DOI: http://dx.doi.org/10.26483/ijarcs.v9i1.5266

Volume 9, No. 1, January-February 2018

ISSN No. 0976-5697

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

A STUDY OF MACHINE TRANSLATION APPROACHES FOR GUJARATI LANGUAGE

Jatin C. Modh
Assistant Professor,
Narmada College of Computer Application,
Bharuch, Gujarat, India.

Dr. Jatinderkumar R. Saini
Professor & I/C Director,
Narmada College of Computer Application,
Bharuch, Gujarat, India.

Abstract: India is a multi-lingual country. At present, there are 22 official languages in India. Gujarat is a state located in the western region of India. The Gujarati language is spoken by nearly 60 million people worldwide, making it the 26th most-spoken native language in the world. In Machine Translation System (MTS), one natural language gets translated to another language using computational applications with minimal human effort or without a real-time human interface. Many attempts have been done in Machine Translation System for Indian languages. Unfortunately, we do not have an efficient Machine Translation System today. This paper gives a brief description of approaches of Machine Translation and the work done for the Gujarati language.

Keywords: Machine Translation System (MTS); Computational Linguistics; English; Gujarati; Natural Language Processing

I. INTRODUCTION

Machine Translation [1] refers to the automated translation of text from one language to another language. Machine Translation System (MTS) is the application of Natural Language Processing (NLP) of Artificial Intelligence. The language of text entered as an input is known as the source language whereas the language of output text is known as the target language. Nowadays Machine Translation System is an emerging area of study for researchers in India. India is multilingual country. Indian government uses Hindi or English language as a communication medium whereas various states of India use their local language as a communication medium. There is a big demand for document conversion from one language to another language. The English language is widely used in all fields. So Machine Translation Systems are needed for translation of local language to English language or vice-a-versa.

The Gujarat language is the official language of the state of Gujarat of India. Indian government publishes and issues official documents in English or Hindi or in both the languages. State government publishes official documents in their regional languages also. Gujarat Government uses the Gujarati language for official documents. In the Gujarat state, local newspapers, magazines and books are published in the local Gujarati language only. For the exchange of information among states, central government, industry, academia, good Machine Translation System (MTS) is required. Manual translation of documents is very time consuming and costly. This paper presents the approaches of Machine Translation and the work done for Machine Translation for Gujarati-English or English-Gujarati language pairs.

II. OVERVIEW OF MACHINE TRANSLATION APPROACHES

Researchers proposed many approaches for the Machine Translation. Overview of main approaches is presented here. There are two broad categories of Machine Translation Systems, namely Rule-Based and Empirical Based Machine Translation Systems. Hybrid Machine Translation system takes the benefits from both Rule-Based Machine Translation System and Empirical Based Machine Translation System. Rule-Based Machine Translation System is further classified into Direct, Transfer and Interlingua, while Empirical Based Translation System is classified into Statistical and Example-based machine translation system.

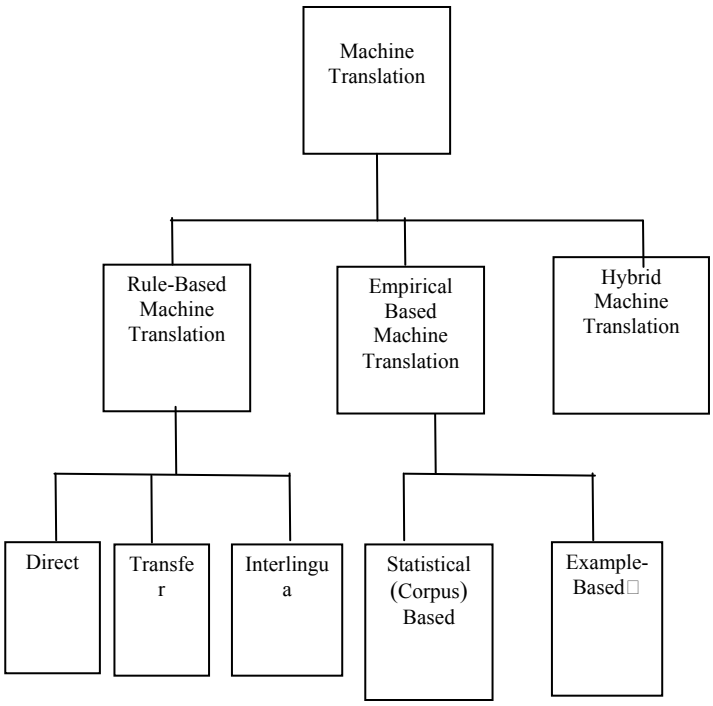


Figure 1. Classification of Machine Translation System

A. *Rule-Based Machine Translation (RBMT)*

Rule-Based Machine Translation is a traditional method of Machine Translation and also known as Knowledge-Based Machine Translation [12]. RBMT uses grammar rules which

