not have any second representations e.g. State bank of India.

III. 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].

IV. ENGLISH-DOGRI PARALLEL CORPUS

Parallel corpora have proved be an important resource for statistical machine translation. A parallel corpus contains a collection of texts in language and their translations into other languages. In most cases, parallel corpora contain data from only two languages where the texts, sentences and words are typically linked with each other. We have built parallel corpus of 83,618 sentences. In these corpora, there are approximately 80,000 multiword present in it. Procedure for developing parallel corpus is like this: firstly, we have collected the news from the Internet which is in Hindi text. Then these Hindi sentences are converted into the English sentences by the Google Translate. Then manual correction is done on the English text so that they have same meaning as in Hindi language. And for the Dogri sentences the Hindi sentences are converted into Dogri sentences, by the translators which had been developed by research scholar of the Department of Computer Science & IT, University of Jammu. Although that translator does not gives accurate results. Manually it is corrected and after that, it is checked by the linguistics. Finally we have parallel text of English-Dogri.

V. STATISTICAL MACHINE TRANSLATION USING MOSES

Moses is statistical machine translation system that allows to automatically train translation models for any language pair. All we must have is a collection of parallel text. Moses is an implementation of the statistical approach to machine translation [6]. This approach is now mostly used in this field at the moment. A collection of tools is used by Moses such as GIZA++ and KenLM. GIZA++ is used to perform word alignment over the parallel corpora. The alignments are then used to learn the phrase transliteration probabilities. KenLM is a toolkit for building and applying statistical language models [10]. We have used KenLM to build statistical language models by moses.

Language Model

A language model gives the probability of a sentence computed using 3-gram. It can be considered as computation of the probability of single word given all of the word that precedes it in a sentence.

Translation Model

It is trained using the parallel corpus of target-source pairs. As there is no corpus is big enough to allow the computation translation model probabilities at sentence level, so the process is broken down into smaller units e.g., words or phrases and their probabilities.

Decoder

This phase of SMT maximizes the probability of translated text. The words are chosen which have maximum like hood of being the translated translation

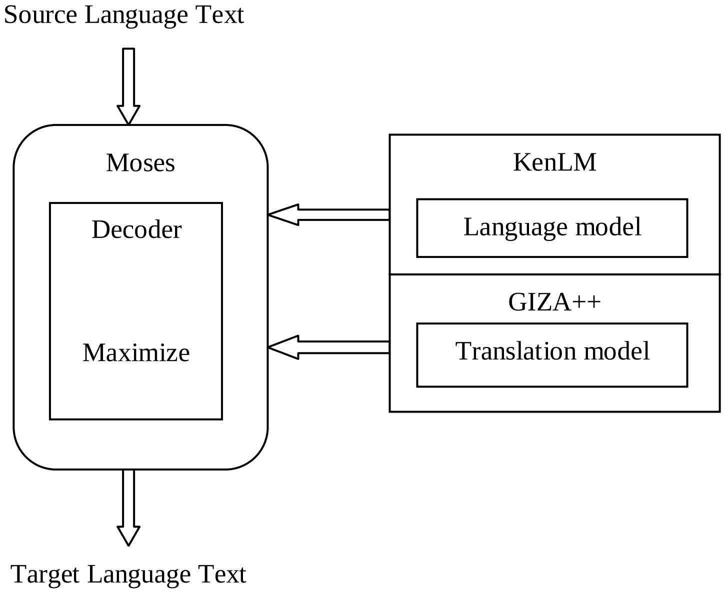


Figure : Architecture of SMT system

VI. RESEARCH METHODOLOGY

This section includes corpus collection, data preparation, development of language model, translation model and training of decoder using moses tool.

A. Corpus Preparation:

For this, we have built English-Dogri parallel corpus with 83,618 sentences. This corpus consists of small sentences related to novel, cricket, story, sports, travel, tourism and currently news.

To prepare the data for the training the translation system, we have to perform the following steps:

Tokenization: This means that we have to add the spaces between the words and the punctuation. For the Dogri language the tokenizer has not built so far. We have built the tokenizer for Dogri.

Truecasing: In this counting of word take place i.e. there no. of occurrence of words in the parallel corpus.

Cleaning: Long sentences are removed from the parallel corpus. We have limited the sentences length to the 80 words. The limits can be changed.

