Resource Creation for Sanskrit ASR (Automatic Speech Recognition)

[1]Devendr Kumar1, [2]Girish Nath Jha
[1]School of Sanskrit and Indic Studies, [2]Jawaharlal Nehru University, New Delhi, India
devneed2@gmail.com & girishjha@gmail.com

ABSTRACT

There are a few works on Automatic Speech Recognition (ASR) for Sanskrit. Generally, developing an ASR system is a time and cost-consuming multi-layered teamwork. An average of basic 90 to 100 hours of annotated speech corpus is required for the development of a basic ASR system. The present paper is part of a doctoral research work that proposes corpora creation to provide a reliable annotated speech dataset for future research on Sanskrit ASR.

Keywords: Resource Creation, Automatic Speech Recognition for Sanskrit, Indian Language, Sanskrit manuscripts, Transcription

INTRODUCTION

Sanskrit is one of the 22 languages enlisted in the Eighth Schedule of the Indian Constitution. Sanskrit has been a vast and vital source of knowledge since ancient times. Sanskrit is an ancient Indo-Aryan language also considered as one of the oldest classical languages of the world. It has a long and rich literary tradition and was the language of scholarship and religious texts in ancient India. Sanskrit is known for its complex grammatical structure and highly refined phonetics system. Ancient knowledge of any discipline could be found in Sanskrit in the form of manuscripts, which are decaying day by day due to improper care and preservation. Therefore, in today's time, the preservation of these manuscripts could be made faster using digitization and other higher-level language processing techniques. One of these techniques are through creating Sanskrit Automatic Speech Recognition (ASR).

As part of this endeavor, 80 to 100 hours of audio data is collected for creating Sanskrit ASR system. Audio data will be recorded by different participants and pre-recorded sets of data will be collected from various sources, including those available on open web sources. A speech corpus will be created using this audio data which plays a crucial role in Sanskrit ASR. Readily available software like PRAAT etc. could be used for annotating the entire audio data. This will help annotate audio and text files at Sentence and Word levels.

MOTIVATION

ASR is one of the successful applications of NLP now a days. ASR has reached almost the mature stage with 90% plus accuracy in some languages such as English and Hindi. ASR technology of Google, Amazon, Apple, and Microsoft etc. has become state of the art with widely used applications and a high number of user inputs. To know the knowledge contained in a language, it is necessary to know that language. Studies in Sanskrit academies, journalism, and manuscripts have a high demand for digitization work. People still find it difficult to type in Sanskrit. Therefore, it becomes necessary to create Sanskrit ASR so that people do not face problems in typing, and so far, very little work has been done in Sanskrit ASR. Sanskrit ASR can be used in many areas, such as e-learning, text digitization, manuscript protection, journalism, and many more. Hence, this is high time to look into the latest ASR models for Sanskrit.

RELATED WORK

The major literature of our work is constituted by general works on ASR and Indian language speech corpora creation under the project of the government of India[1]. Anoop and Ramakrishnan (2019) presented a Sanskrit ASR system built on a trained model of 2:35 hours of data annotated in three levels of 46 phonemes, 8370 words, and 1360 sentences. It reports accuracies of 62% at phoneme level, 89% at word level and 58% at sentence level. Hindi gets special attention in industries in comparison with other modern Indian languages. TDIL[2] has undertaken projects on speech corpora creations for the five languages of Tamil, Telugu, Marathi, Bengali, and Assamese. These datasets were prepared from agricultural commodity domain. A table is given below for the speech dataset with basic information.

Name of Language	Number of Transcribed audio files	Size of corpus file in GB	Number of speakers	Domain
Telugu	1077	5.4	1073	Agricultural Commodity
Tamil	62000	5.7	1000	Agricultural Commodity
Marathi	44500	2.8	1500	Agricultural Commodity
Assamese	3600	1.9	1023	Agricultural Commodity
Bengali	43000	1.8	1000	Agricultural Commodity

[1] http://www.tdil-dc.in/index.php?searchword=Text%20Corpora&searchphrase=all&option=com_search&lang=en&limitstart=40&limit=20

[2] Technology Development for Indian Languages, Ministry of Electronics & Information Technology (MeitY), MC&IT, Govt of India.

