International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 06 | June 2020 www.irjet.net p-ISSN: 2395-0072

Translation of English to Ahirani Language

Uday Chandrakant Patkar[1], Dr. Suhas Haribhau Patil [2], Dr. Prasadu Peddi [3]

[1]Research Scholar, JJTU
[2]Professor, Bharati Vidyapeeth Deemed University College of Engineering, Katraj, and Pune
[3]Professor, Shri Jagdish Prasad Jhabarmal Tibrewala University

---

Abstract - *This article describes how to implement system to translate the English to Ahirani language. Ahirani is a common language in the region Khandesh of Maharashtra of India. The system is developed using POS Tagging. The system uses total 20 documents from social media. This paper details the experiment by discussing the implementation. The experimental result shows that the performance of the system is more accurate.*

***Key Words*: Ahirani, POS Tagging, NLP (Natural language Processing).**

1. INTRODUCTION

The discourse about Indian language text input seems to be changing. Early work continued to focus on refining the conceptual text input model for users and creating simple enough text input mechanisms for people to type. The problem of conceptual confusion is particularly acute on mobile phones. In 2006, Katre said that it took between 18 and 55 taps on the main phone keypad to enter the one word, Maharashtra in Marathi, a word that has four glyph and only 10 Unicode characters. Back in 2011, Jung and others have described the text input in Indian languages as "puzzle". The slower typing speed on Indian keyboards is due to the large character set, the complexity of the script, the complex rules, the alpha syllabary script style, and the high scanning time. The World Wide Web (WWW), a rich source of information, is growing at a tremendous rate. Although English still remains the dominant language in the web, making this huge information repository available in English, making it accessible to non-English users worldwide is a recent important issue, as the overall internet usage statistics are steadily on an increasing number of non-English internet users.

Speech, which is used to communicate between people by means of a set of signs, whether graphical gestures, acoustics, or even musical, from the language of the ancient days, is one of the most important communication in everyday life. The Text-to-speech (TTS) system is capable of converting input text to speech waveforms. The TTS system works in two steps, i.e. text processing and voice generation. Text processing is performed to convert the input text given with the help of synthesis units to its normal form, and the language corresponding to each of these units is the ability to express their thoughts through a set of signs, whether it is graphical gestural, acoustic or musical.

It is the unique nature of man that is the only creature that uses such a structured system. Speech is one of its main components. It is by far the oldest means of communication among humans and also the most widely used. It is no wonder that people have studied it extensively and often tried to build a machine to handle it in an acoustic way. Most of the information in the digital world is accessible to anyone who can read or understand a particular language. Language technology can provide a solution in the form of a natural interface, so digital content can reach mass and facilitate the exchange of information across different people, different these technologies play an important role in an Indian-like society with dialects/native language of about 1652.As information systems are facing the challenges of information overhead management, remarkable progress in the last few years has been seen in information retrieval methods. This information overhead is due to the popularization of the communication network which causes the increase in the use of World Wide Web, www.

This is mainly due to the ever-increasing amount of shapeless data they make available to users. In English, a huge amount of text data is available on the Internet, and the same for Ahirani language. But, most of the tools available are and the methods are English-oriented. This shows that there is a lack of a tool that does effective text mining for both Ahirani and English. To overcome this limitation, in this article we have proposed a software interface that does automatic transliteration and translation of English text into Ahirani language.

Transliteration is becoming increasingly popular worldwide with the advent of Web 2.0 and mobile devices. In different social networks, people create sharing, tags, and search for multi-faceted data in multiple languages, but mostly use the Roman script. Even if we focus only on text data, a huge amount of text is generated on the Internet, much of which is in the transliterated domain. Although these texts are largely informal, they contain a great deal of information and therefore need to be studied. It has wide ramifications in low-resource languages in general, where Web presence is limited, specifically for Indian languages. India is a country

with a large population well versed with vernacular languages but not fluent in English. An Ahirani to English translation system will be helpful to the Ahirani speaking population who need to converse in English. Lot of documents, scripts and scriptures in Ahirani also need to be translated to English and this process is manual. Ahirani to English translation system will help to automate this process and help reduce manual work related to translation.