are nothing but the collection of rules. It also uses a program and bilingual dictionary for the processing of rules. Rule-Based Machine Translation requires immense human effort in coding. Much attention is required in implementing resources like the bilingual lexicon, syntactic parser, target language morphological generator, source part of speech taggers, and source to target transliteration. Rules play important role in the morphological, syntactic and semantic analysis of each language [12]. Direct i.e. Dictionary-based, Transfer-based and Interlingua Machine Translation are the three different approaches that come under the RBMT category.

1) Direct Machine Translation

Direct Machine Translation approach is less popular, made at the word level. Words are converted into the target language without passing through an intermediary representation [12]. For each source and target language pair, there is a need of separate translator in Direct Machine Translation [11]. It is basically bilingual and uni-directional. It is a word-by-word translation approach with little syntactic, semantic analysis and grammatical adjustments.

2) Transfer Based Machine Translation

Transfer-based machine translation entails three stages, Analysis, Transfer and Generation. In the Analysis stage, the source language is parsed, the sentence structure and components of the sentence are identified. Transfer stage applies transformations to the source language parse tree for structure conversion in the target language. The generation stage translates the words and represents the gender, number, tense etc in the target language. Analysis phase constructs a source language dependent representation, so a separate transfer component is needed for each direction of translation for every pair of languages in case of multilingual machine translation system [13].

3) Interlingua Based Machine Translation

The Interlingua based approach is based on the intermediate representation of language. Translators are built that convert text from input language to the intermediate representation – a representation of its 'meaning' in some respect - which could form the basis for the generation of target language [12]. Interlingua based model also constructs a parse tree from the source language. It moves one step further and transforms the source language parse tree into a standard language-independent format, known as Interlingua [2]. The main advantage of this model is that it can be used with any language pair. It uses abstract elements like agent, event, tense etc. The generator component for each target language takes the Interlingua as input and generates the translation in the target language.

B. Empirical Based Machine Translation

Nowadays, Machine Translation research is dominated by Empirical Based approaches. Empirical-based Machine Translation research is corpus-based or data-driven approaches. This is the alternate way to Rule-Based Translation System. A bilingual corpus is required as a base for the text transformation. It is further classified into two sub-approaches namely Statistical Machine Translation system and Example-Based Machine Translation system.

1) Statistical based or Corpus-based Machine Translation

Statistical-Based Machine Translation is a data-oriented approach for translating text from the source language to target language based on the statistical models extracted from bilingual or multilingual textual corpora. Statistical tables are built from the corpora with the help of supervised or unsupervised statistical machine learning algorithm called the learning or training [3]. These tables consist of statistical information like the correlation between languages and characteristics of well-formed sentences. Decoding process uses this statistical data to find the best translation of the entered input text. Statistical Machine Translation uses the available parallel corpora. It does not apply linguistics analysis on entered text. The output of Statistical translation model relies thoroughly on available multilingual corpora; however, this model can be built fast. Statistical Machine Translation systems have the capability to comprehend indirect knowledge incorporated in co-occurrence statistics [4]. Statistical based technique experiences difficulties to manage the situation which needs linguistic knowledge like morphology, syntactic functions and ordering of words.

2) Example based Machine Translation

The main purpose of Example-Based Machine Translation is a translation by similarity. In Example-Based Machine Translation, old examples are used to translate the entered text. The input text is translated on the base of similarity using the bilingual dictionary. It is no use to do the deep linguistic analysis but applies the point to point mapping of source to target sentence [12]. In Example-Based Machine Translation can be viewed as case-based reasoning in which example translations are used to train a system. [17]

C. Hybrid Machine Translation

Various projects use just Rule-Based Machine Translation approach whereas various projects use simply Empirical Based Machine Translation approach. Hybrid Machine Translation technique makes use of the benefits of both the techniques, Rule-Based Machine Translation and Empirical Based Machine Translation. The hybrid technique finds a use of both rules and corpora. Hybrid Machine Translation implements the high accuracy of linguistic analysis from Rule-Based methods as they can successfully manage the overall syntactic structure of sentences and word re-arrangement. Hybrid Machine Translation implements the Statistical-Based Machine Translation methods as they are constructed using the vast collection of previously converted text (parallel text corpora). The Rule-Based Machine Translation system has slower development cycle compared to Statistical-Based Machine Translation. Rule-Based Machine Translation systems need manually built lexicon, grammars and algorithms that limit the supported languages. Statistical-Based Machine Translation systems are accurate in ambiguity resolution but the problem in linguistics. Recently researchers follow Hybrid Machine Translation approach by adding some linguistic rules and sentence structure on Statistical-Based Machine Translation systems [5].