B. Language Model Training:

The model is created with a number of variables that can be adjusted to enable better translation. The models are building with the target language as it is important for it to know the language it outputs should be structured. This will build 3-gram language model.

C. Training the Translation System

Finally we come to our important phase-training the translation model. This will run word alignment, phrase extraction and scoring, create lexicalized reordering tables. This process creates the “moses.ini” file. We can use this file to decode. The file contains default parameter. These parameters will change after the tunning.

D. Tuning

Tuning is the slowest part of the process. Tunning requires a small amount of parallel data. We are going to tokenize and truecase it. Now we again go back to the training directory and execute the tuning process.

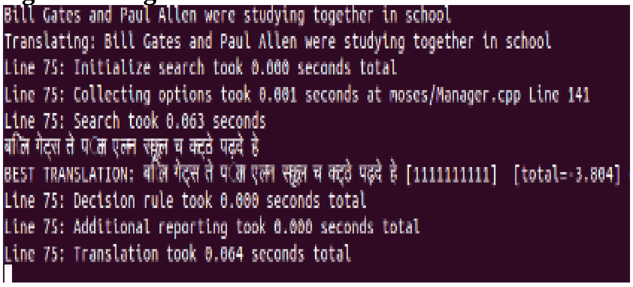
E. Testing

We can now run moses with the command: ~/mosesdecoder/bin/moses –f ~/work/mert-work/moses.in.

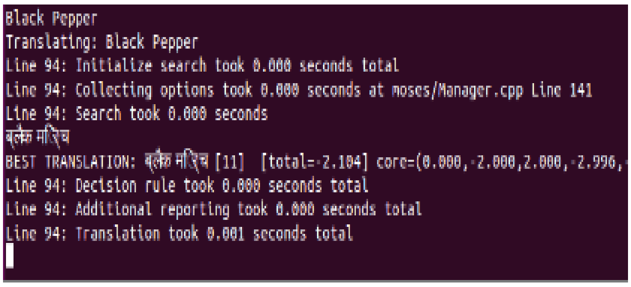
VII. RESULTS AND ANALYSIS

English-Dogri parallel corpus of 83,618 sentences has been implemented in this work, which comprises approximately 80,000 multiword. The parallel corpus is used by the moses in the translation of the text, where text can be English or Dogri. Firstly, input is given in English to the moses and the result of which is in Dogri. In addition to the output it shows the values of the parameter like decision rule, additional reporting and translation time. We have implemented 800 test sentences of different fields and the results are analyzed that are generated by the system. Out of these 800 test sentences, 600 sentences are correctly matched with the English-Dogri patterns. In the incorrect 200 translated sentences, there are misplacements of the word position and also presence of some English content which is not translated. The system gives accuracy of 75% in translating English to Dogri. Secondly, the Dogri input is given and the output of which is in English. For this evaluation of Dogri-English system, 678 test sentences were taken and the results were analyzed. Out of these 678 sentences, 542 sentences are correctly matched with the Dogri-English patterns. This system gives accuracy of 80% in translating Dogri to English system.

