English-Dogri Translation System using MOSES

Avinash Singh
M.Tech Student
Department of Computer Science & IT, University of Jammu
Jammu, India

Asmeet Kour
M.Tech Student
Department of Computer Science & IT, University of Jammu
Jammu, India

Shubhnandan S. Jamwal
PhD, Assistant Professor
Department of Computer Science & IT, University of Jammu
Jammu, India

ABSTRACT
The objective behind this paper is to analyze the English-Dogri parallel corpus translation. Machine translation is the process of translating one language into another language. Machine translation is the biggest application of the Natural Language Processing (NLP). Moses is statistical machine translation system that allows to train translation models for any language pair. We have developed a translation system using Statistical based approach which helps in translating English to Dogri and vice versa. The parallel corpus consists of 98,973 sentences. The system gives accuracy of 80% in translating English to Dogri and the system gives accuracy of 87% in translating Dogri to English system.

Keywords
BLEU, Machine Translation, Moses, Statistical Approach, Parallel corpus.

1. INTRODUCTION
NLP [2] is the ability of a computer to understand human speech as it is spoken. Natural language processing is a field of Artificial Intelligence that deals with the methods of communicating with computers in natural languages like English, Hindi, Dogri etc. [17]. Machine Translation [1] is the process of translating one language to another language. Machine translation is the biggest and important applications of the natural language processing (NLP). The state of Jammu and Kashmir has a diversity of languages. The languages majorly spoken in the state are: Urdu, Dogri and Kashmiri [3]. Dogri is an Indo-Aryan language. Dogri is mother tongue of 422 million people. It is the second prominent language of J&K State. The idea of translation came into existence by 1947. Newer approach of machine translation techniques have been developed for yielding better results. The translation of Hindi and English text into Urdu is available online. No work of machine translation has been done on Dogri language that is the reason we are motivated towards it. Dogri is written using Devanagari script and has thirty eight segmental and five supra segmental phonemes. Segmental phonemes have been divided into two broad groups i.e. vowels and consonants. It has ten vowel phonemes and twenty eight consonant phonemes [3][4]. The vowels, consonants and matras use in the Dogri language are:

1.1 Vowel Symbols

| अ | आ | इ | ई | उ | ऊ | ए | ऐ | ओ | औ |

1.2 Consonants

क, ख, ग, घ, ङ, च, छ, ज, झ, ञ, ट, ठ, ड, ढ, ण, त, थ, द, ध, न, प, फ, ब, भ, म, म्, य, र, ल, व, श, ष, स, ह, ङ, छ, च, ङ |

1.3 Matras

| अ | आ | इ | ई | उ | ऊ | ए | ऐ | ओ | औ |

2. CHALLENGES
Some of the major challenges faced are:

- Lack of lexical resources.
- Multiple translations in English for Dogri words. There are many Dogri words which have different meanings depending upon the context in which the word is present in the sentence.
- Identifying multiword expressions present in the text like State Bank of India, Railway Station.
- Identifying name entities present in the text like Avinash, Asmeet, Sunil Dutt.
- Collection of phrases that cannot be translated word by word and these have different meaning in context.
- Handling grammatical errors after translation.

3. LITERATURE REVIEW
Shubhnandan S Jamwal and Sunil Dutt [1] tested the parameters like Translation Table size, Stack size, Language model, reordering model and word penalty for better speed and quality of Moses. SMT is the dominant approach in the field of machine translation. In SMT translation is trained on large quantities of parallel data. Collection of sentences in two different languages are known as parallel data, which is sentence-aligned, i.e. sentence in one source language is corresponding to translated sentence in the target language which is also known as bi-text. There are two approaches to...
SMT i.e. phrase-based and syntactic based. Tuning is the process by which speed and quality of translation is improved. There are various parameters that affect the performance of Moses. The author observed that if the size of the stack is limited up to 10 the output remains same with log probability 28.923 which is the best scoring by the model with best translation.

Rakesh Chandra Balabantary and Deepak Sahoo [5] used the popular phrase-based SMT techniques for the task of machine transliteration, for Odia-English and Odia-Hindi language pair. GIZA++ is used to perform word alignments over parallel corpora. Moses is an SMT system that allows automatically train translation models for any language. Moses gives better principle methods, both for learning useful phrases and combining them in the process of transliteration. For Odia language Moses model based on syllable based split gives good result than the model generated based on character based split. Author achieved accuracy of 89% for Odia-English and 86% for Odia-Hindi on syllable based split and 70% for Odia-English and 85% for Odia-Hindi on character cased split.