III. LITERATURE SURVEY FOR MACHINE TRANSLATION SYSTEM FOR THE GUJARATI LANGUAGE

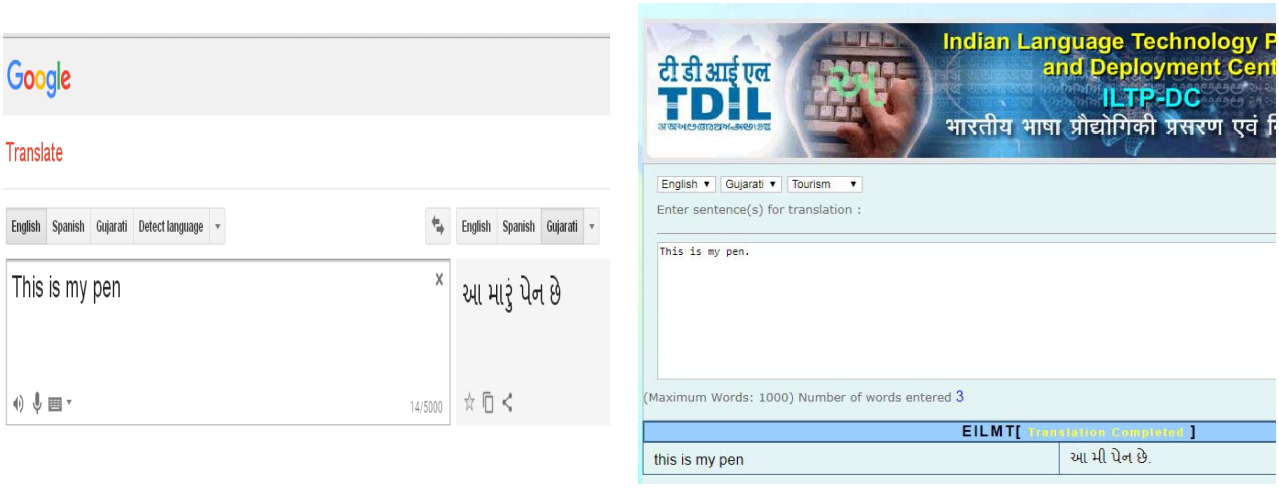
Machine Translation has been an active research field of Artificial Intelligence for years. Various government institutions and private companies constructed many Machine Translation systems for Indian languages. The extent of this paper is limited to two language pairs Gujarati-English and English-Gujarati only. The following Table 1 gives brief information about Machine Translation Systems involving only two language pairs Gujarati-English and English-Gujarati. Four Machine Translation Systems namely MANTRA, Google Translate, ANUVADAKSH/EILMT and AnglaBharati-II are found that uses Gujarati as a source or target language. Out of these four projects, only

ANUVADAKSH/EILMT and Google Translate are found active, whereas MANTRA and AnglaBharati-II are found inactive. ANUVADAKSH/EILMT converts English into Gujarati, whereas Google Translate supports both the language pairs English-Gujarati and Gujarati-English.

Table I. Machine Translation Systems for language pairs Gujarati-English or English-Gujarati

Sr. No	Machine Translation System	Machine Translation Approach	Language pair	Year, Developer	Data Quality and Features
1	MANTRA [15]	Transfer Based	English to Gujarati	1999, CDAC, Pune	Currently, the work for language pairs English-Gujarati, English-Hindi, English-Bengali and English-Telugu is going on. The work for language pairs Hindi-English, Hindi-Marathi, and Hindi-Bengali is going on.
2	Google Translate	Statistical Based	English-Gujarati, Gujarati-English [8]	2006, Franz Josef Och [16]	It is having server-based solutions with good enough accuracy. It is having Parallel corpora (at least, 5,000,000 words /500,000 translation units). It also supports document translation. Currently, Google Translate supports 103 languages. [6][7] Google will support the Gujarati language as a part of GNMT (Google Neural Machine Translation) engine which announced by Google in November 2016 [16]. GNMT engine will translate overall sentence at a time.
3	ANUVADAKSH/ EILMT	Hybrid Machine Translation [5]	English to Gujarati [14]	2009, Consortium-Based Project	ANUVADAKSH/EILMT incorporated 12299 sentences as training corpus and another more 1570 sentences were incorporated for testing and tuning. ANUVADAKSH/EILMT helps translation for 8 language pairs namely English to Gujarati, English to Urdu, English to Marathi, English to Bengali, English to Hindi, English to Oriya, English to Tamil and English to Bodo language.
4	AnglaBharati-I	Pattern-directed rule-based system (Pseudo interlingua) [10]	English to Indian languages (primarily Hindi)	1991, IIT, Kanpur	90% translation task is done by machine and 10% left to the human post-editing.
5	AnglaBharati-II	Hybrid Approach Pseudo interlingua [10][9]	English to Indian Languages including Gujarati	2004, IIT, Kanpur	The efficiency of the system was improved as compared to ANGLABHARTI-I.

