ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

IJARCCE

nCORETech

LBS College of Engineering, Kasaragod

Vol. 5, Special Issue 1, Feburary 2016

English Text to Malayalam Speech Translation

Deepthy P S

P G Student, Computer Science & Engineering, NSS College of Engineering, Palakkad, India

Abstract: Natural Language Processing (NLP) is a field of software engineering, counterfeit consciousness and computational semantics which were worried with collaborations in the middle of PCs and human languages. Machine interpretation is a standout amongst the most essential uses of Natural Language Processing. This is the one of the vital technique that helps people from different places to understand an unknown language without the help of a human translator .The language to be translated is the Source Language (SL) and the language to which source language translated is Target Dialect (TL). While translating, the syntactic structure and semantics structure of both source and target language ought to be considered. Although there has been a ton of works distributed for widespread languages like English, works in language like Malayalam is similarly less. One of the most rising research area is in NLP and machine translation, which is from English to Malayalam. In this field this paper is centering upon the speech translation into Malayalam. It has an assortment of uses in the spaces like magazines,books,film reviews,etc.The approach included in this procedure is as follows. Here we propose a technique for the translation of English sentences to Malayalam and afterward into speech. This machine translation is done by rule based method and statistical method combined. The core process in this translation is mediated by bilingual dictionaries and rules for converting source language structures into the target language structures. The rules that are to be used in this approach are prepared based on the Parts Of Speech (POS) tags and dependency information that are obtained from the parser. There are mainly 2 types of rules are used here, one is transfer link rule and the other is morphological rules. In this method, for generating target structure, transfer link rules are used .Morphological rules used here are for assigning morphological features. The bilingual dictionary used here is an English, Malayalam bilingual dictionary. By this approach, a given English sentence can be translated to its Malayalam equivalent. And finally statistical method is used to reduce the error in the translation. Thus the first phase is completed, and in the next phase the translation of Malayalam text to speech is done, which is the main area on which this paper is focusing on. Hence it is concluded that speech translation is a very important area to be considered in Malayalam Language. Smart sermonis is a translator which convert English text to Malayalam speech. It provide a short but a comprehensive overview of Text-To-Speech synthesis by highlighting its digital signal processing component.

Keywords: Component, TTS, HMM Synthesis, Phoneme, Prosody, Concatenation Synthesis.

1. INTRODUCTION

Machine translation is the translation of text from one language to another, with minimum manual involvement. It eliminates the language barrier allowing people to communicate and access resources available in different languages. The language being translated is the source language and the language to which it gets translated is the target language. The translations are language specific and manual translations are time-consuming. There are various approaches to machine translation (MT) mainly, rule based and corpus based. Rule Based Machine Translation (RBMT) translates text based on syntactic and semantic rules, whereas Statistical Machine Translation (SMT) is corpus dependent and uses probability measures for translation. Statistical translations require no linguistic data and produces correct output if parallel corpus is available. But it cannot be applied on languages with limited computational resources. A RBMT system can be relied in such cases. These systems offer good quality translation results, since it is based on rules and linguistic information. The different methods used in RBMT are Direct, Transfer and Interlingua based translation. The system discussed in [1] translates, based on the syntax tree reordering using transfer approach whereas [2] uses tree substitution grammars. The proposed system uses transfer approach for translation of text from English to Malayalam. In order to improve the quality of output, the translation is made domain specific. The 2 feature is that multiple systems of various domains can be integrated to produce a single system that translates text irrespective of domain .

2. BACKGROUND

Translations from English to Indian languages were of great interest to the Indian government and it was actively undertaken by Technology Development for Indian Languages (TDIL) [4]. The project was held in two phases titled Anuvadaksh and AnglaBharati. The latter translates English to a pseudo Interlingua representation from where it is

