

Development of Technology for Indian Languages: Indian Government Initiatives

Sunil Kumar Srivastava

Government of India
Ministry of Electronics and Information Technology
Electronics Niketan, 6, Lodi Road, New Delhi - 110003
sks@meity.gov.in

Abstract

With twenty two constitutionally recognized languages written in twelve different scripts and several hundred dialects, India faces a major challenge when it comes to the issue of language. Under Technology Development for Indian Languages (TDIL) Programme, Government of India has sponsored several projects which has led to development of technology and tools in the areas like machine translation, automatic speech recognition, optical character recognition, etc. Government is now initiating Natural Language Translation Mission which aims at building and deploying operational machine translation systems for major Indian languages.

बारह अलग-अलग लिपियों में लिखी जाने वाली और कई सौ बोलियों में बोली जाने वाली बाईस संविधान द्वारा मान्यता प्राप्त भाषाओं के प्रयोग में होने के कारण, भारत को भाषा के बिंदु पर एक बड़ी चुनौती का सामना करना पड़ता है। भारतीय भाषाओं के लिए प्रौद्योगिकी विकास (टीडीआईएल) कार्यक्रम के तहत, भारत सरकार ने कई परियोजनाओं को प्रायोजित किया है, जिसके कारण मशीन आधारित अनुवाद, आटोमैटिक स्पीच रिकग्निशन, ऑप्टिकल कैरेक्टर रिकग्निशन, आदि क्षेत्रों में प्रौद्योगिकी और उपकरणों का विकास हुआ है। अब भारत सरकार प्राकृतिक भाषाओं के अनुवाद पर एक मिशन की शुरुआत कर रही है। इस मिशन का उद्देश्य प्रमुख भारतीय भाषाओं के लिए मशीन आधारित अनुवाद प्रणालियों का विकास करना और प्रयोग में लाना है।

Keywords: language technology, machine translation, India

1. Introduction

India is a country of diversity in several aspects including languages. With 22 constitutionally recognized languages written in 12 different scripts and several hundred dialects, the country faces a challenge when it comes to the issue of communication. During the recent past, India has seen an emergence of digital economy. A number of government services are being offered in digital form - ranging from birth certificate to filing petitions in the courts. Almost all kinds of organizations, public or private, commercial or non-commercial have made their presence felt in the digital space. However, one major challenge is the access to information and services in the native languages. About 15% people only can speak and write in English. The remaining population is unable to derive the benefits of IT as most of the solutions developed have interfaces in English. Language becomes a barrier for a large percentage of population.

Government of India has been taking steps towards the development of technology for Indian languages since the eighties. In early nineties, it initiated an R&D Programme titled *Technology Development for Indian Languages* [4]. During the beginning years, the programme was primarily

concerned with the development of standards for Indian languages, device drivers for Indian languages, fonts for various scripts, and localization of the popular open-source tools such as Linux, OpenOffice, etc. These tools were distributed across the country. It was especially useful for those languages where tools were not available from the vendors.

Later on, the focus shifted to technology development in several new areas including, but not limited to, automatic speech recognition (ASR), text-to-speech synthesis (TTS), optical character recognition (OCR) and machine translation (MT). Several projects were sponsored to the academic/R&D institutions in consortium mode where one institution was leading the R&D work in that area (e.g. ASR) and other member institutions were working on the development of technology for specific languages. Some of the lead institutions in the respective areas were IIT Madras (ASR and TTS), IISc Bangalore (OCR), IIIT Hyderabad (MT from Indian languages to Indian languages) and C-DAC Pune (MT from English to Indian languages). The achievements in the individual areas are briefly described below.

2. Status

Several demonstration level prototypes have been developed in the above-mentioned areas. Some of the tools such as TTS have been used by the developers in applications. These are briefly discussed below.

2.1. Automatic Speech Recognition (ASR)

Under the TDIL Programme, IIT Madras has been working on the development of large vocabulary continuous speech recognition (LVCSR) systems. It has developed a system called, Mandi which provides speech-based access to agricultural commodity prices and weather information in 11 Indian languages/dialects. This system provides current commodity prices in local markets and local weather information to users in a convenient manner. It takes the data available on AgmarkNet and AgriMet websites. The system has been built from scratch using open-source tools/software so that it can be used by public institutions with no licensing issue or cost.

2.2 Text-to-Speech (TTS) Synthesis

Text-To-Speech (TTS) synthesis system for 10 Indian languages viz. Tamil, Telugu, Marathi, Bodo, Kannada, Odia, Hindi, Malayalam, Manipuri & Rajasthani have been developed under a consortium project under the leadership of IIT Madras. Mean Opinion Score (MOS) of these TTS systems on the scale of 0 to 5 is 3.2 or more. It has also been made available in open source under creative commons (CC-BY 4.0) license and can be downloaded from [1]. One of the developed systems called m-Vachak, helps visually challenged people in accessing digital information. The system has also been made available on Android based operating system, Indus OS. As of now, there are 1.54 million activations on 8 Mobile Brands (Micromax, Celkon, Swipe, Karbonn, Intex, Trio, Sansui & Datawind) supporting Indus OS. Applications like browser plugin for Mozilla Firefox and Google Chrome, SMS Reader for Indian languages, TTS voices have been made available to the public through TDIL Data centre [3] for use and feedback.

2.3 Machine Translation (MAT)

2.3.1 Indian Language to Indian Language Machine Translation System: The system called *Sampark*, uses both rules-based and dictionary-based algorithms with statistical machine learning approach. It has been developed for 18 language pairs [Hindi↔Punjabi, Hindi↔Urdu, Hindi↔Tamil, Hindi↔Tamil, Hindi↔Bengali, Hindi↔Marathi, Hindi↔Kannada, Telugu↔Tamil, Malayalam↔Tamil]. The

system was developed by a consortium of institutions under the leadership of IIT Hyderabad.

