Grenze International Journal of Engineering and Technology, Jan Issue

Role of NLP for Corpus Development of Endangered Languages

Sushree Sangita Mohanty[1], Shantipriya Parida[2] and Satya Ranjan Dash[3]
[1]KISS Deemed to be University, Bhubaneshwar, India
Email: sushree.s.mohanty@kiss.ac.in
[2]Silo AI, Finland
Email: shantipriya.parida@gmail.com
[3]KIIT Deemed to be University, India
Email: sdashfca@kiit.ac.in
ORCID : 0000-0002-7902-1183

***Abstract*—Language plays a crucial role in preserving culture in our society and acts as a repository of traditional knowledge, memories, values, practices and unique worldviews that have been used, transformed and practiced in the form of language since millennia. It is estimated that there are more than 6000 spoken languages today that are in danger. UNESCO regularly publishes a list of endangered languages, based on a 5-scale classification system such as: vulnerable, definitely endangered, severely endangered, critically endangered & extinct. It is worrying to know that most of Odisha's indigenous languages are coming under these above factors. Safeguarding and reinvigorating these languages has become crucial for maintaining and preserving the cultural diversity of the society. On account of this, the present paper discusses how Natural Language Processing (NLP), in collaboration with linguistics, can help to revive endangered languages by developing a methodology to build a corpus for the lesser-known endangered indigenous languages of Odisha, some of which have no existing script. The purpose of this paper is to serve researchers/professionals working on low resource languages with a complete guideline for classifying languages and collecting corpus considering language diversity, style and achievable issues.**

***Index Terms*— endangered languages, language classification, corpus collection, natural language processing.**

I. INTRODUCTION

Language is the vehicle of our culture, yet 43 percent of the global languages are endangered. Language plays a crucial role in preserving culture in our society and acts as a repository of traditional knowledge, memories, values, practices and unique worldviews that have been used, transformed and practised in the form of language since millennia. It is estimated that there are more than 6000 spoken languages today that are in danger. To make matters worse, there is no conclusive count that has any status as a scientific discovery of living languages [1]. The reasons are not that some isolated part has not been studied linguistically nor explored properly but rather the notion of enumerating languages is not that simple, it's a lot more complex affair than it appears.
UNESCO offers a classification system to show how languages are 'in trouble'. It has categorised 5-scale classification system such as: vulnerable, definitely endangered, severely endangered, critically endangered &

Grenze ID: 01.GIJET.9.1.709

extinct. According to UNESCO, when a child no longer learns the language as mother tongue at home then the language becomes “definitely endangered”. Similarly, when a language is present at the spoken level in the grandparent generation and understanding level at the parental generation but the parents are not speaking the language to the children, then the language become “severely endangered”. When the youngest speakers are grandparents who have little or infrequent use of that language it defines the language as being in “critically endangered” state. And when there are no speakers left, the language becomes “extinct”.

History says, languages have shifted naturally and declined in dormancy. But, the present situation of language loss is beyond its "natural" pace. Austin & Salabank, renowned linguists predicted that 50% to 90% of languages will be severely/critically endangered by the end of this period 2011[2].

This precipitation of language endangerment be indebted mostly to cultural, political, and economic marginalization and the rise of global expansionism/colonialism/imperialism. Across the world, indigenous peoples have suffered from the colonization and have abandoned their mother tongue in favour of another language. In order to attain a higher social status, indigenous/tribal peoples have had to accept the linguistic norms of the colonists. As per Laderfoged [3] we cannot/should not blame indigenous/tribal people for abandoning their languages in order to secure a better life under intense socio-economic pressures. Rather as a scientist it is our prime duty or responsibility to develop certain mechanism which can be helpful for this power imbalance nature of the society and create a space for the indigenous languages. Besides, a loss of language means loss of identity, culture, knowledge and memory [4].

Study shows, little known/endangered languages are even less signified in the Natural Language Processing (NLP) literature. Sometimes it also merged with many dominant languages. Take an instance of Indian languages, as we all know, India is known for its multicultural, multi-ethnic and multi-lingual mosaic. A place often called by the linguistic experts as the heaven for pluralism. However, it is important to mention here that India both in pre and post independent era has no or partial record on language details based on two formal Linguistic Survey. Sir Abraham Grierson’s project ‘Linguistic Survey of India’, for instance, was conducted for 30 years from 1898 to 1927 and published the data in 11 different volumes. Similarly, in 2010 ‘People’s Linguistic Survey’ was lunched to update the existing knowledge about the spoken languages of India. This survey was led by noted linguistic Ganesh N. Devy and the data was published in December 2012 in 50 different volumes.

Apart from this, 2011 census has also revealed that 123 languages are present in India grouped in 2 categories [5]. The first category is schedule languages and the second category is non-schedule languages, which has 22 and 100 languages respectively. A staggering 96.71% of the speakers are from the Schedule language category, whereas 3.29% speakers are from the non-schedule language category. It is important to note here that, these 3.29% speak a whole lot of 100 types of languages with numerous varieties of dialects given in Fig 1.

