International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 8 (2018) pp. 6394-6398

Issues in Chhattisgarhi to Hindi Rule Based Machine Translation System

Vikas Pandey[1], Dr. M.V Padmavati[2] and Dr. Ramesh Kumar[3]

[1]*Department of Information Technology, Bhilai Institute of Technology, Durg, India.*

[2]*Department of Computer Science and Engineering, Bhilai Institute of Technology, Durg, India.*

[3]*Department of Computer Science and Engineering, Bhilai Institute of Technology, Durg, India.*

Abstract

There is an increasing demand for machine translation systems for various regional languages of India. Chhattisgarhi being the language of the young Chhattisgarh state requires automatic languages translating system. This paper proposes rule based Chhattisgarhi to Hindi machine translation (MT) system that takes Chhattisgarhi as source language and Hindi as target language. It also discusses the issues to be considered for the translation. As there is not much structural difference between these two languages so formation of production rules, adding and changing of production rule is easier in Rule Based System since rule base exists for Hindi language.

Keywords: Machine Translation, Chhattisgarhi, Rule Based System

INTRODUCTION

India is a multi linguistic country in which 22 languages and 720 dialects are spoken by the people. For such multi linguistic and morphological rich country, language understandability is a big problem. Such problem can be solved by machine translation (MT) system. They are automatic system that takes a source language and converts it into target language [6]. Some work has already done for some regional Indian languages [3] [4]. These regional Indian languages can be broadly categorized into high and low resource languages. High resource languages are those languages whose grammar rule and other literary work is available in public domain like Marathi, Tamil, and Malayalam etc. There are some regional Indian languages which are called low resource languages like Bhojpuri, Magahi, and Nimadi etc., as the grammar rule and other literary work is not available in public domain.

For making machine translation system for regional languages, there are various machine translation approaches for automatic conversion of source language to target language. Some of which are:

Direct Machine Translation

Direct MT technique was developed during 1950s to make use of newly invented computers for MT. A direct translation system carries out word-by-word translation with the help of bilingual dictionary.

Hindi to Punjabi machine translation system based on direct approach has been proposed by [7]. The system architecture consists of pre-processing module, Hindi-Punjabi dictionary, morphological analysis module, transliteration and post processing modules.

Rule Based Machine Translation (RBMT)

RBMT system works on two components: lexicon and rules. The rule-based MT is used to remove major shortcomings of direct machine translation system. It parses the source text and produces an intermediate representation, which may be a parse tree or some abstract representation. The target language text is generated from the intermediate representation.

Punjabi to English machine translation system based on rule based approach has been proposed by [1]. The system architecture consists of three main components namely: Analysis, Translation and Synthesis component

Statistical Machine Translation

Statistical machine translation (SMT) system is based on bilingual corpora which consist of both source and target language .There are three phases in SMT: language modeling, translation modeling and decoding. In the first phase the probability of target language is determined denoted by P(T).In the second phase the conditional probability of target language is determined given the source language(T|S) and in the last phase the product of language model and translation mode is computed which gives most appropriate target sentence i.e. $P(S, T) = P(T)(S|T)$.

English to Malayalam machine translation system based on statistical machine translation approach has been proposed by [5]. The system architecture consists of suffix separator that uses to separate the suffix from Malayalam words in the sentence from the Malayalam corpus. With the help of decoder the English sentences gets converted to Malayalam.
For Chhattisgarh state, Chhattisgarhi is the state language. It is a low resource language. Government of Chhattisgarh is promoting Chhattisgarhi language in the administrative functioning of government. But, many citizens of Chhattisgarh state and government officers who are non

Chhattisgarhi speaking are facing problem in Hindi to Chhattisgarhi and Chhattisgarhi to Hindi conversion. The main objective of this paper is to address various issues related to Machine Translation. Since Chhattisgarhi is a low resource language due to which literary work of this language is not much available. Another challenge with the Chhattisgarhi Hindi machine translation system is the formation of Chhattisgarhi corpus and bilingual dictionary so that machine translation tools required for conversion can be made. Chhattisgarhi Hindi dictionary consisting of 56,819 bi lingual pair and a grammar for Chhattisgarhi language has been made by [2][8] .

ISSUES IN CONVERSION

The two important issues with the conversion of Chhattisgarhi to Hindi is the (i) Making Chhattisgarhi to Hindi Dictionary (ii) Formulation of production Rule.

For complete conversion of Chhattisgarhi to Hindi Chhattisgarhi Hindi bilingual pair from the dictionary [2], was take which were in Kruti Dev Hindi font and conversion is done into Unicode because it is a standard character set encoding technique that can support various types of character. Unicode uses different types of bit encoding like 8 bit and 16 bit. This encoding technique has been developed so that a single charter set can support all character from all scripts as well as some common symbols.

Chhattisgarhi to Hindi online dictionary developed is shown in Figure 1 and the database for the same is shown in Figure.2

Figure1: Chhattisgarhi-Hindi Dictionary

Figure 2: Chhattisgarhi Hindi database in Unicode

The following are some of the sub issues related to Chhattisgarhi to Hindi machine translation:

Lexical differences: Sometimes, a word used in one language has no single-word equivalent in another language which results into lexical differences between languages.

