Hindi Chhattisgarhi Machine Translation System Using Statistical Approach

Vikas Pandey*
Department of Information Technology, Bhilai Institute of Technology, Durg, India.
E-mail: vikas.pandeymtech2009@gmail.com

Dr.M.V. Padmavati
Department of Computer Science and Engineering, Bhilai Institute of Technology, Durg, India.
E-mail: vmetta@gmail.com

Dr. Ramesh Kumar
Department of Computer Science and Engineering, Bhilai Institute of Technology, Durg, India.
E-mail: rk_bitd@rediffmail.com

Received November 05, 2020; Accepted December 15, 2020
ISSN: 1735-188X
DOI: 10.14704/WEB/V18SI02/WEB18067

Abstract

Machine Translation is a subfield of Natural language Processing (NLP) which uses to translate source language to target language. In this paper an attempt has been made to make a Hindi Chhattisgarhi machine translation system which is based on statistical approach. In the state of Chhattisgarh there is a long awaited need for Hindi to Chhattisgarhi machine translation system for converting Hindi into Chhattisgarhi especially for non Chhattisgarhi speaking people. In order to develop Hindi Chhattisgarhi statistical machine translation system an open source software called Moses is used. Moses is a statistical machine translation system and used to automatically train the translation model for Hindi Chhattisgarhi language pair called as parallel corpus. A collection of structured text to study linguistic properties is called corpus. This machine translation system works on parallel corpus of 40,000 Hindi-Chhattisgarhi bilingual sentences. In order to overcome translation problem related to proper noun and unknown words, a transliteration system is also embedded in it. These sentences are extracted from various domains like stories, novels, text books and news papers etc. This system is tested on 1000 sentences to check the grammatical correctness of sentences and it was found that an accuracy of 75% is achieved.

Keywords

Chhattisgarhi, Hindi, Machine Translation, Moses, Corpus.

Introduction

Machine Translation (MT) is an automated system that converts source language to target language without any human intervention [1]. India, which is a multi linguistic country having total 179 languages and 544 dialects machine translation system can play a very important role in conversion of digital information from one language to another language. People of one state when visit other state in form of student, government officer and tourist generally face language problem when they access digital information for their interest which is in regional language of that state [2]. Machine Translation plays an important role in solving this problem, as it can convert a source language in target language [3].Thus machine translation system helps in breaking the language barrier. A Simple machine translation system a only perform word substitution and do not work on alignment of words and their reordering as shown in table 1

Table 1 Word to Word Substitution in Simple Machine Translation System

English Sentence	Vijay	Plays	Cricket
Hindi Sentence	विजय	खेलता है	क्रिकेट
Syntactic rearrangement	विजय	क्रिकेट	खेलता है

In Simple machine translation system alignment and rearrangement of words is an important issue. To overcome this limitation basically four different machine translation approaches are used: Rule based approach, Corpus based approach, Hybrid approach and Neural Network approach [4][5].

In Rule Based approach various rules related to the syntax of the language are used which helps in conversion from source to target language. It is based on the specification of rules for morphology, syntax, lexical selection, semantic analysis, and transfer and generation process [6]. AnglaHindi MT system was developed by IIT, Kanpur in year 2003 is based on Rule Based MT approach.

In Corpus based approach big bilingual corpus is required for translation purpose [7][8]. The corpus based approach is further classified into two types: Statistical Machine Translation and Example Based Machine Translation. A collection of structured text to study linguistic properties is called corpus. It is based on the principle of n-gram which is a continuous sequence of n items from a given sequence of text or speech. Statistical Machine Translation system is based in conditional probability called Bayes Probability Rule:

$$P(C|D)=\frac{P(C,D)}{P(D)}$$ Where 0 <= P (C) <=1

P (C)-Probability that word C present in the text
P (C, D)-It is the probability that words C and D are present in the text
P (C|D)- It is the probability that word C is presents in the text when D is already present in the text

Let's consider the example of Hindi to Chhattisgarhi SMT system. Every Chhattisgarhi sentence *C* is a possible translation of a Hindi sentence *H*. P (C|H), which is the probability that a translator when presented with a Hindi sentence H, will produce *C* as its Chhattisgarhi translation. The goal of SMT is to find the sentence *C* that the native speaker in his mind when he produces *H*. Hindi to Chhattisgarhi SMT system can be modeled by Bayes Probability Rule:

$$P(H \mid C) = \frac{P(C \mid H)P(H)}{P(C)} \propto P(C \mid H)P(H)$$

Example based machine translation system uses previous translation examples to generate translations for an input sentence. When a sentence is inputted to the system, it first retrieves a similar source sentence from the example-base and its translation. The system then matches the example translation to generate the translation of the input sentence.