2.3.2 English to Indian Languages Machine Translation System (AnglaMT): The system called AnglaMT, is a rule based machine translation system for English to 8 Indian languages [English↔Hindi, English↔Malayalam, English↔Bengali, English↔Urdu, English↔Punjabi, English↔Tamil, English↔Assamese, English↔Nepali] developed by a consortium of institutions under the leadership of CDAC Noida.

2.3.3 English to Indian Languages Machine Translation System (Anuvadakh): The system called Anuvadakh uses statistical and example-based machine translation techniques for translation from English to 8 Indian languages [English↔Hindi, English↔Marathi, English↔Bengali, English↔Urdu, English↔Odia, English↔Tamil, English↔Gujarati, English↔Bodo]. It was developed by a consortium of institutions under the leadership of CDAC Pune.

2.4. Optical Character Recognition (OCR)

OCR system has been developed for 13 Indian languages- Assamese, Bangla, Gurmukhi, Hindi, Kannada, Malayalam, Tamil, Telugu, Urdu, Gujarati, Oriya, Manipuri and Marathi. The preprocessing routines such as adaptive binarization, noise cleaning, skew corrections routines were developed by different consortium partners. Different classifiers such as SVM, KNN, LSTM were used. These can be used under Windows, Linux and Web version.

2.5. Language Technology Tools

A large number of tools have been developed for all major Indian languages. These tools have been used in several applications for Indian languages. The centres have also developed linguistic resources like dictionaries, taggers, spell checkers, CLDR, grammar checkers, sorting utilities, thesauri, tagged lexicons, and information extraction & retrieval and standards.

2.6. National Platform for Language Technology

The outcomes of the projects undertaken under TDIL Programme have been showcased at the Indian Language Technology Proliferation & Deployment Centre portal. This portal has been acting as a national repository for linguistic resources, tools and applications being developed under the various TDIL sponsored projects. Now, the portal

has been redesigned and is being launched as National Platform for Language Technology (<http://nplt.gov.in>). The portal will work as an e-marketplace for linguistic resources and tools.

3. Natural Language Translation Mission

During March 2019, nine national science & technology missions were announced by the Principal Scientific Adviser to Government of India. These missions have been recommended by Prime Minister's Science, Technology & Innovation Advisory Council (PM-STIAC) [2].

3.1. Objectives:

The objectives of the mission are the following:

- 3.1.1.** To build a high-quality speech to speech machine translation (SSMT) system for major Indian languages;
- 3.1.2.** To create and nurture an ecosystem involving start-ups, central/state government agencies working together to develop and deploy innovative products and services in Indian languages;
- 3.1.3.** To increase the content in Indian languages on Internet substantially in the domains of public interest, particularly science & technology, education, healthcare, governance, and law & justice, etc.

3.2. Implementation Strategy:

There are five elements in the mission: R&D consortia, CoEs, start-ups, sub-missions and National Hub for Language Technology (NHLT). Under TDIL Programme, several R&D consortia were created to develop technology for Indian languages and these have been working for almost a decade. In the present mission, four consortia will be working on the respective areas viz. Speech Technology (ASR & TTS), English to Indian Languages Machine Translation (EILMT), Indian language to Indian language Machine Translation (ILMT), and Optical Character recognition (OCR). The consortia will work towards upgrading the technology and will also provide technical assistance to CoEs, start-ups, etc.

CoEs, the second element will be responsible for translating the lab prototypes into commercial products. It

has been seen that the academic institutions have not been able to take the lab prototypes from the lab to the land as their focus has been on R&D. These centers, essentially, would be Center of Engineering which will do necessary software engineering for converting the lab prototype into a commercial product. CoEs will also provide support to the developers, especially start-ups who would use the technology to develop solutions to meet the requirements of the users.

Start-ups are the third strategic element in the mission. The start-ups are being envisaged to be primary vehicle for developing the applications and providing services in language technology space. These will also be used to create the required high volumes of language data. The academic institutions will provide technical guidance to the start-ups in the process of resource creation.

Sub-Missions on individual languages are the fourth element of the Mission. These will be launched with the participation of the states. Each sub-mission will focus on one of the recognized languages. The sub-mission will be handled by the state where it is used. In case of the languages which are spoken across several states, all the states will be participating. In order to increase the content in Indian languages on the internet, the content available on the Internet will be translated into Indian languages using machine translation systems followed by review by human translators. Once corrected, the content will be made available for training of the machine translation systems.

Finally, it is proposed to create a National Hub for Language Technology (NHLT) to provide services and central facilities including National Machine Translation Service through Bahu-Bhashak Platform, National Language Technology Platform for resource sharing. NHLT will also be responsible for conducting Grand-Challenges and contests in the area of language technology.

4. Conclusions

The article has described the activities which have been undertaken towards the development of technology for Indian languages. Efforts have been made since the nineties. However, though some good prototype systems have been developed, very few have reached to the end users. In order to make use of the advances in language technology space, Government of India is initiating a mission on natural language translation which aims at

developing machine translation systems for all major Indian languages.

5. Bibliographical References

1. Indic TTS Project.
<https://www.iitm.ac.in/donlab/tts/>
2. PMSTIAC - Missions.
<http://psa.gov.in/pmstiac-missions>
3. TDIL Data Centre. <http://www.tdil-dc.in>
4. TDIL Programme. [<http://tdil.meity.gov.in/>