IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Vol. 09, Issue 5 (May. 2019), ||S (XI) || PP 50-55

Language Translation using Predictive Dictionary and Corpus

Pinky Gangwani[1], Prof. Samir Ajani[2]
[1]*MTech Scholar, Computer Science Engineering, Jhulelal institute of technology, Nagpur, Maharashtra, India*
[2]*Asst. Professor, Computer Science Engineering, Jhulelal institute of technology, Nagpur, Maharashtra, India*

Abstract : *Language is an effective medium of communication. It basically represents the ideas and expressions of human mind. There are more than 5000 languages exist in the world which reflects the linguistic diversity. It is difficult for an individual to know and understand all the languages of the world. Hence, the methodology of language translation (LT) was adopted to communicate the messages from one language to another language. Several online translation tools are now available which support translation of text into one or more languages such as Bing Translator of Microsoft, Google Translate from Google etc. Language Translation (LT) is one of the most important applications and research tasks of NLP which investigates the use of software to translate text or speech from one natural language to another natural language using computers with or without human assistance. Language Translation (LT) systems have been developed for translation from English to Indian Languages and from regional languages to regional languages. The LT system which generates translation between two specific languages are called bilingual LT systems. English is a SVO (subject-verb-object) language while Indian regional languages are SOV (subject-object-verb) and are relatively of free word-order. In this paper, we are presenting a Dictionary (unclassified) that contains English word to their corresponding Sindhi language word, Dictionary (classified: Animals, birds, colors, days of week etc.), Phrases, Proverbs, Parts of speech and Translation of some sentences from English to Sindhi language.*

Keywords: *Language Translation (LT), Natural Language Processing (NLP), Dictionary, Phrases, Proverbs, Parts of speech, Translation.*

I. Introduction

Language Translation (LT) can be defined as an automated system that analyses text from a Source Language (SL), applies some computation on that input and produces equivalent text in a required target language (TL) ideally without any kind of human intervention. It is one of the most interesting and the hardest problem in the field of NLP (natural language processing).

We will develop a dictionary of vocabulary and grammar rules from English to Sindhi in MS.SQL Server. We will develop an app that will translate inputted English phrases into Sindhi by scanning through dictionary and programming language API's. As much as comprehensive is dictionary the most accuracy will be shown. In short, Accuracy is based on training dictionary. Higher Training higher accuracy.

Language Translation (LT) is one of the most important applications and research tasks of NLP which investigates the use of software to translate text or speech from one natural language to another natural language using computers with or without human assistance. The LT system which generates translation between two specific languages are called bilingual LT systems. The bilingual LT system may be either one direction or both directions. The language translation is as old as that of computers and it was the first computer based applications related to NLP. The first non-military computers were developed in 1947, from that time the idea was proposed to translate text from a source language (SL) to a target language (TL) using a computer.

At present, it is a very challenging research tasks in the area of computational linguistics and NLP in the world as well as in India. The research scenario in India is relatively young and language translation gained momentum in India only from 1980 onwards with institutions like IIT Kanpur, IIT Bombay, IIIT Hyderabad, University of Hyderabad and NCST Mumbai. The Technology Development for Indian Languages (TDIL), Centre for Development of Advanced Computing (CDAC) and Ministry of Communications and Information Technology are playing a major role in developing the LT systems.

II. Literature Review

We researched that, in a large multilingual society like India, there is a great demand for translation of documents from one language to another language. Most of the state government works in there provincial languages, whereas the central government's official documents and reports are in English and Hindi. In order to have an appropriate communication there is a need to translate these documents and reports in the respective

provincial languages. Natural Language Processing (NLP) and Machine or Language Translation (LT) tools are upcoming areas of study the field of computational linguistics.

Machine or Language translation is the application of computers to the translation of texts from one natural language into another natural language. It is an important sub-discipline of the wider field of artificial intelligence. There are certain machine translation systems that have been developed in India for translation from English to Indian languages by using different approaches. It is this perspective with which we shall broach this study, launching our theme with a brief on the machine or language translation systems scenario in India through data and previous research on machine translation.

Language Translation (LT) is one of the most important applications and research tasks of NLP which investigates the use of software to translate text or speech from one natural language to another natural language using computers with or without human assistance. The LT system which generates translation between two specific languages are called bilingual LT systems.