Chhattisgarhi is an Indo-Aryan language which is spoken in the state of Chhattisgarh by about 11.5 million speakers. This language is closely related to the Bagheli and Awadhi languages that are classified in the east central zone of the Indo-Aryan languages. Chhattisgarhi language like Hindi uses the Devnagari script. Chhattisgarhi is the official language of Chhattisgarh state. State government works in Chhattisgarhi languages. There is no machine translator available for Chhattisgarhi language. This is main motivation behind the design of Chhattisgarhi to Hindi machine translator. The research will try to find the different issues involved in the translation. There are five Types of Chhattisgarhi spoken in Chhattisgarh State [14][15]. These are as follows:

a. Northern Chhattisgarhi (Sargujia)-This type of Chhattisgarhi is spoken in Raigarh, Jashpur, Korea northern part of Sarguja district. It has effect of Bhojpuri and Bagheli.
b. Eastern Chhattisgarhi (Lariya)- This type of Chhattisgarhi is spoken in Mahasamund and eastern part of Raipur. It has effect of Oriya.
c. Southern Chhattisgarhi (Bastariya)- It is mixture of Habli and Gondi.It has effect of Gondi in south and Marathi in western part. It is Spoken in Baster and Dentewara district.

d. Western Chhattisgarhi (Khaltahi)-It is used in Kawardha, Western part of Rajnandgaon district and North West part of Bilaspur district. It has effect of Marathi and Bundeli.
e. Central Chhattisgarhi-It is used in western part of Kawardha and Rajnandgaon district. Aprt from the eastern part of Mahasamund and north west part of Bilaspur used in Korba, Jangagir, Durg, Dhamtari and Kanker.

Machine Translation Approaches

There are different types of machine translation approaches as discussed by (Arnold et al., 1994). Different machine translation approaches are shown in Figure 1. They include direct MT approaches, rule-based MT approaches, knowledge-based approaches and statistical MT approaches.

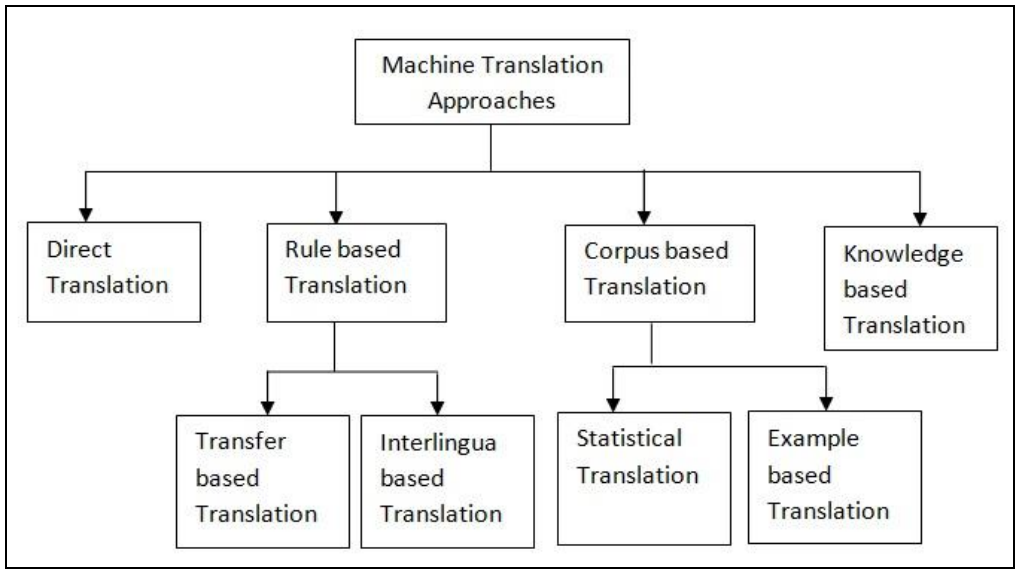

Figure 1 Machine Translation Classification

Direct Machine Translation

During 1950s the first MT approach invented was Direct MT approach. It is called direct approach as there is one to one mapping between source and target words. It requires use of bilingual dictionary for translation purpose.

