IJCAT - International Journal of Computing and Technology
Volume 1, Issue 1, February 2014
www.IJCAT.org

English Text to Malayalam Speech-A Survey

[1] Jemy Joy, [2] Devika M D, [3] Della T J, [4] Afsheena M

[1, 2, 3, 4] Department of Information Technology, Jyothi Engineering College
Cheruthuruthy,Thrissur, Kerala,India

Abstract - The biggest innovation in the field of Natural language processing is a field of computer science, artificial intelligence and linguistics concerned with interaction between computers and human (natural) languages. Machine translation is one of the most commonly researched task in Natural language processing. Machine translation automatically translate text from one human language to another. Machine translations become complicated when dealing with morphologically rich language likes malayalam. Machine translation to morphologically rich languages is challenging, due to lexical sparsity on account of grammaticaly features being expressed with morphology English to malayalam translation is still under research. 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

Natural language processing is the branch of information science that deals with natural language information .Many challenges in NLP involve natural language understanding -- that is, enabling computers to derive meaning from human or natural language input. Short for **n**atural language **p**rocessing, a branch of artificial intelligence that deals with analyzing, understanding and generating the languages that humans use naturally in order to interface with computers in both written and spoken contexts using natural human languages instead of computer languages. One of the challenges inherent in natural language processing is teaching computers to understand the way humans learn and use language. The phrase "natural language **processing**" may or may not be taken as synonymous with "natural language understanding","Processing" most naturally is used for both **interpretation** and **generation**, while one would think "understanding" is better used for only the **interpretation** part. A system is capable of natural language understanding does not imply that the system can generate natural language, only that it can interpret natural language. To say that the system can process natural language allows for both understanding (interpretation) and generation (production). But the phrase "natural language understanding" seems used by some authors as synonymous with "natural language processing," and on this use includes interpretation and generation. Language translation becomes complicated when dealing with morphologically rich languages like malayalam.

2. 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 language.

3. Literature Survey

3.1 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.

Apart from that, the growth and popularity rise of internet has given users access to practically any written, visual and audio material from anywhere in the world. Still, the language barrier is the only for this vast information to be fully shared by all users.

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 orderto estimate a set of statistical models**.**

3.2 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 phrase-based 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 English-Malayalam 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.

3.3 Text to Speeech 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.

3.4 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 inorder toobtain text to speech system.

4. Consolidation

PAPER	REMARK
An understanding to english-malayalam statistical machine translation	• Refinement in corpora is needed. • Reorder the English sentence as per Malayalam. • Use the suffix of Malayalam words.
Rule based reordering and morphological processing for English-Malayalam statistical machine translation	• Malayalam is a morphologically based language • Need to improve the morphological analyzer • Need to get more corpora from different domain
Text to speech system for malayalam	• Makes a brief studyon, Malayalam linguistics, and gives a comparison between two prominent methodologies for speech synthesis concatenative based synthesis and HMM synthesis.
Speech synthesis techniques	• To provide a short overview of text-to-speech synthesis by highlighting its digital signal processing component. • It includes rule-based synthesis techniques and concatenative synthesis.

5. Conclusion

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 orginal text in their own language.

References

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

[2] Yamadaand knight,"A syntax based statistical translation model",2001..

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

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