Nayan Jyoti Kalita and Baharul Islam [6] used Moses for Bengali to Assamese machine translation. Other translation tools like IRSTLM for language model and GIZA for translation model are utilized within this framework which is accessible in Linux situations. They have developed a parallel corpus of 17100 sentences in Bengali and Assamese. The statistical machine translation (SMT) system is based on the view that every sentence in a language has a possible translation in another language. A sentence can be translated from one language to another in many possible ways. A language model gives the probability of a sentence computed using n-gram model.

Nguyen Quang Phuoc, Yingxiu Quan and Cheol-Young Ock [7] attempted at constructing a bidirectional English-Vietnamese statistical translation system using Moses, an open-source toolkit for statistical machine translation. The quality of a statistical machine translation system depends on the parallel corpus. The language model is trained on the huge amounts of monolingual corpus in target language. In case of English to Vietnamese translation, the BLEU score is higher than Google BLEU score 14.5% and higher than Bing BLEU score 32.3% since using the large amount of Vietnamese monolingual corpus with over 11 million sentences to train the Vietnamese language model. In case of Vietnamese to English translation, the author used English sentences that were extracted from parallel corpus to train English language model, the BLEU score is almost the same as Bing’s one and higher than Google’s score 17%.

4. MOSES

Moses is statistical translation system that train models for any language pair. For training the system we should have a collection of parallel text [8]. This approach is now mostly used in this field at the moment. A collection of tools used by Moses are GIZA++ and KenLM. GIZA++[17] is used to perform word alignment over the parallel corpora. The alignments are used to learn the phrase transliteration probabilities [9]. KenLM is a toolkit for building and applying statistical language models. We have used KenLM to build statistical language model by Moses.

Language Model

The language model gives the probability of a sentence computed using tri-gram. It can be considered as computation of the probability of single word given all of the words that precede it in a sentence [9].

Translation Model

It is trained using the parallel corpus of target-source pairs. The translation model [8] can be phrase-phrase rules, or hierarchical (perhaps syntactic) rules these can be compiled into a binarised form so as to enable fast loading. These models can be supplemented or appended with features to add extra information to the translation process.

Decoder

Decoder maximizes the probability of translated text. The words are chosen that have maximum possibility of being the translated translation [10].

5. STEPS IN TRANSLATION

For translation using statistical approach using Moses different steps are needed, which are shown in this section. It includes corpus preparation, development of language model, translation model and training of decoder using Moses tool.

5.1 Corpus Preparation

Parallel corpus of 98,973 entries had been built. The parallel corpus consists of 30,000 named entities, 35,000 multiwords and remaining are one line sentences.

**Fig.5.1: English Corpus**

**Fig.5.2: Dogri Corpus**

Tokenization: This means that spaces have to be inserted between words and punctuation. We have done tokenization over 98,973 sentences of both English and Dogri language pair. For tokenization of English sentences we have used the English tokenizer and for the tokenization of Dogri sentences we have developed Dogri tokenizer.

Results after the tokenization are shown in the screenshots:

**Fig.5.3: English tokenized data**
5.3 Training the Translation System
Finally comes an important phase—training the translation model. This will run word alignment (using GIZA++) phrase extraction and scoring, create lexicalized reordering tables and create Moses configuration file [13] [14]. Once it is finished, a moses.ini file will be created in the directory /work/train/model. This moses.ini file is used to decode.

5.4 Tuning
A small amount of parallel data is to be collected which is different from the training data. First step is to tokenize and truecase it, just as we have done in the training process. Now going back to the training directory and then execute the tuning process [15]. After the tuning process is finished, a moses.ini file is created with train weights, which is in the directory ~/work/mertwork/ moses.ini and default values are changed now.

5.5 Testing
Run moses with the command: ~/mosesdecoder/bin/moses –f ~/work/mert-work/moses.ini.
6. RESULTS

Results of our system are as follows:

![Fig.6.1: Dogri to English Translation](image1)

![Fig.6.2: English to Dogri Translation](image2)

The system gives accuracy of 80% in translating English to Dogri and the system gives accuracy of 87% in translating Dogri to English system. 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 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. 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.

7. CONCLUSION & FUTURE WORK

Machine translation is the translation from one language into another language. Machine translation is the biggest application of the Natural Language Processing (NLP). 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. Translation with English-Dogri parallel corpus is not yet done [16]. The efficiency of this approach can be further improved by including more English-Dogri parallel corpus sentences. So, in the future course of this study, more parallel corpus sentences can be included and work on different type of MWE’s could be done to achieve better results than the present system.

8. REFERENCES