Rule-Based Machine Translation

Due to use of different syntax and semantic rules, these systems are known as rule-based MT systems. The system use to parses the source text and produces a parse tree. The parse

tree structure also gets produced in target side based on target rule base, after which reordering of sentences take place using reordering module.

Corpus-Based Machine Translation

Corpus-based MT systems use to operate on set of parallel corpus. Corpus is structured information in text form. The corpus-based approach is of two types: statistical based (SMT) and example based.

SMT is based on works on bilingual text corpora and follows mathematical model of probability. In SMT source and target words are paired together having maximum probability.

The Example based Machine Translation (EBMT) approach came into existence in year 1984 and is based bilingual parallel corpus with parallel texts. EBMT follows the model that the new translation is based on previously translated sentences.

Literature Review

In India there are 30 languages that are recognized by Government of India, out of which there are 22 languages that come under article 8 of our constitution. These are the official state languages through which various administrative works can be done. These languages are used by state government for grafting public notices and also in public awareness drives. Various national level exams are conducted through theses languages. There are some languages like Malayalam, Bengali, Manipuri, Nepali, Oriya, Marathi, Bodo, Dogri, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Maithili, Punjabi, Sanskrit, Santali, Sindhi, Tamil, Telugu, Assamese, and Urdu [9].

The development of machine translation systems started in 90's in India and it finds its application in various areas like in administrative work, State Assemblies and Parliament, Education and News paper industry and Advertisement industry. There are various institutions like IITs and NITs playing important role in developing the machine translation systems. Various MT systems have been developed in India which is based on different approaches of Machine Translation (MT).

Research on machine translation started in India at IIT Kanpur. It proposed 'AnglaBharti' is a rule based multi-lingual MT system which converts English language to various Indian languages.

'AnglaBharti' a new rule based MT system was developed in 2004 to overcome the shortcomings of the previous architecture. Later, 'AnglaBharti-II' was developed which was hybrid MT system [4].

'Anusaaraka' MT system was developed by IIIT Hyderabad in 2003 and is based on the principles of Grammar given by Paninian, and similarity of Indian languages.

'MANTRA' MT system was developed in 1999, by C-DAC for translating documents from English to Hindi and is based on Tree Adjoining Grammar (TAG). It's works on auto-phrase-detection algorithms

Shiva and Shakti were two MT systems being developed by Carnegie Mellon University US, IISC, Bangalore and IIIT, Hyderabad in 2005, which use to convert English languages to Hindi, Marathi and Telugu languages [5].

English to Bengali Phrasal EBMT was proposed in 2006, based on shallow analysis and retrieves the target phrases by example based approach [9].

A Hindi to Punjabi direct machine translation system was proposed in 2010,based on direct MT approach in which Hindi Punjabi bilingual dictionary was used along with preprocessor and transliteration module [10][11][12].

Bengali to Assamese SMT approach was proposed in 2015 using Moses and other tools like GIZA++ which runs on open source platform. It works on Bengali-Assamese parallel corpus. They used different corpuses from various domains. The performance of translation system is based on BLEU score [13].

Approach of MT System

This section describes different software tools used for developing the proposed MT system. It also describes system architecture and description of each module of the system.

Tools to Develop MT System

Various open source software tools are used to develop Hindi Chhattisgarhi MT system which works on LINUX platform. MOSES, which is a phrasal based machine translation system, is used to perform the translation task as discussed in [16].Through this statistical machine translation tool, translation module can trained by using Hindi Chhattisgarhi parallel corpus as shown in Table 2.

Table 2 Hindi Chhattisgarhi parallel corpus

S. No.	HINDI SENTENCES	CHHATTISGARHI SENTENCES
1	मेरा नाम मोहित है।	मोर नाव मोहित हावे।
2	मैं खैरागढ़ में रहता हूं।	मे खैरागढ़ म रथो।
3	यह किताब है।	एहा किताब हावे।
4	बर्फ गिर रही है।	बरफ गिरत हावे।
5	मैं जा रहा हूं।	मे हा जावत हो।
6	तुम गाड़ी चला सकते हो क्या।	तै गाड़ी चला सकथस का।
7	भाजी मुरझा गया है।	भाजी अइला गेहे।
8	मैं अकेला हूं।	मे हा अकेल्ला हो।
9	मैं तुम्हें फोन करूंगा।	मे तोला फ़ोन करहु।

The working of MOSES is based on beams search as discussed by [17]. A bilingual dictionary can be produced from Hindi Chhattisgarhi Parallel corpus using GIZA++.In order to develop this MT system SRILM tool is used.