A bilingual LT systems have dictionary comprising of words-dictionary, sentence-dictionary, phrases-dictionary and proverbs-dictionary is used for the language translation. Each of the above dictionaries contains parallel corpora of sentence, phrases and words. The basic premise is that, if a previously translated sentence occurs again, the same translation is likely to occur again. A sentence may be seen as a combination of phrases. To translate, each sentence is divided into its constituent phrases and words, and these smaller units are translated by looking up in the sentence, phrase and word dictionaries. So, we have dataset / corpus that contains some words i.e. dictionary, phrases, proverbs, sentences and a basic tutorial of parts of speech.

III. Dataset / Corpus

Table: 1 English-Sindhi Dictionary (unclassified)

English Word	Sindhi Word
Acquire	हासिल करण
Across	आरपार
Act	अमल करण
Actual	असली
Actuality	असलियत
Actually	हक़ीक़त में
Address	पतो
Additional	बियो बे
Adjacent	लगो लग्
Admit	क़बूलण
Admire	साराहिण
Adult	बालिग़

Table: 2 English-Sindhi Dictionary (classified: days of a week)

English Word	Sindhi Word
Monday	सूमर
Tuesday	मंगल
Wednesday	बुधर
Thursday	विस्पत
Friday	जुमओ
Saturday	छंछर
Sunday	आरत्वार

Table: 3 English-Sindhi Phrases

English Phrase	Sindhi Phrase
Wear the clothes	कपड़ा पाइ
Read the book	किताब पड़

See the rose	गुलाब डिस
Eat the apple	सूफ़ खा
Take the handkerchief	रुमाल वठ
See the picture	फोटो डिस
Have fun	मौज कर
Servant came	नौकर आयो

Table: 4 English-Sindhi Proverbs

English Proverb	**Sindhi Proverb**
Union is strength	ब् त बारहां
Every dog is valiant at its own door	पहिंजे घर मे बिलि भि शेर
East or west, home is best	पहिंजो घर गुरुअ जो दर
When you are in Rome, do as Romans do	जहिड़ो देश तहिड़ो वेश
As you sow, so shall you reap	जहिड़ी करणी तहिड़ी भरणी

Table: 5 Translation of some sentences from English to Sindhi language

English Sentence	**Sindhi Sentence**
How many potatoes?	घणा पटाटा?
Don't pluck the flower.	गुल नअ पट
Don't always see the faults.	हमेशाह ऐब नअ डिस
I am a girl.	मां छोकरी आहियां
I am a boy.	मां छोकरो आहियां
India is my country	भारतु मुंहिंजो देशु आहे
My name is Pinky.	मुंहिंजो नालो पिंकी आहे

Table: 6 Parts of Speech

१) इस्मु (Noun) कंहिं बि साहवारे, माण्हूअ, शइ, वक़्तु, जाइ, कम, हालत या खासियत(गुण), जे नाले खे इस्मु चइबो आहे |
२) ज़मीरू (Pronoun) जुमिले मैं को बी लफ़्ज़, इस्म जे बदिरां कमु अचे त उनखे ज़मीरू चइबो आहे |
३) सुफति (Adjective) जेको लफ़्ज़ु इस्म यां ज़मीर सां लगी उनजो गुणु, अवगुणु, अंदाज़ु, रंगु, कदु, किस्मु, खासियत, मकिदार या हालत डेखारे तंहिंखे सुफति चइबो आहे |
४) फइलु (Verb) कमु बुधाईंदड़ लफ़्ज़ खे फइलु चइबो आहे | फइल जी माना आहे कमु | फइल मां हूअण, करण,सहण, थियण ऐं पवण जी खबर पवे थी |
५) ज़र्फु (Adverb) उहो लफ़्ज़ जेको फइल, सुफति, या बिए ज़र्फ सां लगनि ऐं जिनमां वक़्त, जाइ, रीति, कदुरू, सुवाल, या नाकार जी माना निकिरे तंहिं खे ज़र्फु चइबो आहे |
६) हर्फु जरि (Preposition) उहो लफ़्ज़ जेकी इस्म यां ज़मीर जे पुठियां कमि अचनि ऐं उन जो बिए कंहिं इस्म यां ज़मीर सां लागापो डेखारीनि तिनी खे हर्फु जरि चइबो आहे |
७) हर्फु जुमिलो (Conjunction) जेको लफ़्ज़ बिनि लफ़्ज़नि या जुमिलनि खे गंढे, तंहिंखे हर्फु जुमिलो चइजे थो |
८) हर्फु निदा (Interjection) उहे लफ़्ज़ खुशी, शोक, नफरत, अजब जहिड़ा भाव डेखारीनि तिन खे हर्फु निदा सडिजे थो |

IV. Implementation

Machine or Language Translation is the branch of computational linguistics deals with translating of one language to another or from source language (SL) to target language (TL) with or without assistance of human. There has been advanced research is going on in the field of computational linguistics that changes the way of dealing with the data as a means of effective communication in the society. In a multilingual country like India the demand for translation tool as a means of exchange of information between the regions [20].

