English to Marathi Translator with retaining structure of data

Miss. Swati M.Raut[*] **, Mr. Zakir M.Shaikh**[**]

[*] Department of Computer Science & Engineering,N K Orchid College of Engineering Solapur, India
[**] Department of Computer Science & Engineering,N K Orchid College of Engineering Solapur, India

Abstract- People of different linguistic background could not able to interact with each other. This concept of translation will help people to communicate comfortably. Also it will help to fill communication gap between two linguistically different back-grounds. It will help to the people in the villages, who have taken education of English. Majority of the Indian population is not familiar with English while most of the information available on web or electronic information is in English. So, to reach out to the common man across various sections and field, an automatic language translator is important. So the main objective of Machine Translation (MT) is to break the language barrier in a multilingual nation like India.MT gives several approaches to translate source language to target language. In this paper we are translating English to Marathi with retaining structure of data.

Index Terms- Approaches. Communication, Language, Lexicon, Machine Translation(MT) .

I. INTRODUCTION

Hindi is a widely spoken language and it is the principal official language of the Republic of India. On the other hand, English is internationally popular language. In India, English as a language has played a major role in administration, legal and education sector since British period. Presently, an awareness has been developed in this country for using regional languages such as Marathi , Kannada etc. for government document writing, for primary and higher education and every other domain of public life. In this context, it has become very important to build system which can translate English to various Indian languages. With the existence of huge text resources in internet and India being one of the most prominent users of web, even commercial companies are finding it necessary to venture out for building machine translation.

There are three types of machine translation framework :

Rule-Based Machine Translation (RBMT) systems use large collections of rules which is manually developed over time by human experts mapping structures from the source language to the target language. The first RBMT systems were developed in the early 1970s. The human factor in rule-based systems helps deliver fairly good automated translations with predictable results. Rule-based systems can be quite costly and time consuming to implement as well as to maintain. As rules are added and updated these systems have the potential of generating ambiguity and translation degradation over time[1].

Statistical Machine Translation systems use computer algorithms to produce a translation that looks best statistically from millions of permutations. The first ideas of statistical machine translation were introduced by Warren Weaver in 1949, including the ideas of applying Claude Shannon's information theory. Statistical machine translation was re-introduced in 1993 by researchers at IBM's Thomas J. Watson Research Center and has contributed to the significant resurgence in interest in machine translation in recent years. Statistical models consist of words and phrases learned automatically from bilingual parallel sentences, creating a bilingual "database" of translations. The attractiveness of statistical systems comes from the level of automation in building new systems using its machine learning capabilities, leading to rapid turnaround time and the low cost of processing power required for constructing and operating these statistical models.

Hybrid Machine Translation:

Hybrid-based approach is developed by taking the advantage of both statistical and rule-based translation methodologies which has proven to have better efficiency in the area of MT systems. At present, several governmental and private based MT sectors use this hybrid-based approach to develop translation from source to target language, which is based on both rules and statistics. The hybrid approach can be used in a number of different ways. In some cases, translations are performed in the first stage using a rule-based approach followed by adjusting or correcting the output using statistical information. In the other way, rules are used to pre-process the input data as well as post-process the statistical output of a statistical-based translation system. This technique is better than the previous two techniques and has more power, flexibility, and control in translation [1].

II. NEED OF TRANSLATION

The official language of India is Hindi. According to the Constitution, Parliamentary business may be conducted in either Hindi or English. In corporation world English language is used for official work. At state level different languages are used, for example Marathi language is used in Maharashtra , Kannada is used in Karnataka etc. So there is need to translate English to local language for communication/ circulation and interaction purpose. In office work there is need to translate documents into local language.

III. PROBLEM STATEMENT

This paper aim to translate English to Marathi word by word convertor with retaining structure of input file. For example if

input file is document file with English text so we are going to translate it in Marathi with retaining same structure of input file means if input file contains bold or italic font style words, after convertor it will display with same property.