Example 1: The word अँइठ in Chhattisgarhi has two different meaning in Hindi.

अँइठ → 1. ऐंठने की क्रिया या भाव 2. अकड़

Gender resolution: In Hindi there are two types of gender masculine and feminine, but in Chhattisgarhi, it is difficult to identify the gender in interrogative sentences.

Example 2: In Chhattisgarhi, in interrogative sentences, the verb is suffixed by **थस**, and is difficult to interpret the gender. In Hindi sentences, gender can be easily identified from the verb. **रही हो** is used for feminine and **रहे हो** is used for masculine.

In Chhattisgarhi if it is **ते हा जा थस का**? , then for Hindi it can be 1.**क्या तुम जा रही हो**? or 2.**क्या तुम जा रहे हो**?

Increase in number of words in target language:

During translation from Chhattisgarhi to Hindi there are some cases of increase in the number of words in the target language.

Example 3:

Chhattisgarhi: मैदान म पाहट खड़े हे ।

Hindi: मैदान में भैसो का समूह खड़ा हैं ।

Decrease in number of words in target language: During translation from Chhattisgarhi to Hindi there are some cases of decrease in the number of words in the target language.

Example 4:

Chhattisgarhi: मे ह एक ठन आमा खाये हों ।

Hindi: मैं एक आम खाया हूँ ।

Conversion of idioms:

During translation from Chhattisgarhi to Hindi there are some cases where the system encounters Chhattisgarhi idioms; the conversion of theses idioms into equivalent Hindi idioms is a big challenge.

APPROACH FOLLOWED

Above all issues are considers during the design of the machine translation system for the Chhattisgarhi to Hindi.

The paper proposes that following approach can be adapted for conversion from Chhattisgarhi to Hindi:

Pre Processing

In the pre processing stage the compound noun phrases are converted in simple noun phrases. There are some noun phrases in Chhattisgarhi which are mixture of two words for which single word will be searched in Hindi.

Example: In Chhattisgarhi the word टुरा मन is consist of two word टुरा + मन for which single equivalent word लड़के exist in Hindi database.

Identification of Named Entities

In this stage named entities are identified by the help of their previous word like श्री and श्रीमती etc. The words that succeed theses words will be name like श्री विकास पांडेय, here विकास पांडेय will be transliterated.

Tokenization

In tokenization stage the whole text can be divided into sentences with the help of line splitter program where splitting will be done on encountering a delimiter, for Chhattisgarhi sentences पूर्णविराम [|] will act as delimiter.

Tagging and Morph Analysis

In the tagging phase all the untagged words can be tagged by the Sanchay tool. Sanchay tool is an open source platform made by Language Technologies Research Centre (LTRC) of IIIT Hyderabad, for working on Indian languages, using computers and also for developing Natural Language Processing (NLP) based applications. It is used in syntactic annotation interface (used for Hindi dependency annotation), it has several other useful functionalities as well. Font conversion, language and encoding detection, n-gram generation are a few of them [9]. In morph analysis the grammar category of words that gender, number, person, case will be stored in morph database. The field which is not applicable will be left empty.

Parsing

In parsing process the system deals with grammatical structure of a sentence and the relationship of the words with each other. The main objective of this analysis is to visualize syntactic structure of a sentence which is usually viewed in form of a parse tree. The syntactic structure is useful to understand the meaning of a sentence [10]. A Chhattisgarhi rule base has been designed through which the syntactic structure of the Chhattisgarhi sentences can be viewed in form of parse tree.

ARCHITECTURE OF CHHATTISGARHI HINDI MACHINE TRANSLATION SYSTEM

The complete architecture of Chhattisgarhi Hindi Machine translation system is shown in Figure 3.

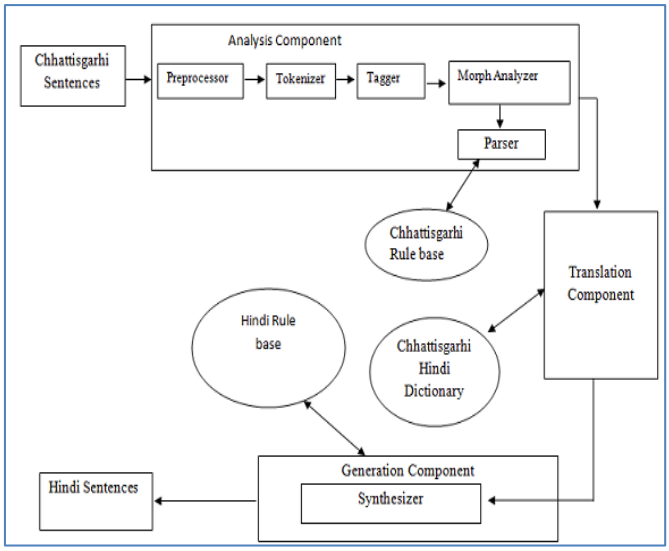


Figure 3: Complete Architecture of Proposed Chhattisgarhi to Hindi Machine Translation System.