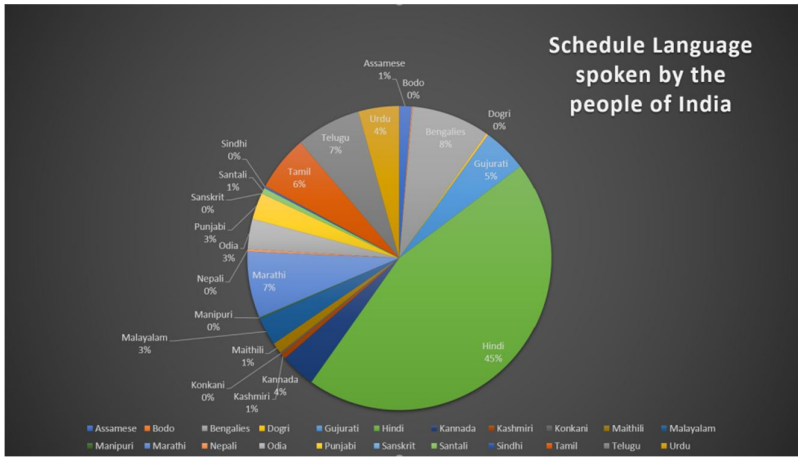

Fig 1: Shows the percentage of people speaks schedule language in India

II. Related Works

According to a theoretical linguist Prof. Ayesha Kidwai, Hindi has emerged as the dominant language spoken by the greatest number of people (4,363 out of every 10,000)[8]. She added that the truths uttered by the Census data are a massive fiction. Hindi is a major spoken language only because the census classification groups 56

other languages plus a category of 'others' under Hindi. It includes many dialects/mother tongue such as; Awadhi, Bagheli/Baghelkhandi, Bagri Rajasthani, Banjari, Bhadrawahi, Bharmauri/Gaddi, Bhojpuri, Brajbhasha, Bundeli/Bundelkhandi, Chambeali, Chhattisgarhi, Churahi, Dhundhari, Garhwali, Gojri, Harauti, Haryanvi, Hindi, Jaunsari, Khairari, Khari Boli, Khortha/Khotta, Kulvi, Kumauni, Kurmali Thar, Labani, Lamani/Lambadi, Laria, Lodhi, Magahi/Magadhi, Malvi, Mandeali, Marwari, Mewari, Mewati, Nagpuria, Nimadi, Pahari, Panchpargania, Pangwali, Pawari/Powari, Rajasthani, Sadan/Sadri, Sirmauri, Sondwari, Sugali, Surgujia, Surjapuri; and many others.

Likewise, many indigenous languages are clubbed under major dominant language as the example given above. According to Joshi more than 88% of the world languages are neglected which is spoken by around 1.2 billion people, that mean the language technology aspect is overlooked or ignored these speakers [6]. Therefore, it is high time to develop linguistic NLP tasks for example morphology analysis which is more inclusive than machine translation NLP tasks as explained by Blasi [7]. Today, as we all are witnessing, is the age of Digital Word. The traditional ways of communication or communicating ideas with each other are becoming out-dated and digital culture mostly the NLP techniques has taken over on it. People are using it widely on the internet. Research says, most of the internet content that we are exposed to every day is processed or generated by NLP techniques. However, when we discuss about the low source language / little known language / endangered language the digital interface has failed to address these languages as a result the enraged languages are facing sever challenges for its survival. On the other hand, most of the NLP researches are biased towards dominant/high sourced languages and neglected the diverse forms of dialectical variations and often depends upon the availability of large scale data.

III. Our Approach

Therefore, the present work is trying to address these above said gaps by developing a model where we are adopting a 6 layered data gathering process with a complete guideline for classifying languages and collecting corpus considering language diversity, style and achievable issues given in Fig 2.