DOI 10.17148/IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

IJARCCE

nCORETech

LBS College of Engineering, Kasaragod

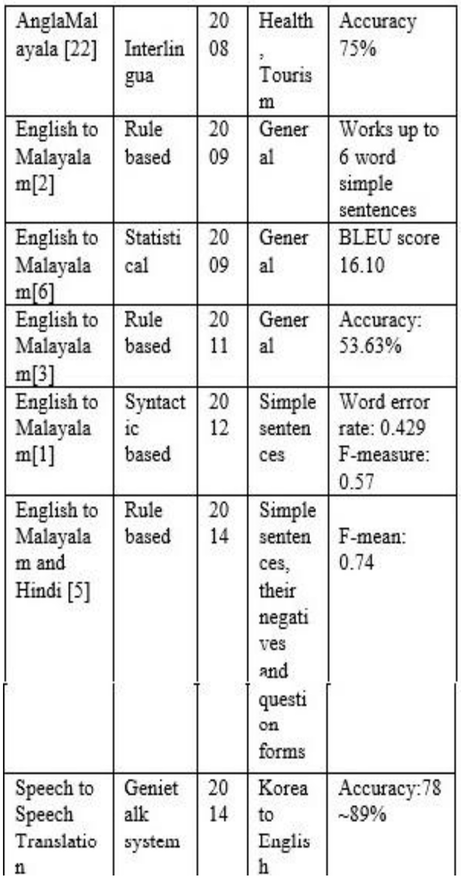
Vol. 5, Special Issue 1, Feburary 2016

translated to Indian languages. AnglaMalayalam developed by CDAC [Centre for Development of Advanced Computing] using AnglaBharati translation engine translates texts from English to Malayalam but is limited only to Health, Tourism domains. Some of the RBMT systems for language other than Malayalam are given in [5], [6]. The English to Malayalam translation systems are limited to a few. In [9] the translation works for 10 specified sentence patterns. In [7] translation works up to six worded sentences. The table I. shows an analysis of existing English to Malayalam MT systems [8]. Since these systems were evaluated using different methods a comparison on the improved translation quality of output is irrelevant. Also they were evaluated for different input texts and system like AnglaMalayalam provides one or more outputs

2.1 Statistical Machine Translation

Statistical machine translation is a data oriented statistical framework for translating text from one natural language to another based on the knowledge extracted from bilingual corpus. SMT systems make use of a combination of one or more translation models and a language model. In this paper we explore how a direct, well aligned corpus (English – Malayalam) will work on SMT system, it's decoding and evaluation and how to improve the translation by adding rules, and by adding morphological information to the Malayalam. The information society we live in is undoubtedly a globalized and multilingual one. Every day, hundreds and thousands of documents are being generated, and in many cases one or several translations for them are needed in order to cover the linguistic variety of the target population. The majority of work carried out by professional translators is related to non-literary documents (technical reports, legal and financial documents, user manuals, political debates, meeting minutes, and so on),where translation tends to be mechanical and domain-specific. However, the high translation cost in terms of money and time is a bottleneck that prevents all information from being easily spread across languages. To a large extent, much of the optimism being shared in the MT research community now a days has been caused by the revival of statistical approaches to machine translation, or in other words, the birth of purely Statistical Machine Translation3. In contrast to previous approaches based on linguistic knowledge representation, SMT is based on large amounts of human-translated example sentences (parallel corpora) in order to estimate a set of statistical models.

TABLE I. ANALYSIS OF EXISTING ENGLISH-MALAYALAM MT SYSTEMS.



AnglaMalayala [22]	Interlingua	2008	Health, Tourism	Accuracy 75%
English to Malayalam[2]	Rule based	2009	General	Works up to 6 word simple sentences
English to Malayalam[6]	Statistical	2009	General	BLEU score 16.10
English to Malayalam[3]	Rule based	2011	General	Accuracy: 53.63%
English to Malayalam[1]	Syntactic based	2012	Simple sentences	Word error rate: 0.429 F-measure: 0.57
English to Malayalam and Hindi [5]	Rule based	2014	Simple sentences, their negatives and question forms	F-mean: 0.74
Speech to Speech Translation	Genietalk system	2014	Korea to English	Accuracy:78 ~89%

DOI 10.17148/IJARCCE

2.2 Reordering and Morphological Processing

The main ideas which have proven very effective are (i) reordering the English source sentence according to Malayalam syntax, and (ii) using the root suffix separation on both English and Malayalam words. The first one is done by applying simple modified transformation rules on the English parse tree, which is given by the Stanford Dependency Parser. The second one is developed by using a morph analyzer. This approach achieves good performance and better results over the phrasebased system. Our approach avoids the use of parsing for the target language (Malayalam), making it suitable for statistical machine translation from English to Malayalam, since parsing tools for Malayalam are currently not available.

Statistical Machine Translation for Malayalam language gives poor result, if we provide parallel corpus directly, because of the following reasons;(i)English follows SVO (Subject – Verb – Object) word order but Malayalam follows SOV (Subject – Object – Verb) word order;(ii)Malayalam language is morphologically quite rich and (iii)huge tagged parallel corpus are not available for EnglishMalayalam language pairs. So the technique of including rule based reordering and morphological processing for statistical machine translation (SMT) for Malayalam gives more accuracy .In this paper, we present our work by including rule based reordering and morphological information for English to Malayalam.