2. REVIEW OF LITERATURE SURVEY

Rajesh S Prasad and Kale Sunil Diagamberrao [1], the author propose a strategy for identifying the authorship of documents written in Marathi. Here they use a set of fine-grained lexical and stylistic features for text analysis and use them to develop two different models (Statistical similarity model and SMORDT-sequential minimal optimization using the rule-based decision tree method). Here they tested the feature extraction method to show consistent significance in each model used in these experiments. The performance of this method has been evaluated based on recall values, accuracy, F-measures, and accuracy. Pubali Chatterjee, Soumen Santra, Ananya Paul, Subhajit Bhowmick, Arpan Deyasi Partha Mukherjee [2], they have to developed a text-to-speech synthesizer that analyzes and processes text using NLP and transforms text into synthesized speech using digital signal processing (DSP) technology. Here they can convert the text entered into a synthetic speech, and then save it as an mp3 file, read out to the user and save it in the form of a simple application.

Mohammed Arshad Ansari and Sharvari Govilkar [3], the proposed system is an effort to automatically classify Hindi and Marathi text documents using supervised learning methods (KNN), Naive Bayes and support vector machines (SVMs). The use of machine learning algorithms as an application is often better than the performance of an analytical approach. Dinesh Kumar Prabhakar and Sukomal Pal [4], in this article, we will examine recent work in the field of transliteration. Following the various deterministic and non-deterministic approaches they used to tackle Translational-related issues in machine translation and Information Retrieval, at the different end, they study the performance of their technology and present comparative analysis of them. Vishal Kaushik, P Mohith, Hussain Rangoonwala, and Dhana Lakshmi Samiappan [5], in this paper, propose a complete system that can convert text into speech, convert text files into speech, convert text in various languages into speech, convert images into text, and convert images into speech using MATLAB as a programming tool. The various methods used are pre-processing, Unicode conversion, segmentation, concatenation, rhyming, and application for easy access and usability, and then to be combined sliding motive in the background to develop this system is to combine various modules by means of modular approaches.

Nutan B. Zungre, Nagmani Wanjaria and G. M. Dhopavkarb [6], paper proposes a rule-based system for correctly identifying sentence boundaries in Marathi. The task of identifying the end of a sentence in Marathi is to use certain rules to correctly determine the end of a sentence, as in Marathi the system dictates the beginning of a sentence so that English has a capital letter to indicate the beginning of a new sentence. D. B. K. Kamesh, Radhika Rani, S. Venkateswarlu and J. K. R. Sastry [7], Instead of reading text images, this paper presents an innovative and efficient real-time cost-effective technology that can listen to the contents of text images. It combines the concepts of Optical Character Recognition (OCR) and text to speech synthesis (TTS) in the Raspberry Pi. An effective vocal interface by a computer to interact with the visually impaired person of such a system. Kara text to speech is a way to scan and read the English alphabet and numbers that are in the image using OCR technology and change it to voice.

Ancy Anto and Nisha K. K [8], this paper is to help people translate English text into their own language and implement a minority language, Malayalam TTS. It is achieved by combining both machine translation and TTS. When an English text is given, it is translated into Malayalam with the help of a parser, using the grammatical rules, applying morphology and the bilingual dictionary. From each of the translated Malayalam text, the syllables are separated. Serkan Bilecen [9],in this study, an already available English text-to-speech application(which is used to read text by voice in a computer environment)was developed by using the means provided by the French Java programming language and the Voce API (Application Programming Interface), in which a string to Speech application is used to input a string to speech object. special pronunciation incidents such as vowels and liaisons are also evaluated and applied to the outcome. In the end, the words defined and pronounced in the 63out ot72 application were pronounced correctly and the overall success rate of 87.5% was achieved. Shashank Ahire, Sanjay Ghosh, Girish Dalvi and Nagraj Emmadi [10], author present the empirical results of these keyboards and discuss them with respect to their designs. They found that the keypad with logical layout work a bit better than the keyboard with the partially frequency layouts. The results also showed that users did not work well on keyboards that have word prediction features compared to keyboards that do not have prediction features when entering Marathi. Here, they assume that this performance difference is due to the “cognitive fee” that users pay for using word prediction. They also identify several areas for future research.