The proposed architecture consists of following components:

(i) Analysis component-This component is divided into following components:

a) Preprocessor: It uses to split the sentence into tokensby the help of delimiter.

b) Tokenizer: It use to break the sentence in form of tokens.

c) Tagger: It uses to assign a particular part of speech tag to every word which is in form of tokens.

d) Morph Analyzer: It use to give morph information that is information related to person, Number and Gender from the morph database.

e) Parser: With the help of production rule it use to make the parse tree.

(ii)Translation component: It takes input from analysis component and helps in translation process by help of Chhattisgarhi Hindi dictionary.

(iii) Synthesis Component: It use to take the parse tree of the source language and convert it into parse tree structure of the target language by the help of transfer link rule file,

which is a file consisting of mapping information between source and target words .

The complete conversion process of the system can be well understood by the following steps:

1st step: Getting basic part-of-speech information of each source word:

वो = सर्वनाम; हा = विभक्ति; घर = संज्ञा; जाथे = क्रिया

2nd step: Getting syntactic information about the verb "जाथे ":

Here: जाथे – Present Simple, 3rd Person, Singular, Active Voice

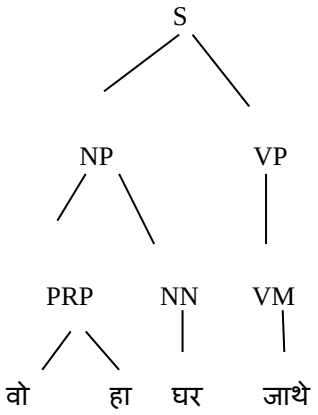
3rd step: Parsing the source sentence:

By the production rule from the rule base the shallow parsing will be done

S->NP VP

NP->PRP NN

VP->VM



4th step: translate Chhattisgarhi words into Hindi

वो (category = सर्वनाम) => वह (category = सर्वना)

हा (category = विभक्ति)

घर (category = संज्ञा) => घर (category = संज्ञा)

जाथे (category = क्रिया) => जाता (category = क्रिया)

है (category = स क्रिया)

5th step: Mapping dictionary entries into appropriate forms the help of transfer link rule file

(सर्वनाम) (विभक्ति) (संज्ञा) (क्रिया) =>

1 2 3

[Source Rule]

(सर्वनाम)(संज्ञा) (क्रिया) (स. क्रिया)

1 2 3

[Target Rule]

Transfer link rule mapping => 1:1 2:2 3:3

वो हा घर जाथे। => वह घर जाता है।

Since there is not much structural difference between Chhattisgarhi and Hindi as both derive from Devnagari script.

CONCLUSION AND FUTURE WORK

In this paper, we have discussed different issues considered during the design of machine translation system from Chhattisgarhi to Hindi. It also discusses different phases of rule based machine translation system. Conversion of Chhattisgarhi to Hindi sentences has been done using Chhattisgarhi to Hindi bilingual dictionary and production rules. Neural based Machine translation system is the most promising approach which can be done on the availability of parallel corpus. Hindi to Chhattisgarhi MT system is going to designed for which the dictionary is almost prepared.

REFERENCES

[1] Batra. K.K. and Lehal.G.S. 2010. *Rule based machine translation of noun phrases from Punjabi to English*. International Journal of Computer Science Issue.7, Vol. 5, pp. 409-412.

[2] Chandrakar.K. 2010. *Manak Chhattisgarhi vyakaran*. Stakshi Publication. ISBN No.:8189545086.

[3] Kalyani .A and Sajja P.S. 2015. *A Review of Machine Translation Systems in India and different Translation Evaluation Methodologies.* International Journal of Computer Applications, Vol. 23, pp. 0975 – 8887.

[4] Antony.P.J. 2013. *Machine translation approaches and survey for Indian languages.* Computational linguistics and Chinese language processing.18 (1). pp.47-48.

[5] Sebastian. M. P, Kurian. S and Kumar. S. G. 2010. *Statistical Machine Translation from English to Malayalam.* National Conference on Advanced Computing, pp.1-6.

[6] Kumar. E. 2013.*Natural Language Processing*. I.K. International Publishing House. ISBN No.:9789380578774.

[7] Goyal .V and Lehal G.S. 2011. *Hindi to Punjabi Machine Translation System*. International Conference for Information Systems for Indian Languages, Patiala, pp. 236-241.

[8] Chandrakar.K. 2012. Vrihad Chhattisghari shabda kosh. Chhattisgarh Hindi Granth Academy. ISBN No.:9788192169125.

[9] Agrawal, R., Ambati, B., & Singh, A.Singh.(2012). *A GUI to Detect and Correct Errors in Hindi Dependency Treebank.* In Proc.of Eighth International Conference on Language Resources and Evaluation, Istanbul, Turkey, 1907-1911.

[10] Tayal. M, Raghuwanshi. M & Malik. L. 2014 *Syntax Parsing: Implementation Using Grammar-Rules for English Language*. International Conference on Electronic Systems, Signal Processing and Computing Technologies, pp. 376-381, DOI: 10.1109/ICESC.2014.71.