2.3 Text to Speech System For Malayalam

Text-to speech (TTS) systems which mainly meant for speech synthesis are, used for one of the South Indian languages called Malayalam. The paper makes a brief study on, Malayalam linguistics, and also gives a Comparison between two prominent methodologies for speech synthesis, viz Concatenative based synthesis and HMM based synthesis. As a result, the paper mentions some of the problems facing by Concatenative based TTS systems and thereby, the research goes on with HMM synthesis. The paper also done a proposal for TTS system for Malayalam which is statistical based using HMM's (Hidden Markov Model). Natural language processing (NLP) is a field of computer science, artificial intelligence (also called machine learning), and linguistics concerned with the interactions between computers and human (natural) languages. Internet domain consists of huge amount of various data.

So, we need applications for processing this large amount of texts. Thus we requires of NLP expertise usually called computational linguistics. Speech Synthesis which is a prominent area under NLP that is having so much importance in researches, has introduced Text to Speech Systems (TTS) for almost all foreign and Indian languages. Among the applications of speech technology, the automatic speech production, which is referred to as text-to speech (TTS) system is the most natural sounding technology. The text-to-speech (TTS) system will convert ordinary orthographic text into acoustic signal which is indistinguishable from human speech.

2.4 Speech Synthesis Techniques

Text-To-Speech synthesis is by highlighting its digital signal processing component .First two rule-based synthesis techniques (formant synthesis and articulatory synthesis) are explained the concatenative synthesis is explored. Concatenative synthesis is simpler than rule based synthesis, since there is no need to determine speech production rules. However, it introduces the challenges of prosodic modification to speech units and resolving discontinuities at unit boundaries. Prosodic modification results in artefacts in the speech that make the speech sound unnatural. Unit selection synthesis, which is a kind of concatenative synthesis, solves this problem by storing numerous instances for each unit with varying prosodies. The unit that best matches the target prosody is selected and concatenated. To resolve mismatches speech synthesis system combines the unit selection method with Harmonic plus Noise Model (HNM). This model represents speech signal as a sum of a harmonic and noise part. The decomposition of speech signal into these two parts enables more natural sounding modifications of the signal. Finally Hidden Markov model(HMM)synthesis combined with an HNM model is introduced in order to obtain text to speech system.

3. OVERVIEW

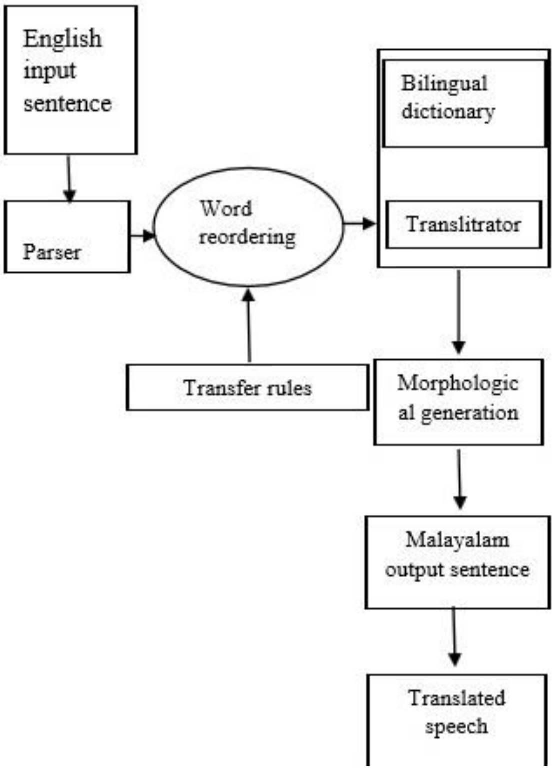
Machine Translation is the task of automatically converting one natural language into another, preserving the meaning of the input text, and producing fluent text in the output language. While machine translation is one of the oldest subfields of artificial intelligence research, the recent shift towards large-scale empirical techniques has led to very significant improvements in translation quality.

Machine translation is the process of translating from source language text into the target language. Natural Language Processing or Computational Linguistics deals with understanding and developing computational theories of human language. Such theories allow us to understand the structure of language and build computer software that can process

DOI 10.17148/IJARCCE

language. The design architecture of the proposed system is as shown in Fig.1. The architecture shows the stepwise procedure of how input English sentences get translated to Malayalam sentences.

Fig. 1. Design Architecture of RBMT system.