Sushma R, Niranjana Krupa B and: Dhananjaya M S [11], the authors propose an algorithm to translate Kannada text language into speech. Here, a direct concatenation of the speech coefficients extracted from the pre-recorded voice is used for the conversion. The proposed algorithm is also compared with another speech synthesizer, which is widely

used to evaluate its performance. Soudamini Pawar Avanti Patange, Madhuri Potey [12], In this article, they propose methods for converting documents written in ancient Marathi and English into modern Maratha and English using a cross-temporal information search algorithm. Researchers are developing an interest in temporary information retrieval as the amount of data on the Internet grows exponentially in this 21st century, making it difficult for the user to obtain the relevant documents. IR studies often ignore lexical drift. But in the growing area of huge digitized book collections, the risk of dictionary mismatch due to language change is high. Some collections contain text written in the folk languages of older centuries. The time measurement available in documents must be integrated with document ranking for efficient retrieval.

Dr. Sanjay Mathur, Shilpi Kannojia and Ghana Priya Singh [13], this article presents the improvement of prosody in acoustic unit on the basis of concatenation of any language of Devanagari writing. In the system of text and speech synthesis syllabic synthesis is quite clear in comparison with systems based on the Di-phone. Designed so the system takes a written text in any language of the Devanagari script using the utilities of MS word using MATLAB, which is then converted into the Romanized script in the analysis of the text.

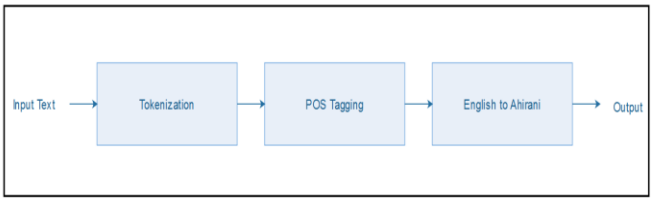
Akshay Bansode, Adita Kulkarni, G.V. Garje and Suyog Gandhi [14], to develop a software system that would translate simple affirmative and interrogative sentences of Marathi into appropriate English sentences. The translation quality of the existing system is very rough. Since there are no full-featured translation systems from Marathi to English, they intend to develop one such system to provide a higher quality translation. Rithika. H and B. Nithya santhoshi [15],this article is based on a prototype that helps the user to hear the contents of text images in the desired language. It involves extracting text from an image and converting the text into translated speech in the user's desired language. This is done using the Raspberry Pi and camera module, using the Tesseract OCR concept [optical character recognition] engine, Google Speech API [application interface], which is a text-to-speech engine and Microsoft translator. This makes it easier for travelers as they can use this device to hear English text in their own desired language. It can also be used by the visually impaired.

3. METHODOLOGY

The main aim to develop a system to translate the English to Ahirani Language using POS Tagging. The system is designed by using Java 1.8 framework on Windows platform. The Net bean IDE is used as a development tool.

The figure 1 shows the proposed system architecture. Translation is the communication of meaning from one language (the source) to another language (the target). Translation refers to written information, whereas interpretation refers to spoken information.

Figure 1: System Architecture



Input: The system takes English language as an input. Then perform tokenization.

Tokenization: Tokenization is a method of dividing a string into several parts. String Tokenizer is a utility class for extracting tokens from strings. However, the Java API documentation is at the mercy of its use, not the recommended method of splitting, which is similar to the String class. The Split method uses regular expressions. The table 1 shows the example of tokenization.

Table 1. Example of Tokenization

Token	Informal Description
abed	characters a, b, e, d
ability	characters a, b, i, l, i, t, y

POS Tagging: Part of speech tagger (POS tagger) is a software application that reads text in several languages and assigns part of speech to each word (and other tokens) such as noun, verb, adjective, etc. POS tagging is also essential for building lemmatizes that are used to reduce words to the root form. POS tagging is the process of marking up words in a corpus into corresponding parts of speech tags based on their context and definition. The table 2 shows the example of POS Tagging.