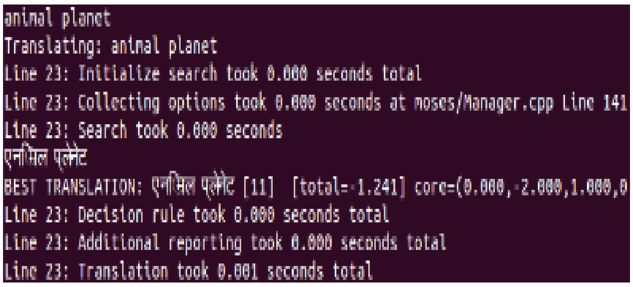
English to Dogri Translation



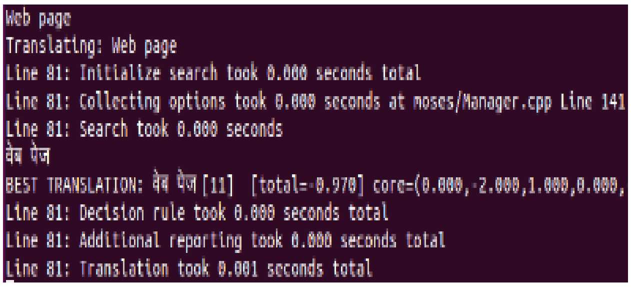
```
Bill Gates and Paul Allen were studying together in school
Translating: Bill Gates and Paul Allen were studying together in school
Line 75: Initialize search took 0.000 seconds total
Line 75: Collecting options took 0.001 seconds at moses/Manager.cpp Line 141
Line 75: Search took 0.063 seconds
बिल गेट्स ते पॉल एलन स्कूल च कट्ठे पढ़दे हे
BEST TRANSLATION: बिल गेट्स ते पॉल एलन स्कूल च कट्ठे पढ़दे हे [111111111]  [total=-3.804]
Line 75: Decision rule took 0.000 seconds total
Line 75: Additional reporting took 0.000 seconds total
Line 75: Translation took 0.064 seconds total
```



```
Black Pepper
Translating: Black Pepper
Line 94: Initialize search took 0.000 seconds total
Line 94: Collecting options took 0.000 seconds at moses/Manager.cpp Line 141
Line 94: Search took 0.000 seconds
ब्लैक मरिच
BEST TRANSLATION: ब्लैक मरिच [11]  [total=-2.104] core=(0.000,-2.000,2.000,-2.996,
Line 94: Decision rule took 0.000 seconds total
Line 94: Additional reporting took 0.000 seconds total
Line 94: Translation took 0.001 seconds total
```

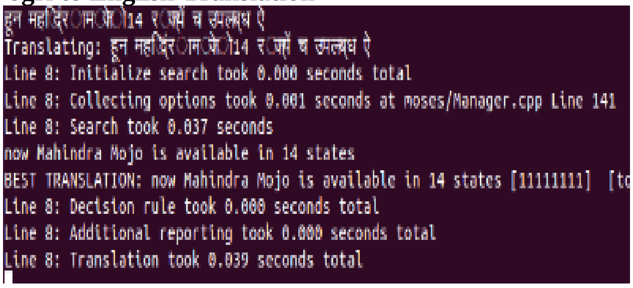


```
animal planet
Translating: animal planet
Line 23: Initialize search took 0.000 seconds total
Line 23: Collecting options took 0.000 seconds at moses/Manager.cpp Line 141
Line 23: Search took 0.000 seconds
एनिमल प्लेनेट
BEST TRANSLATION: एनिमल प्लेनेट [11]  [total=-1.241] core=(0.000,-2.000,1.000,0
Line 23: Decision rule took 0.000 seconds total
Line 23: Additional reporting took 0.000 seconds total
Line 23: Translation took 0.001 seconds total
```

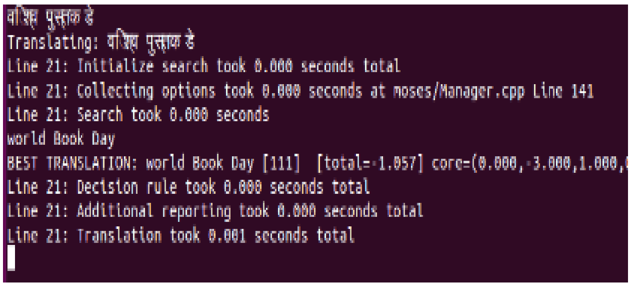


```
Web page
Translating: Web page
Line 81: Initialize search took 0.000 seconds total
Line 81: Collecting options took 0.000 seconds at moses/Manager.cpp Line 141
Line 81: Search took 0.000 seconds
वेब पेज
BEST TRANSLATION: वेब पेज [11]  [total=-0.970] core=(0.000,-2.000,1.000,0.000,
Line 81: Decision rule took 0.000 seconds total
Line 81: Additional reporting took 0.000 seconds total
Line 81: Translation took 0.001 seconds total
```