IV. METHODOLOGY

As per our problem statement we are presenting English to Marathi convertor on basis of word. For translating words we created dictionary for retrieving the corresponding Marathi word. First open the document file which we want to translate from location. Then read file document file so here we identify the property of each word in file such as font style, font size etc. Extract the words from document file and search it in database. When it found write it in output file with all properties.
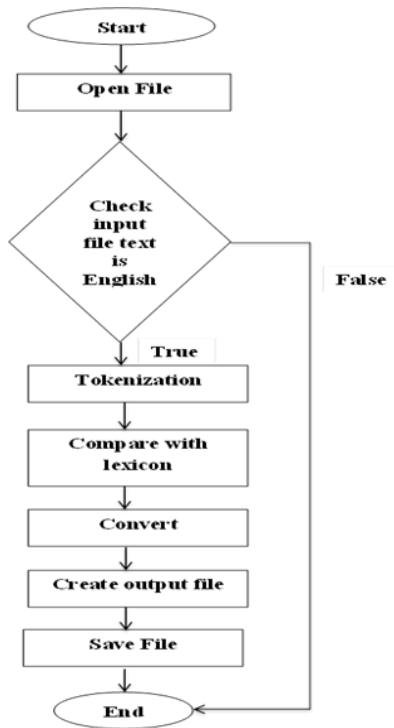

Fig.2 Home Page

On the next screenshot given in Fig.3 which shows the contents of file. When we clicked on open menu item it will open dialog box through which we open document file. As given in fig.3 it shows contents of file here the contents are,

This is *example*

Actual input file contains this text in which first word "This" is bold, "is" is plain text and "example" is bold and italic style.

Fig.3 Open document file

Fig.4 is convert file means this fig shows that after tokenization we translate words into Marathi

"है आहे *उदाहरण*"

So we are going to write this line in output file. Then this file with same extension of input file.

Fig.1 Flowchart of proposed system

The flowchart given above in fig.1 which shows all steps taken. As stated below first step is to open the document file. The file which is open should contain only English text.so first check is document contains English text. If the file contains English text, then tokenization is done. Here words are separated and then identify the properties of words. Some words can be bold or italic. Then compare word with lexicon. So for that we need large amounts of database. After searching appropriate words convert into Marathi language. Create same input file extension and write to output text and save file. So in this we get output file.

The screenshots given below Fig.2 shows that it is home page so it gives options like open, Translate, save file. For getting help there is menu provided Help and last one is Exit.
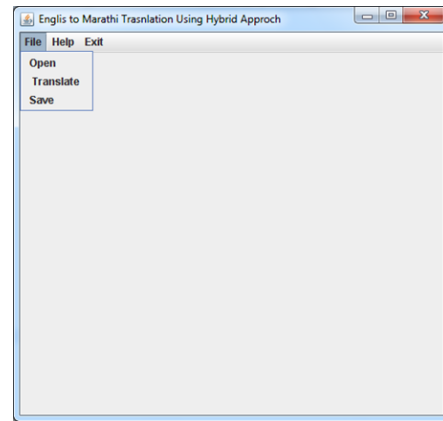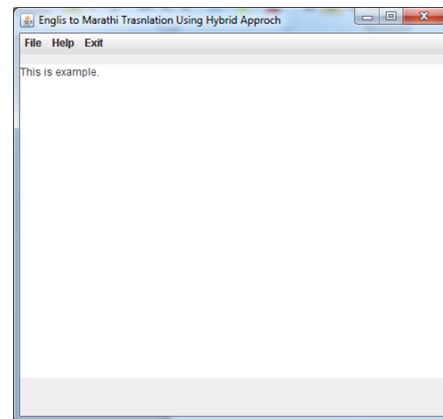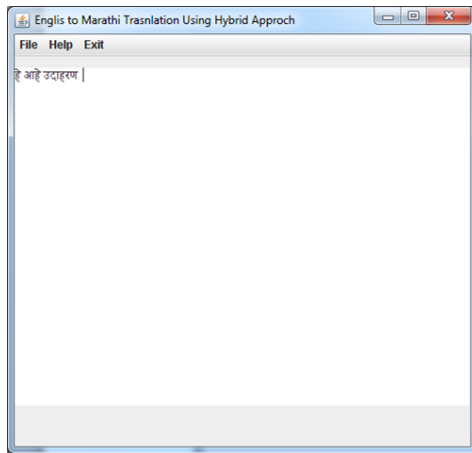
Fig.4 Translate document file