Table 1. Dataset for Agriculture Commodity Domain

HYPOTHESIS

ASR is a successful technology. Sanskrit ASR application may have multi-dimensions. Generally, a baseline ASR system requires 100 hours of annotated data. Sanskrit is a consistent language in the sense that it maintains a standard grammar for over 2500 years. Developing an ASR system could be lesser tedious task for the language engineers as this follows phonetic patter of writing system (sound- letter correspondence). With easy accessibility of internet, Sanskrit raw text and audio data in Unicode is also available online. Such online data can be mined, refined and used for ASR system development. A domain-wise Sanskrit text from the contemporary usage can be identified for audio recording from different people.

METHODOLOGY FOR RESOUURCE CREATION

Collecting data for training and evaluation of ASR system for Sanskrit could be challenging due to the availability of limited spoken data. Some options for collecting data for a Sanskrit ASR system may include:

Transcription: One way to collect data is to transcribe a large number of audio recordings of spoken Sanskrit manually. This can is time-consuming and labor-intensive, but the data will be of high-quality and good for linguistic researches. This data is also very useful for developing an ASR system.

Crowdsourcing: Another method is through crowdsourcing platforms where multiple people can remotely make simultaneous audio recording of the running Sanskrit text. This is faster and more cost-effective than manual transcription and data quality is pretty good.

Pre-transcribed data: There are some pre-transcribed audio datasets available online that include spoken Sanskrit, such as the Vakyansh and Vāksañcayaḥ[3] dataset. These datasets are very useful for training and evaluating an ASR system, but they may not always be representative of the type of audio the system will encounter in real-world use.

Data augmentation: Data augmentation techniques can also be used to increase the amount of training data available for a Sanskrit ASR system. This may artificially generate new data from the existing data, for example by editing noise or applying other transformations to the audio.

Real-time data collection: An ASR system for Sanskrit can also be trained and evaluated using data collected in real-time from users interacting with the system. This can help ensuring that the system is accurate and relevant for the task it is being used for.

[3] https://www.cse.iitb.ac.in/~asr/

DATA COLLECTION AND NATURE OF DATA

The data we have prepared for Sanskrit speech recognition –are in the form of text data, audio data and transcribed data.

Text Data: In text data, we have collected plain Sanskrit text from Sanskrit Wikipedia[4] with the help of web crawler. This includes 12 thousand articles from Wikipedia. This text has been segmented into sentences and recorded from 150 different speakers. 50 hours of speech data is collected through recording of 30,000 sentences.

Audio Data: Sanskrit Audio data is downloaded from YouTube and Sanskrit Bharti[5] website which has many videos in conversational Sanskrit. Audio has been extracted from all these videos for transcription and then annotated with the help of PRAAT.

Transcribed Data: Parallel speech and text data (transcribed data) of the news that is broadcasted daily and uploaded on the website of AIR[6] (All India Radio), which can be easily downloaded. There are 40 audio files collected for the transcribed data, each audio is approximately 30 minutes long. There is about 20 hours of data collected from All India Radio website ready for transcription and annotation.

DATA PREPROCESSING

Data pre-processing is an important step in the development of any Automatic Speech Recognition (ASR) system. Some specific considerations for pre-processing of data for ASR includes:

Noise reduction: Audio data for an ASR system may contain background noise or other distractions that can interfere with the performance of the system. Noise reduction techniques is used to remove or reduce unwanted noise from the audio background.

Audio segmentation: Audio data is needed to be divided into smaller segments, such as individual words or sentences, in order to be used effectively in training and evaluating the ASR system.

Feature extraction: Proper formatting of audio data in desired format is necessary to serve as an input to the ASR system. This often involves thorough feature extraction and feature selection according to the toll in use. Mel-Frequency Cepstral Coefficients (MFCCs) or spectrograms from the audio are among few very important features that is required for ARS system training.

[4] https://sa.wikipedia.org/wiki/%E0%A4%AE%E0%A5%81%E0%A4%96%E0%A5%8D%E0%A4%AF%E0%A4%AA%E0%A5%83%E0%A4%B7%E0%A5%8D%E0%A4%A0%E0%A4%AE%E0%A5%8D
[5] https://samskritabharati.in/videos
[6] https://newsonair.gov.in/regional-audio.aspx

Data balancing: It is important to have a balanced training data. This means that it must be representative in terms of ratio of examples from each class or category. This will help to prevent the ASR system from being biased towards certain classes.

Script: Sanskrit, a few centuries ago was written in a different script and presently it is following Devanagari. Additionally web based data often extracts many neighboring foreign data from the website which is written in any script other than Sanskrit. Therefore, It is necessary to pre-process the data and ensure that it is properly encoded and readable by the ASR system.