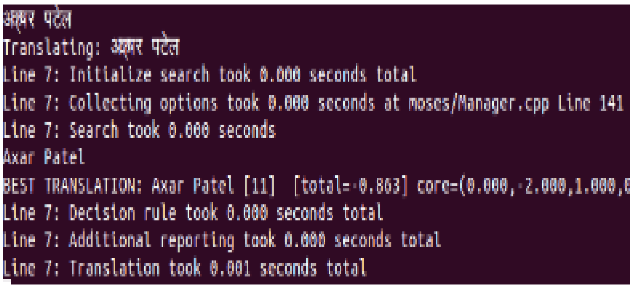
Dogri to English Translation



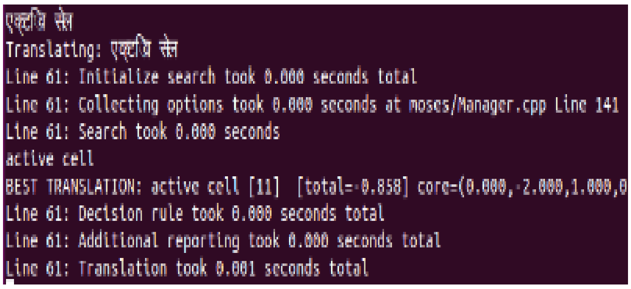
```
हुन महिंद्रा मोजो 14 रज्जें च उपलब्ध ऐ
Translating: हुन महिंद्रा मोजो 14 रज्जें च उपलब्ध ऐ
Line 8: Initialize search took 0.000 seconds total
Line 8: Collecting options took 0.001 seconds at moses/Manager.cpp Line 141
Line 8: Search took 0.037 seconds
now Mahindra Mojo is available in 14 states
BEST TRANSLATION: now Mahindra Mojo is available in 14 states [11111111]  [to
Line 8: Decision rule took 0.000 seconds total
Line 8: Additional reporting took 0.000 seconds total
Line 8: Translation took 0.039 seconds total
```



```
विश्व पुस्तक डे
Translating: विश्व पुस्तक डे
Line 21: Initialize search took 0.000 seconds total
Line 21: Collecting options took 0.000 seconds at moses/Manager.cpp Line 141
Line 21: Search took 0.000 seconds
world Book Day
BEST TRANSLATION: world Book Day [111]  [total=-1.057] core=(0.000,-3.000,1.000,
Line 21: Decision rule took 0.000 seconds total
Line 21: Additional reporting took 0.000 seconds total
Line 21: Translation took 0.001 seconds total
```



```
अक्षर पटेल
Translating: अक्षर पटेल
Line 7: Initialize search took 0.000 seconds total
Line 7: Collecting options took 0.000 seconds at moses/Manager.cpp Line 141
Line 7: Search took 0.000 seconds
Axar Patel
BEST TRANSLATION: Axar Patel [11]  [total=-0.863] core=(0.000,-2.000,1.000,0
Line 7: Decision rule took 0.000 seconds total
Line 7: Additional reporting took 0.000 seconds total
Line 7: Translation took 0.001 seconds total
```



```
एक्टिव सेल
Translating: एक्टिव सेल
Line 61: Initialize search took 0.000 seconds total
Line 61: Collecting options took 0.000 seconds at moses/Manager.cpp Line 141
Line 61: Search took 0.000 seconds
active cell
BEST TRANSLATION: active cell [11]  [total=-0.858] core=(0.000,-2.000,1.000,0
Line 61: Decision rule took 0.000 seconds total
Line 61: Additional reporting took 0.000 seconds total
Line 61: Translation took 0.001 seconds total
```

To test, how good the translation system is, another parallel data of 2,000 sentences has been built. Bilingual Evaluation Understudy or BLEU is one of the most popular metric for automatically evaluating machine translation system output quality. The central idea behind this metric is, how closer the machine translation is to a professional human translation. The primary programming task in a BLEU implementation is to compare n-grams of the candidate with the n-grams of the reference translation and count the number of matches. These matches are position-independent. The more the matches, the better is the candidate translation. BLEU score of English to Dogri Translation system is 22.26 while that of Dogri to English translation system is 25.09.

VIII. CONCLUSION & FUTURE WORK

The English-Dogri machine translation system is based on the statistical approach. Method of machine translation is the difficult task because of the unavailability of English-Dogri corpus. Multiword expressions are distinctive word whose meaning cannot be determined from its word. MWEs are the key issue and a current weakness for natural language processing applications. So, they need to be handled properly. In future work, different type of MWEs can be extracted. Further, by adding more data in the parallel corpus, better results can be achieved.

REFERENCES

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

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

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

[4] Md J. Abedin, B. S. Purkayastha, "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, 2013.

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

[6] Suraiya Jabin, Suos Samak and Kim Sokphyrum, "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, 2013.

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

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

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

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

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

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

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

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

[15] Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake and Dan Flickinger, "Multiword Expressions: A Pain in the Neck for NLP".

[16] Pallavi and Dr. Anitha S Pillai, "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, 2013.

[17] Available from: http://www.statmt.org/moses/?n=Moses.Baseline

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

[19] Availabe from: www.ciil-lisindia.net/Dogri/Dogri.html.