The following figure 2 shows the output received from ANUVADAKSH and Google Translate from English to Gujarati language translation. Simple sentence "This is my pen" is given for translation. Google Translate does translation as "આ મારું પેન છે". ANUVADAKSH does translation as "આ મી પેન છે". The correct translation in the Gujarati language is "આ મારી પેન છે".



(a) Google Translate (English to Gujarati) (b) ANUVADAKSH/EILMT (English to Gujarati only)

Figure 2. Comparison of Google Translate with ANUVADAKSH (Gujarati to English)

IV. CONCLUSION

Machine Translation refers an automated process of translation from one language to another language. It is very hard to achieve human-level translation quality. Based on the study of various Machine Translation approaches, we found that each approach has the capability to translate from source language to target language with some limitations, but statistical based approach brings to maximum accuracy. In the era of the Internet, web-based interactive translation is getting popularity.

It is concluded that nowadays the lot of research work is going in the area of Machine Translation, but we found only two Gujarati Machine Translation systems namely ANUVADAKSH and Google Translate into working mode and available online. ANUVADAKSH is translating from English to Gujarati only, whereas Google Translate is translating from English to Gujarati and Gujarati to the English language. Google translate clearly gives the better result than ANUVADAKSH for English to Gujarati translation, but gives an incorrect result for Gujarati idioms to English translation. Google Translate, web-based Machine Translation supports the Gujarati language which follows Statistical Machine Translation System.

V. REFERENCES

[1] J.Hutchins and H.Somers, "An introduction to Machine Translation", Academic Press, 1992

[2] Shachi Dave, Jignashu Parikh, and Pushpak Bhattacharyya, "Interlingua based English Hindi machine translation and language divergence", Journal of Machine Translation, 2002

[3] Zhang, Y., "Chinese-English SMT by Parsing", 2006, Available Online: www.cl.cam.ac.uk/~yz360/mscthesis.pdf

[4] Sneha Tripathi & Juran Krishna Sarkhel, "Approaches to Machine Translation", International Journal of Annals of Library and Information Studies, Vol. 57, 2010, pp. 388-393

[5] Neeha Ashraf & Manzoor Ahmad, "Experimental Framework using Web-based Tools for Evaluation of Machine Translation Techniques", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 6, Issue 4, April 2016, Available online: www.ijarcsse.com

[6] Mamta & Tanuj Wala, 2015, "Survey of Approaches Used in Machine Translation System", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Volume 4 Issue 5, May 2015

[7] Och, Franz Josef, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague, Association for Computational Linguistics, pp. 858-867, June 2007

[8] Google Translate, Google Corporation Ltd.; Available Online: https://translate.google.co.in/

[9] Sudip Naskar, Sivaji Bandyopadhyay, "Use of machine translation in India: Current status", Available Online: https://www.researchgate.net/profile/Sudip_Naskar/publication/228527440_Use_of_machine_translation_in_India_Current_status/links/02e7e517eabc7ea3dc000000.pdf

[10] 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

[11] Deepak Khemani, "A First Course in Artificial Intelligence", 2013, pages 702-703

[12] M. D. Okpor, "Machine Translation Approaches: Issues and Challenges", IJCSI International Journal of Computer Science Issues, Vol. 11, Issue 5, No 2, September 2014

[13] Somya Gupta, "A survey of Data Driven Machine Translation", Department of Computer Science and Engineering Indian Institute of Technology, Bombay, Mumbai, 2012

[14] Anuvadaksh, "Anuvadaksh", Machine Assisted Translation Tool; Available Online: http://eilmt.rb-aai.in/

[15] Sitender, Seema Bawa, "Survey of Indian Machine Translation Systems", International Journal of Computer Science and Technology (IJCST), March 2012

[16] Wikipedia, Available Online: https://en.wikipedia.org/wiki/Google_Translate

[17] Wikipedia, Available Online: https://en.wikipedia.org/wiki/Example-based_machine_translation