Table 2. Example of POS Tagging

Word	Tag
abed	Noun
ability	Noun

Output: The system translates the English language into Ahirani language. Translation involves a change of language altogether. The table 3 shows the English to Ahirani translated words.

Table 3. English to Ahirani translated words

English	Ahirani(Converted to English)
abed	Khat
ability	Shamta

4. CONCLUSION

The statistics show that the Ahirani language is one of the most important languages in the Maharashtra region. The need for an create an automatic, efficient and effective translation of English to Ahirani languages. In this paper system proposed a technique to translate the English to Ahirani language. The system use tokenization for dividing a string into several parts. Then by using POS tagging reads text in several languages and assigns part of speech to each word. After identifying for each word, its exact transliteration and a proper translation in English to Ahirani language is done.

REFERENCES

[1] Kale Sunil Diagamberrao, Rajesh S Prasad, "Author Identification using Sequential Minimal Optimization with Rule Based Decision Tree on Indian Literature in Marathi", International Conference on Computational Intelligence and Data Science (ICCIDS 2018), 2018.

[2] Partha Mukherjee, SoumenSantra, Subhajit Bhowmick, Ananya Paul, and ArpanDeyasi, "Development of GUI for Text-to-Speech Recognition using Natural Language Processing", TENCON IEEE -2018.

[3] Mohammed Arshad Ansari and Sharvari Govilkar, "Sentiment Analysis of Mixed Code for the Translated HINDI AND MARATHI Texts", International Journal on Natural Language Computing (IJNLC) Vol. 7, No.2, April 2018.

[4] Dinesh Kumar Prabhakar and Sukomal Pal, "Machine transliteration and transliterated text retrieval: a survey"; Indian Academy of Sciences, Springer-2018.

[5] Hussain Rangoonwala, Vishal Kaushik, P Mohith and Dhana Lakshmi Samiappan; "Text to Speech Conversion Module; International Journal of Pure and Applied Mathematics, Volume 115 No. 6, 2017.

[6] Nagmani Wanjaria, Prof. G. M. Dhopavkarb, Nutan B. Zungre, "Sentence Boundary Detection for Marathi Language", International Conference on Information Security & Privacy (ICISP2015), 11-12, Nagpur, INDIA, Elesevir-2016.

[7] S. Venkateswarlu1, D. B. K. Kamesh, J. K. R. Sastry and Radhika Rani, "Text to speech conversion", Indian Journal of Science and Technology, Vol 9(38), -2016.

[8] Ancy Anto and Nisha K. K, "Text to Speech Synthesis System for English to Malayalam Translation"; International Conference on Emerging Technological Trends [ICETT]-2016.

[9] Serkan Bilecen, ": Interpretation of English Text to Speech Application French Language", IEEE-2016

[10] Girish Dalvi, Shashank hire, Nagraj Emmadi and Sanjay Ghosh, "Dose Prediction Really Help in Marathi Text Input? Empirical Analysis of a Language study"; 18th International Conference on Human-Computer Interaction with Mobile Devices and Services, ACM-2016.

[11] Dhananjaya M S, Sushma R and Niranjana Krupa B; "Kannada Text to Speech Conversion: A Novel Approach"; International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT), IEEE-2016.

[12] Avanti Patange, Madhuri Potey, Soudamini Pawar, "Cross Temporal Information Retrieval for Marathi and English Language"; International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 5, Issue 5, May 2016.

[13] Shilpi Kannojia, Ghana Priya Singh, and Dr. Sanjay Mathur, "A Text to Speech Synthesizer Using Acoustic Unit Based Concatenation for Any Indian Language of Devanagari Scripts", 11th International Conference on Industrial and Information System (ICIIS)-2016.

[14] G.V. Garje, Akshay Bansode, Suyog Gandhi and Adita Kulkarni, "Marathi to English Sentence Translator for Simple Assertive and Interrogative Sentences"; International Journal of Computer Applications (0975 – 8887) Volume 138 – No.5, March 2016.

[15] Rithika, B. Nithyasanthoshi, "Image Text to Speech Conversion in the Desired Language by Translating with Raspberry Pi", IEEE-2016.