Architecture of MT System

In the Hindi Chhattisgarhi MT system a transliteration module is embedded into statistical MT model. This transliteration module will help us to improve the performance of MT system by transliterating name, place and organization and some foreign words. The system architecture is described in Figure 2.

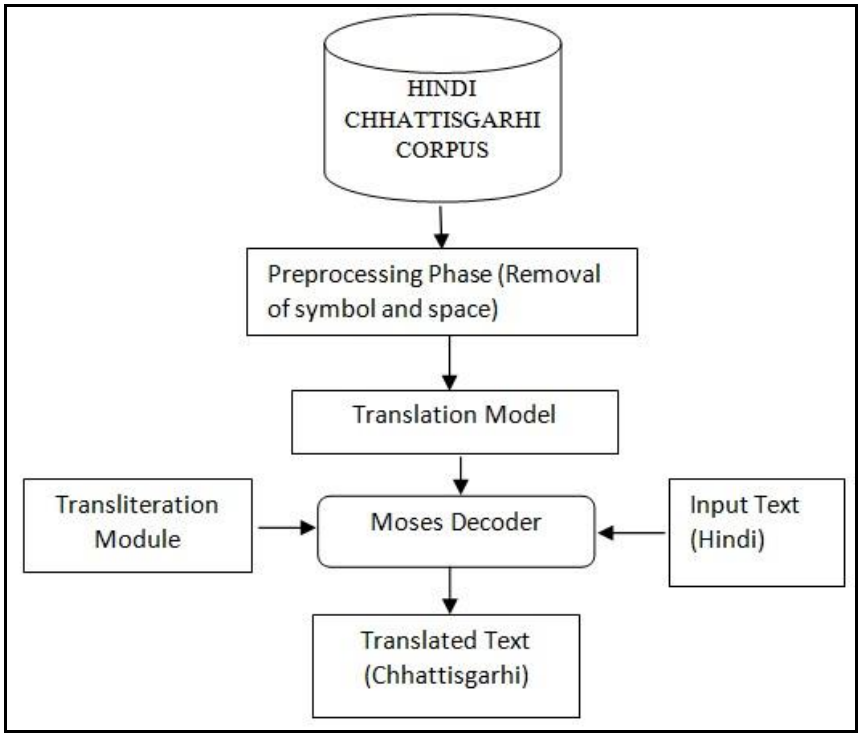

Figure 2 System Architecture of Hindi Chhattisgarhi SMT based Model

Corpus Data

Since digital documents for Chhattisgarhi language are not available in sufficient amount in public domain. So in order to prepare parallel corpus we take help of Chhattisgarhi Story books, text books, news papers and websites and various Chhattisgarhi sentences are collected from these domains as shown in Table3:

Table 3 Parallel Corpus Collected from Various Domains

S.NO.	Domains	Number of Sentences
1	News	10,000
2	Stories	20,000
3	Text Books	20,000

The process of creating parallel corpus is shown in Figure 3:

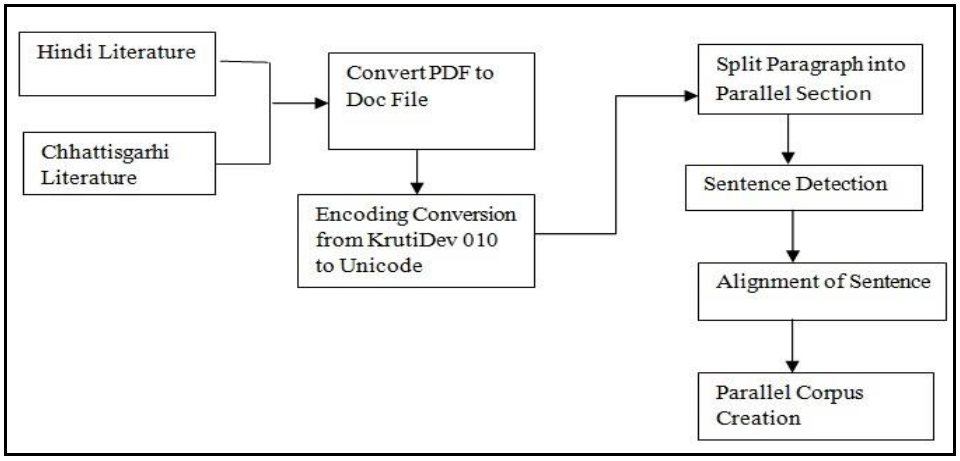

Figure 3 Creation of Parallel Corpus

In order to process this parallel corpus so that it become suitable for our MT system following steps are follows:

a. File Format Conversion – The Hindi font in Devnagari script collected from different sources were in KritiDev010 will be converted to Unicode format of UTF-16.
b. Corpus Cleaning - The parallel corpus will be cleaned by removing all blank spaces, dots, commas and dash lines.