Various methodologies have been developed to automate the translation process [17]. However, the objective has been "to restore the meaning of original text in the translated verse". In general, the process of translation has two levels:

1. Meta-phrase (word)

Meta-phrase means "word-to-word" translation. It relates to "formal equivalence", i.e., the translated version will have "literal" translation for each word in the text. However, the translated text may not necessarily convey the meaning of the original text. That means sometimes the semantics may differ from the original text.

2. Para-phrase (text)

Para-phrase means "dynamic equivalence", i.e., the translated text would contain the central idea of the original text but may not necessarily contain the word-to-word translation.

Different methods of language translation are used to translate source text into target text. In this paper, we are using two methods such as:

- **Dictionary based:**

This method of translation is based on entries of a language dictionary. The word's equivalent is used to develop the translated verse. The first generation of translation was entirely based on machine-readable or electronic dictionaries. To some extent this method is still helpful in translation of phrases but not sentences.

- **Corpus based:**

Corpus based approach for translation has emerged as one of the widely explored area in translation. Because of high level of accuracy achieved during the translation, this method has dominated over other approaches.

Below figure 1 shows translation process in which source text is translated into target language by using database/corpus. In this, firstly source text is searched in the database / corpus by simply checking each alphabet in English word dictionary and provides the corresponding Sindhi word in target text.

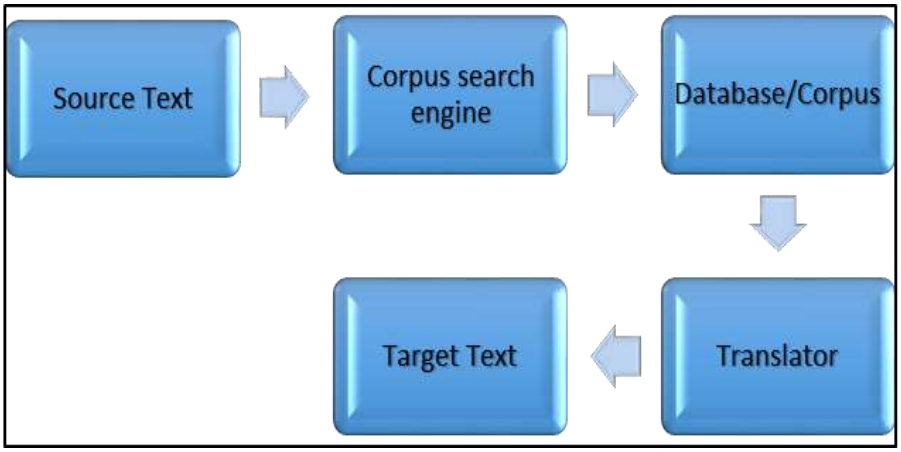

Figure 1: Language Translation Process

The LT system is very important for human for the following reasons:

- ✓ Huge amount of text can be translated from one natural language to another natural language using a LT system.
- ✓ It can be used to reduce the human efforts and to give the translation results quickly.
- ✓ The use of LT system can increase the volume and speed of translation throughput.

- ✓ Manual translation for translating the huge amount of text document is not only time consuming, but also need a more expense. Therefore, LT system can be used to save time and reduce cost.

Hence, from a scientific point of view, machine or language translation remains the classic acid test of how much we understand about human language [16].

V. Conclusions

Many of the translation software available are used to execute literal translation of the text. But none of the solutions is perfect to create dynamic equivalence between the translated and original text. Every method of language translation has its own advantages and drawbacks [17].

The different approaches for language translation suggest that the translation at meta-phrase level is attainable through the currently available translation software but achieving para-phrase level or dynamic equivalence between the source and target language still appears to be a far-fetched dream for computer linguists.

Language is evolutionary in nature; hence, it is difficult to say that one approach would be sufficient to handle the translation process. Knight [16] stated that translation as one of the elusive goals for computer science research. A decade has passed since the statement was made but it still appears to be relevant looking at the level of accuracy being achieved in translation. Linguistic irregularities, ambiguities, lack in universality of grammar and lexicon, are some of the reasons behind the failure of systems to achieve 100 percent accuracy in the translation.