The English input sentence is preprocessed first and is given to the parser. The system requires pre-processing for some of the technical terms used in cricket domain. Such terms are connected with hyphens so that the parser considers them as a single term; examples are: Hawk's-Eye, cover-drive etc. The system depends on Stanford Parser [10] tool for obtaining the Part-of-Speech (POS) tag of words in the input sentence as well as for obtaining source parse tree. The words in the input sentence appear as the leaf nodes of the tree.

The source parse tree is reordered using the transfer rules to obtain target parse tree. The rules are devised such that they follow the Subject – Object – Verb (SOV) order of Malayalam language. The words from parse tree are mapped to bilingual English – Malayalam dictionary to obtain equivalent Malayalam words. The named entities are transliterated using phonetic based transliteration module. Later, necessary inflections are appended to the words in order to produce meaningful translation. The final Malayalam sentence provides the translation of inputted text.

The machine translation of text from English to Malayalam can be explained in three phases such as SOV reordering phase, Dictionary mapping phase and Morphological inflection generation phase. The prerequisite to be satisfied by input text is that they should conform to grammatical structures of English language. The lengthy sentences with more than 15 words are manually split into smaller sentences at the connectives like coordinating conjunctions..

4. CONCLUSION

This RBMT machine translation is an effective technique for translating simple English sentence to Malayalam. When translating, there are situations in which an English word can have multiple Malayalam meaning. This will give multiple translated output sentence. This problem can be solved by the efficient implementation of Word Dictionary File. The main idea of this paper is to translate English text to Malayalam speech. In present generation usage of internet had increased widely where people are getting used to it. Machine translation will helps people to understand the context of original text in their own language.

DOI 10.17148/IJARCCE

ACKNOWLEDGEMENT

This work is done as part of the academic project work supported by teaching staff of NSS College of Engineering. I then thank all my friends and colleagues who helped me, my parents and almighty.

REFERENCE

[1] Nair, A.T.; Idicula, S.M., "Syntactic Based Machine Translation from English to Malayalam," Data Science & Engineering (ICDSE), 2012 International Conference on , vol., no., pp.198-202, IEEE 2012.

[2] Jiajun Zhang; Feifei Zhai; Chengqing Zong, “Syntax-Based Translation With Bilingually Lexicalized Synchronous Tree SubstitutioGrammars,” Audio, Speech, and Language Processing, IEEE Transactions on , vol.21, no.8, pp.1586,1597, Aug. 2013 Communications (ICACC), 2014 Fourth International Conference on , vol., no., pp.223- 226, IEEE 2014.

[3] Andreas Eis et al., “Using Moses to Integrate Multiple Rule-Based Machine Translation Engines into a Hybrid System,”Proceedings of the Third Workshop on Statistical Machine Translation, pp 179–182, Columbus, Ohio, USA, June 2008.

[4] Technology Development for Indian Languages, DIT, Government of India, Available http://www.tdildc.in/index.php

[5] Meera, M.; Sony, P., "Multilingual Machine Translation with Semantic and Disambiguation," Advances in Computing and

[6] Pratik Desai, Amit Sangodkar, Om P. Damani, “A Domain-Restricted, Rule Based, English-Hindi Machine Translation System Based on Dependency Parsing” - 11th International Conference on Natural Language Processing, ICON 2014. Translation: English to Malayalam: A Survey,”

[7] Rajan, R.; Sivan, R.; Ravindran, R.; Soman, K.P., "Rule Based Machine Translation from English to Malayalam," Advances in Computing, Control, & Telecommunication Technologies, 2009. ACT '09.International Conference on , vol., no., pp.439-441, IEEE 2009.

[8] Aasha V.C, Amal Ganesh, “Rule Based Machine in Proc. 3rd International Conference on Advanced Computing, Networking, and Informatics (ICACNI), Bhubaneshwar, June 2325, 2015., Smart Innovation, System and Technologies[ISSN: 2190-3018], Springer Verlag, in press. doi: 10.1007/978-81-322-2538

[9] KevinKnight,”A statistical MTTutorial Workbook”, prepared in connection with the JHU Summer workshop April 30, 2004.

[10] Yamadaand knight,”A syntax based statistical translation model”,2001..

[11] Michael Collins,Philipp Koehn, and Ivona Kucerova,Clause Restructuring for Statistical MachineTranslation, Proceedings of ACL,2003.

[12] Philip Koehn, Franz Josef Och, and DanielMarcu,StatisticalPhrase-based Translation, Proceedingsof HLT-NAACL,2003.

DOI 10.17148/IJARCCE