Translation System

Hindi Chhattisgarhi statistical machine translation system runs on parallel corpus collected from various domains. The corpus consist of 40,000 Hindi Chhattisgarhi bilingual parallel

sentences. In order to develop this system various open source software tools were used. MOSES is used to perform machine translation task as discussed by [16]. This SMT tool is helpful to train the translation system of Hindi Chhattisgarhi MT system. MOSES is based on beam search algorithm that can find the highest probability of matching of a source word from a group of target words. Another Statistical machine translation tool GIZA++ was used to align the parallel corpus which is generally stored in form of alignment file discussed as discuss by [19] GIZA++ was developed in Johns-Hopkins University. The alignment file consists of Hindi Chhattisgarhi bilingual words along with their probability as shown in Table 4:

Table 4 Probability of Matching between Hindi Chhattisgarhi Words

S.NO.	HINDI	CHHATTISGARHI	PROBABILITY OF MATCHING
1	जितना	जितना	1
2	मुफ्त	फ़ोकट	1
3	रही	NULL	0.695286
4	रहा	NULL	0.992582
5	हम	हमन	1
6	हो	NULL	0.723232
7	पर	फेर	0.500011
8	कोई	कोनोच	0.333333
9	यहां	ए	0.571429
10	स्थान	ठउर	1
11	उसका	ओखर	0.666667
12	भाषा	भाषा	1
13	गाड़ी	गाड़ी	0.5
14	राम	राम	1
15	जाना	जाबे।	1
16	रहे	NULL	0.926236
17	मत	झन	1
18	से	तारा	0.250032
19	रुक	ठउर	1
20	आग	आगि	1
21	ऐसा	अइसन	1
23	जगह	ठउर	1
24	उसके	ओखर	1
26	उंगली	दिखत	0.318234
27	दिख	दिखत	1

GIZA++ an open source software tool can be installed by following command on Ubuntu platform as shown in Figure 4:

```
cd ~
cd Documents/transl/giza-pp/
./GIZA++-v2/plain2snt.out hindi cgi
ls
./mkcls-v2/mkcls -phindi -Vhindi.vcb.classes
./mkcls-v2/mkcls -pcgi -Vcgi.vcb.classes
ls
./GIZA++-v2/GIZA++ -S cgi.vcb -T hindi.vcb -C cgi_hindi.snt -o hindi_cg -outputpath ./
```

Figure 4 Commands for installation of GIZA++ and generation of different files

After installation of GIZA++ and running above commands following files will be generated as shown in Figure 5:

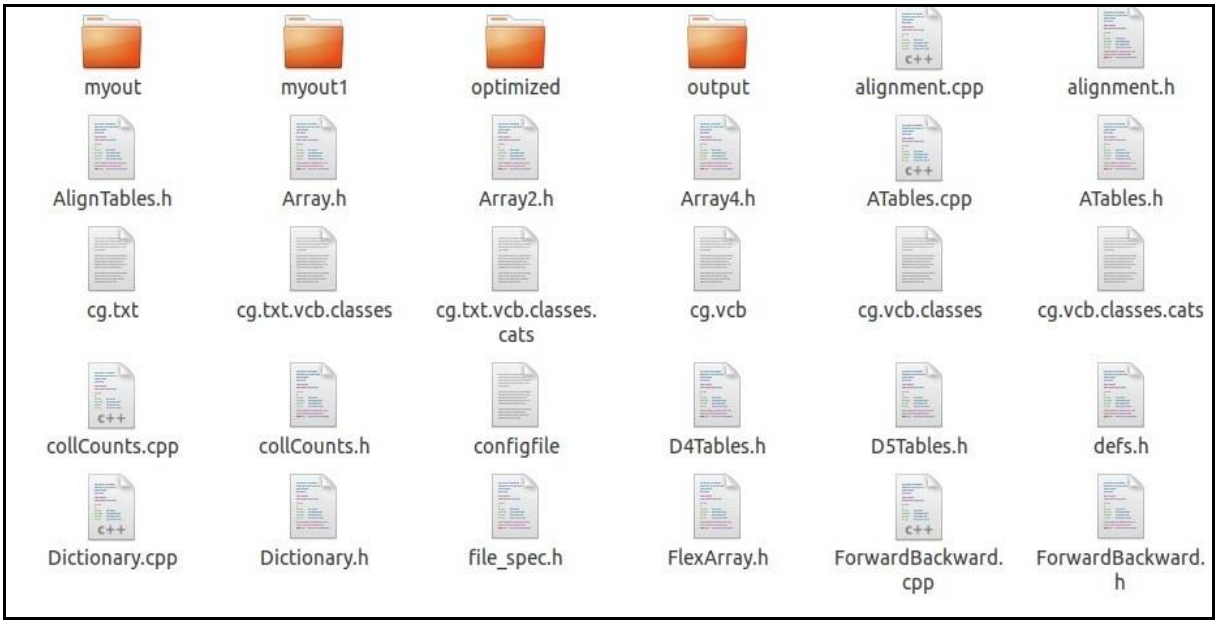

Figure 5 Generation of different files

Following steps are required for creation of files:

Step 1

Create a Hindi Chhattisgarhi parallel corpus.

Step 2

Create files needed for GIZA++:

(a) Run plain2snt.out located within the GIZA++package ./plain2snt.out cg hindi

(b) Files created by plain2snt: hindi.vcb, cg.vcb, cghindi.snt

The Files Created by plain2snt are as follows

hindi.vcb file contains:

i. Words from the Hindi corpus.

ii. Word frequency count.
iii. Word with unique id.

cg.vcb files consist of:

i. Chhattisgarhi word frequency count.
ii. Chhattisgarhi words with unique id.

cghindi.snt file consists of:

i. From each Hindi and Chhattisgarhi sentences, unique number for each word.

Step3

i. GIZA++ creates mkcls files by the help of following commands:

mkcls –phindi–Vhindi.vcb.classes mkcls –pcg–Vcg.vcb.class

ii. These files are created by mkcls: hindi.vcb.class, hindi.vcb.class.cats, cg.vcb.class, cg.vcb.class.cats.

The output of Hindi Chhattisgarhi Statistical Machine Translation system tested for some sentences is shown in Table5:

Table 5 Output of Hindi Chhattisgarhi Machine Translation System

S. no.	Hindi Sentences	Chhattisgarhi Sentences
1	मेरा नाम मोहित है।	मोर नाव मोहित हावे।
2	मैं खैरागढ़ में रहता हूं।	में खैरागढ़ म रथो।
3	यह किताब है।	एहा किताब हावे।
4	बर्फ गिर रही है।	बरफ गिरत हावे।
5	मैं जा रहा हूं।	में हा जावत हो।
6	तुम गाड़ी चला सकते हो क्या।	तै गाड़ी चला सकथस का।
7	भाजी मुरझा गया है।	भाजी अइला गेहे।
8	मैं अकेला हूं।	में हा अकेल्ला हो।
9	मैं तुम्हें फोन करूंगा।	में तोला फ़ोन करहु।
10	उसे छूना मत।	ओला छूबे झन।
11	मुझे दर्द हो रहा है।	मोला पीरा होवत हे।
12	कोई नहीं आया।	कोनोच नई आइस।

The above output is shown in Figure 6:

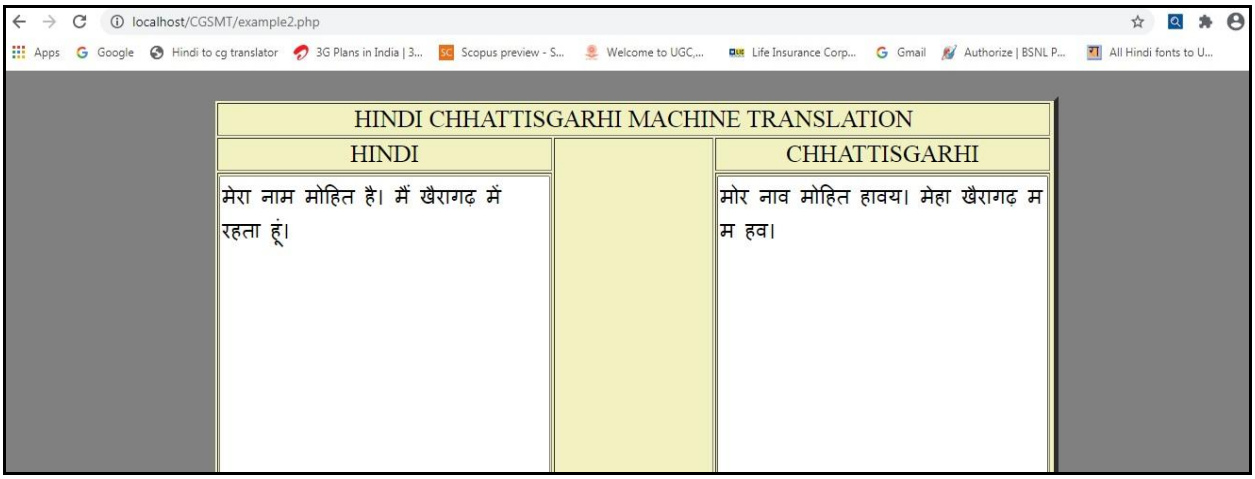

Figure 6 Hindi Chhattisgarhi Machine Translation System

System Evaluation

In case of Automatic evaluation metrics each sentence in system translation is compared against human translations. This human translation is called Reference translation. In order to evaluate the Hindi Chhattisgarhi machine translation system three parameters namely: Precision, Recall and F1 score are used.

Precision: For a translation task done by a machine translation system, precision is number of correct translated sentences divided by the total number of all translated sentences.

$P = (\text{relevant} \cap \text{retrieved}) / \text{retrieved}$

Recall: For a translation task done by a MT system, Recall is number of correct translated sentences divided by the total number of existing correct translated sentences.

$R = (\text{retrieved} \cap \text{relevant}) / \text{relevant}$

F-score is one of the important factors to judge the accuracy of machine translation system. It is calculated from the precision and recall. There are various types of F score, for the accuracy test F1score is being used which the harmonic mean of precision and recall.

The Precision, Recall and F1 score of Hindi Chhattisgarhi Statistical Machine Translation system calculated for some sentences is shown in Table 6:

Table 6 Precision, Recall and F1 score calculated for some sentences

S. No.	True Sentences	Predicted Sentences	Precision Value	Recall Value	F1 Score
1	मोर नाव मोहित बने।	मोर नाव मोहित हावे।	0.75	0.75	0.75
2	ए किताब बने।	एहा किताब हावे।	0.33	0.33	0.33
3	भाजी अइला होंगे बने।	भाजी अइला गे हे।	0.5	0.5	0.5
4	मेहा अकेल्ला हव।	मैं अकेल्ला हो।	0.33	0.33	0.33
5	मेहा तोला फ़ोन करहू।	मैं तोला फ़ोन करहू।	0.75	0.75	0.75
6	ओला छूबे झन।	ओला छूबे झन।	1	1	1
7	कोनो नई आइस।	कोनोच नई आइस।	0.66	0.66	0.66
8	मेहा भुला गैव।	मैं भुला गैव।	0.66	0.66	0.66
9	ए चिरई के खोंदरा बने।	ए चिरई के खोंदरा हावे।	0.8	0.8	0.8
10	चंदा बहुत दुरिहा बने।	चंदा अब्बड़ दुरिहा हावे।	0.5	0.5	0.5
11	और चाउर खा लेवा।	अउ चाउर खा लेवा।	0.75	0.75	0.75
12	एकर दे चानी कर देवा।	एकर दू चानी कर देवा।	0.8	0.8	0.8
13	तोला अभिचेच जाये चिये।	तोला अभिचेच जाना चिये।	0.75	0.75	0.75
14	अइसन झन बोल।	अइसन झन बोल।	1	1	1
15	तोर मया क मारे।	तोर मया के मारे।	0.75	0.75	0.75
16	कन किरिया खा।	तै किरिया खा।	0.66	0.66	0.66
17	ओला भीतर ले आओ।	ओला भीतर ले लान।	0.75	0.75	0.75
18	तोर करा कोनो चिन्ह बने।	तोर करा कोनो चिन्ह हावे।	0.8	0.8	0.8
19	ओहा मोर संगवारी बने।	ओ मोर संगवारी हे।	0.5	0.5	0.5
20	जेवण पाक होंगे बने।	जेवण पाक गे हावे।	0.5	0.5	0.5
21	सामान इति वोती सुते बने।	सामान इति वोती परे हावे।	0.6	0.6	0.6