Language-specific considerations: There may be other language-specific considerations that need to be taken into account when pre-processing the data for Sanskrit. For example, system should be able to handle inflections, punctuations, plurals, sentence segmentations, and other grammatical changes that are specific to Sanskrit.

SANSKRIT DATABASE PREPARATION

Some general steps that are involved in preparing a database for a Sanskrit ASR system. That includes:

Data Collection: The first step in preparing a database is data collection. sufficient amount language of data is required for any system development. This involves transcribing audio recordings manually, using crowdsourcing platforms to have multiple people transcribe the same audio, or using pre-transcribed audio datasets such as Vakyansh and Vāksañcayaḥ. The collected data is then cleaned, formatted and prepared for its training and evaluation. Data cleaning or Corpus cleaning involves steps such as noise reduction, audio segmentation, feature extraction etc. The data is organized and stored in a way that is easy to access.

Corpus Validation: Data is needed to be checked for errors or inconsistencies. Formatting and other typographic are corrected. This may involve manual quality control checks or automated tools to identify and fix errors.

Annotating data: Validated corpus if then annotated or labeled in order to be used effectively for training the system. This also includes metadata information about the speaker like language being spoken, or the content of the audio.

Data Splitting: Data is then splitted into different sets, such as a training set, a validation set, and a test set following the standards of the module the data will be undergoing to. These sets will be used at different stages in the development of the ASR system.

CONCLUSION

Collecting data for automatic speech recognition (ASR) systems in Sanskrit presents new challenges and considerations. Sanskrit is a complex language with a rich history and a large body of literature, and ASR systems developed for this language must be able to accurately recognize

and transcribe a wide range of accents, dialects, and textual styles. Several potential methods for collecting ASR data for Sanskrit is used under this research namely voice recordings, crowdsourcing, and collecting data from well-known Sanskrit speaking speakers and web crawling.

A high quality speech data is required for the development of Sanskrit ASR with balanced corpus and representative training and test sets that are gives better accuracy/system performance and makes the model more reliable. Best data collection practices are followed while carefully considering the specific needs of the ASR system and target audience.

FUTURE WORK

Present work involves selecting a suitable ASR model architecture by employing a training set for system training, validation set to tune the model's hyperparameters and then evaluate the model's performance on the test set. Once the model is trained and evaluated, it will be deployed for use in a real-world application based on its performance and scope of improvements.

REFERENCES:

Jha, G. N. (2010). *Sanskrit Computational Linguistics* (Vol. 6465).

Bharati, A., Chaitanya, V., & Sangal, R. (2010). *Natural Language Processing*. New Delhi: PHI Learning Private Limited.

Bhattacharya, K. (2006). On the Language of Navya-Nyāya: an Experiment with Precision through a Natural Language. *The Journal of Indian Philosophy, IIIIV*(1/2), 5-13.

Bhattacharya, S. (1990). Some Features of the Technical Language of Navya-Nyāya. *Philosophy East and West, XX*(2), 129-149.

Brigg, R. (1985). Knowledge Representation in Sanskrit and Artificial Intelligence. *AI Magazine*, 32-39.

Deshpande, M. (1991). Prototypes in Pāṇinian Syntax. *Journal of the American Oriental Society, III*(3), 465-480.

Huet, G. (2016). Sanskrit signs and Pāṇinian Scripts. *Sanskrit Computational Linguistics,* 53-76. New Delhi: D.K. Publishers.

Kadvany, J. (2016). Pāṇini's Grammar and Modern Computation. *History and Philosophy of Logic*, 325-346.

Kiparsky, P. (2002). On the Architecture of Pāṇini's Grammar. Hydrabad: Central Institute of English and Foreign Languages.

Kuhn, T. (2013). A Principled Approach to Grammars for Controlled Natural Languages and Predictive Editors. *Journal of Logic, Language, and Information*,*XXII*(3), 33-70.

Kuhn, T. (2014). A Survey and Classification of Controlled Natural Languages. *Computational Linguistics*,*XL*(1).

Parr, T. (2013). *The Definitive ANTLR 4 Reference*. Texas: Pragmatic Bookshelf.

Sharma, R. N. (2002). *The AD of Panini* (2nd ed. Vol. 1). New Delhi: MunshiramManoharlal Publishers.

Williams, M. (1872). *A Sanskṛit-English dictionary etymologically and philologically arranged: With special reference to Greek, Latin, Gothic, German, Anglo-Saxon, and other cognate Indo-European languages*.Oxford: The Clarendon Press.