Now Fig.5 shows the actual output file after converting it into Marathi text. The output file contains same contents with translating into Marathi language along with their properties.
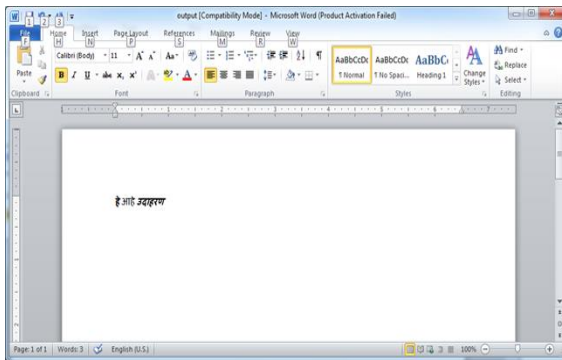

Fig.5 Output File

V. CONCLUSION

Language translators are very important in day to day life as we presented basic translator which is word by word translator. In an office there are many circulars and document which has to translate into local language. In that there is need to translate documents should be retained with structure.

ACKNOWLEDGMENT

This acknowledgement is humble attempt to honestly thank all those who are directly or indirectly involved in my project and are of immense help to me. I would sincerely like to thank you our guide Prof. Z. M. Shaikh for giving me valuable guidance also thanks to principal and all staff of NK Orchid College of Engineering.

REFERENCES

[1] Antony P. J.∗,"Machine Translation Approaches and Survey for Indian Languages", Computational Linguistics and Chinese Language Processing Vol. 18, No. 1, March 2013, pp. 47-78.

[2] Zhang, Y. (2006). Chinese-English SMT by Parsing. www.cl.cam.ac.uk/~yz360/ mscthesis.pdf.

[3] Oepen, S., Velldal, E., Lønning, J. T., Meurer, P., Rosen, V., & Flickinger, D. (2007) Towards Hybrid Quality-oriented Machine Translation on Linguistics and Probabilities in MT. In Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation, 144-153.

[4]   Dave, S.,Parikh, J., & Bhattacharya, P. (2001) Interlingua-based English-Hindi Machine Translation and Language Divergence. Journal of Machine Translation, 16(4), 251-304.

[5] Rekha Sugandhi, Devika Pisharoty, Priya Sidhaye, Hrishikesh Utpat, Sayali Wandkar and, Rajendra Khope ",Extending Capabilites of English to Marathi Machine" International Journal of Engineering, Science and Research (IJESR), Volume 1, Issue 3, 2012

[6] Alawneh, M.F.; Sembok, T.M., "Rule-Based and Example-Based Machine Translation from English to Arabic," Bio-Inspired Computing: Theories and Applications (BIC-TA), 2011 Sixth International Conference on , vol., no., pp.343,347, 27-29 Sept. 2011

[7] Sugandhi, R.S.; Shekhar, R.; Agarwal, T.; Bedi, R.K.; Wadhai, V.M., "Issues in parsing for Machine Aided Translation from English to Hindi," Information and Communication Technologies (WICT), 2011 World Congress on , vol., no., pp.754,759, 11-14 Dec. 2011

AUTHORS

First Author – Miss. Swati Raut recived B.E. Infromation Technology in 2013 from Pune University,India and pursuing the M.E.Computer Science and Engineering from Solapur University Solapur. She is doing her dissertation work under the guidance of Mr. Zakir Mujeeb Shaikh. Email-raut.swati142@gmail.com

Second Author – Mr. Zakir Mujeeb Shaikh recived B.E.Computer Science and Engineering in 2000 from Dr. Baba Saheb Ambedkar Marathwada University Aurangabad,Maharashtra, India and M.E.Computer from BVU Pune in 2004. He is presently working as assistant professor in NK Orchid college of engineering and technology Solapur , Maharashtra, India. His research area includes Image Processing, Algorithm Design, Software Engineering. Email-zakirshaikh@nkorchidenggmgmt.ac.in