Conclusion

In this research various approaches of developing a machine translation system has presented. It is observed that for low resource languages like Chhattisgarhi Statistical approach is more suitable. An attempt has been made to design and implement Chhattisgarhi-Hindi MT system which will be beneficial for Chhattisgarh state. Since each approach has its own pros and cons, the accuracy of the system depends on bilingual dictionary and size of parallel corpus. This machine translation system works on parallel corpus of 40,000 Hindi-Chhattisgarhi bilingual sentences. In order to overcome translation problem related to proper noun and unknown words, a transliteration system is

also embedded in it. These sentences are extracted from various domains like stories, novels, text books and news papers etc. This system is tested on 1000 sentences to check the grammatical correctness of sentences and it was found that an accuracy of 75% is achieved.

References

Antony, P.J. (2013). Machine translation approaches and survey for Indian languages. *In International Journal of Computational Linguistics & Chinese Language Processing, 18*(1), 47-48.

Batra, K.K., & Lehal, G.S. (2010). Rule based machine translation of noun phrases from Punjabi to English. *International Journal of Computer Science Issues (IJCSI)*, *7*(5), 409-412.

Arnold, D., Balkan, L., Meijer, S., Humphreys, R., & Sadler, L. (1994). *Machine translation: an introductory guide*. Blackwell Publishing Company. ISBN No.:185554217 XS.

Nair, L.R., & David Peter, S. (2012). Machine translation systems for Indian languages. *International Journal of Computer Applications*, *39*(1), 25-31.

Rao, D. (2001). Machine translation in India: A brief survey. *In proceedings of SCALLA Conference*, Bangalore, India.

Sinha, R.M.K., & Jain, A. (2003). Angla Hindi: an English to Hindi machine-aided translation system. *MT Summit IX, New Orleans, USA, 494*, 497, 1-5.

Baum, L.E. (1972). An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. *Inequalities*, *3*(1), 1-8.

Sinha, R.M.K., Sivaraman, K., Agrawal, A., Jain, R., Srivastava, R., & Jain, A. (1995). ANGLABHARTI: a multilingual machine aided translation project on translation from English to Indian languages. In *1995 IEEE International Conference on Systems, Man and Cybernetics. Intelligent Systems for the 21st Century*, *2*, 1609-1614. https://doi.org/10.1109/ICSMC.1995.538002.

Naskar, S.K., & Bandyopadhyay, S. (2006). A phrasal EBMT system for translating English to Bengali. *Proc. Workshop on Language, Artificial Intelligence, and Computer Science for Natural Language Processing Applications*, Bangkok, Thailand: 373-379.

Sinha, R.M.K., Shivaraman, K., & Agrawal, A. (1995). ANGLABHARTI: A Multilingual Machine Aided Translation Project on Translation from English to Indian Languages. *IEEE International Conference on Systems, Man and Cybernetics. Intelligent Systems for the 21st Century*, 1609-1616. https://doi.org/10.1109/ICSMC.1995.538002

Josan, G.S., & Lehal, G.S. (2012). Direct Approach for Machine Translation from Punjabi to Hindi. *CSI Journal of Computing*, *1*(1), 3-19.

Goyal, V., & Lehal, G.S. (2011). Hindi to Punjabi machine translation system. In *International Conference on Information Systems for Indian Languages*, 236-241. Springer, Berlin, Heidelberg.

Kalita, N.J., & Islam, B. (2015). Bengali to Assamese Statistical Machine Translation using Moses (Corpus Based). *International Conference on Cognitive Computing and Information Processing, 1*(1). https://doi.org/10.13140/RG.2.1.4331.7289.

Chandrakar (2010). *Manak Chhattisgarhi vyakaran.* Stakshi Publication. ISBN No.:8189545086.

Chandrakar (2012). *Vrihad Chhattisghari shabda kosh.* Chhattisgarh Hindi Granth Academy. ISBN No. 9788192169125.

Koehn, P. (2009). *Statistical machine translation*. Cambridge University Press. ISBN No: 2010 0521874157.

Barman, A., Sarmah, J., & Sarma, S.K. (2014). Assamese WordNet based Quality Enhancement of Bilingual Machine Translation System. *CSI Journal of Computing, 1*(1).

Stolke, A. (2002). SRILM-an extensible language modeling toolkit. *In Proceedings of the ICSLP*, Orlando, FL, United States.

Och, F.J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational linguistics*, *29*(1), 19-51.