Some of recommendations are made to achieve the utmost accuracy in translation are developing universal meta-language; standardization of lexical organization and content and development of translation oriented multilingual textual corpus. However, the suggested measures appear to be unachievable looking at vast number of variants of existing languages.

The process of translation involves a thorough analysis of the source text which require various course of action such as study of grammar; semantic and syntax analysis, etc. A thorough knowledge as well as understanding of the target language is also must for a translator (whether human or machine supported).

Thus, information professionals may depend upon these tools (Google Translate etc.) only to certain extent i.e. just to gather the central idea about the document. Then they can take help of language experts about the content of document before classifying or making it available for users. Thus, these tools may be helpful for initial sorting of the documents not as regular tools for analysis of the content of the other language document.

References

[1] Nirmala Chawla and Bharati Batheja, "*Sindhi Devnagri*", 2019, pp. 50-52.

[2] Nadeem Jadoon Khan, Waqas Anwar & Nadir Durrani, "Machine Translation Approaches and Survey for Indian Languages", 2017.

[3] ALPAC "Language and Machines: Computers in Translation and Linguistics". A report by the Automatic Language Processing Advisory Committee (Tech. Rep. No. Publication 1416), 2101 Constitution Avenue, Washington D.C., 20418 USA: National Academy of Sciences, National Research Council, 1966.

[4] Balajapally, P., Bandaru, P., Ganapathiraju, M., Balakrishnan, N., & Reddy, R., "Multilingual Book Reader: Transliteration, Word-to-Word Translation and Full-text Translation", 2006.

[5] Dwivedi, S. K., & Sukhadeve, P. P., "*Machine Translation System in Indian Perspectives*", Journal of Computer Science, 6(10), 1111-1116, 2010.

[6] 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.

[7] Hasler, E., Haddow B., and Koehn, P., "*Sparse lexicalised features and topic adaptation for SMT*", In Proceedings of the seventh International Workshop on Spoken Language Translation, pages 268–275, 2012.

[8] Och, F., "A systematic comparison of various statistical alignment models", Computational Linguistics, 29(1):19–5, 2003.

[9] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "*BLEU: a Method for Automatic Evaluation of Machine Translation*", In Proceedings of 40th Annual meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 311–318.

[10] Bisazza, A. and Federico, M., "*Chunk-based verb reordering in VSO sentences for Arabic-English statistical machine translation*", In Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR, WMT '10,15-16, july 2010, pp. 235–243.

[11] J. Eisner, "*Learning non-isomorphic tree mappings for machine translation*", In Proceedings of the ACL Interactive Poster/Demonstration Sessions, 2003, pp. 205–208.

[12] Dash, Niladri Sekhar, Chaudhuri, Bidyut Baran, "*Why do we need to develop corpora in Indian languages*?" A paper presented at SCALLA 2001 conference, Bangalore.

[13] Rao, Durgesh, "*Machine Translation in India: A Brief Survey*", SCALLA 2001 conference, Bangalore.

[14] Naskar, S., & Bandyopadhyay, S, "*Use of Machine Translation in India: Current Status*", In Proceedings of MT SUMMIT X; September 13-15, 2005, Phuket, Thailand.

[15] Bandyopadhyay S., "*ANUBAAD - The Translator from English to Indian Languages*" In proceedings of the VIIth State Science and Technology Congress, Calcutta. India, 2000, pp. 43-51.

[16] Knight K, "*Automating Knowledge Acquisition for Machine Translation*", Artificial Intelligence Magazine, Vol. 18 (4) (1997).

[17] Sneha Tripathi and Juran Krishna Sarkhel, "*Approaches to machine translation*", Annals of Library and Information Studies, Vol. 57, December 2010, pp. 388-393
[18] SAIFUL ISLAM and BIPUL SYAM PURKAYASTHA, "*A Review on Electronic Dictionary and Machine Translation System Developed in North-East India*", Vol. 10, June 2017, No. (2): Pgs. 429-437.
[19] Richard Zens, Franz Josef Och, and Hermann Ney, "*Phrase-Based Statistical Machine Translation*", 2002, pp. 18–32.
[20] Sindhu D.V and Sagar B M, "*Study on Machine translation approaches for Indian languages and their challenges*", 2016, pp. 262-267.