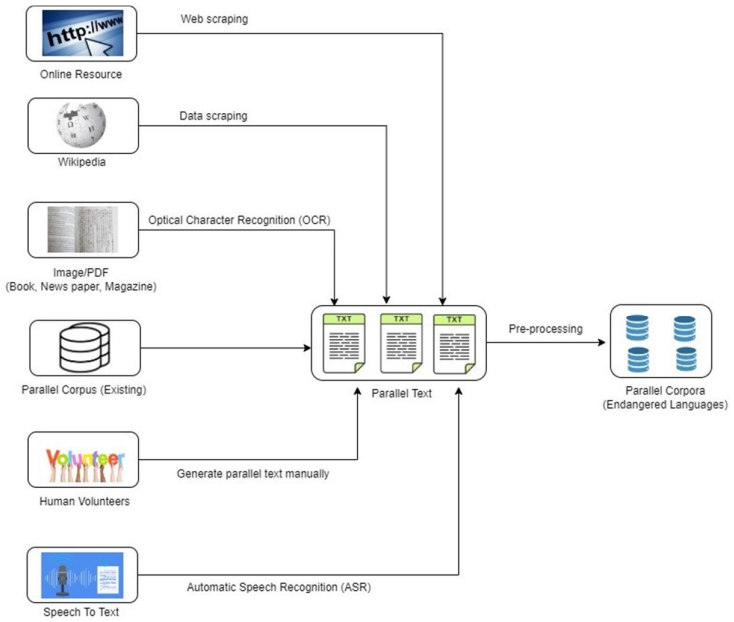

Fig 2: Proposed model to prepare the parallel corpora

A. Web Scraping from Online Resources

Web scraping would be essential for collecting large amounts of data from the online resources and this data could be unstructured and in HTML format. The data from web scraping would be converted into spreadsheet data for using them in several other applications. Web scraping could be performed with several different ways and data can be directly be obtained from websites or web applications. Online services such as APIs can be used for collecting data from web resources [20]. API should be best way to scrape data from web resources;

moreover, Python is most compatible with web scraping as it can handle most processes itself. Python offers Scrapy as popular open-source web crawling framework and it is ideal for data extraction with help of APIs. Beautiful Soup is another python library for web scraping applications; it can create a parse tree so that data can be extracted from HTML [5]. In this model, web scraping would be used as a technique to collect language resource from online web applications or websites. The web scraping data would be structured into parallel text which would be stored as parallel corpora.

B. Data Scraping from Wikipedia

Data Scraping from Wikipedia would be another source for creating text dataset for low-resource and endangered languages. The Wikipedia is considered as online encyclopedia where people keep record of several different items throughout open collaboration and online volunteers. Some researchers could consider Wikipedia as vague source of data considering several people could bring bias into the information or lack of depth could exist. However, Wikipedia is supported with sources or citations that can provide benefit of trust over information. Data scraping for low-resource and endangered language from Wikipedia is useful as community who speaks in those languages can store text data in Wikipedia.

C. Optical Character Recognition

Optical Character Recognition is to be used in this context for collecting data from images. The images of textbook, storybook, magazine, newspaper, article, blogs, and other printed sources would be fed into OCR for extracting data from those images [22]. Use of OCR is highly beneficial in this model as collecting massive data from millions of books and printed sources would be easier. Scanning images would be effortless, automated with easier access to convert the printed sources into machine-readable text data. This technique would be useful for storing data into parallel text dataset.

D. Parallel Corpus

Parallel corpus in this context is existing dataset or corpora available for the low-resource language. Existing parallel corpora is not detailed with low-resources and specific knowledge domain is not available online hence, parallel corpora should be considered as a basic source. Mostly, data should be collected from existing printed resources and from people's inputs.

E. Human Volunteers

Human Volunteers or community should be taken into the data collection phase as they can input text data on their own. Their input would be collected manually from their own inputs. The collected data would be directly added to the parallel text dataset in a suitable format. Human volunteers' input should be validated with experts as mistakes during inputting should be avoided. Manual entry can take more time than machine-generated data inputs therefore; this volunteer-driven data entry should be initiated before web scraping and use of OCR.

F. Automatic Speech Recognition

Automatic Speech Recognition (ASR) should be considered as a technology that can allow users to speak and it would conduct the data-entry to the machine in that specific language. In recent decades, researchers are more interested into automatic speech recognition (ASR) as it is a specific method for communication between people [9]. ASR can be defined as process of collecting data as speech data and converting it into text data. Current studies have shown that there are problems such as background noises, large vocabularies, speech pronunciations, and dialects. Research works are considering audio-visual techniques for making the approach more robust [11]. In medical and educational sectors, ASR is used for conducting research work.

Major challenges in ASR lies in performance of ASR as there exists background noises covering the speech. Many algorithms are used to deal with noisy environments where, these algorithms failed due to presence of noise only. In noisy environment, acoustical features would be degraded in short-frame detection [10]. Noise cancellation or noise reduction methods are developed however; these methods cannot work when noise sources are not known. Diverse noise sources can interrupt the text data collection using ASR.

Overlapping speech or conversation would be another challenge for using ASR; this problem can occur when data is being collected from a group of speakers [12] [8]. ASR faces this difficulty in detecting the target speech. This problem is considered as much critical as others.

Performance of speech capture could influence ASR performance; speech data collection through poor quality microphone can reduce the performance of ASR systems. Poor quality microphone would introduce noises into the speech that would complicate the speech detection and recognition process [13] [18]. Hence, number of microphones should be higher than number of speakers.

Commonly spoken or low-resource languages have dialect influence. Dialect is a huge dilemma for Natural Language Processing as a diverse issue. Diversity of dialect makes the ASR challenging for specifically low-resource languages [15] [17]. Audio-visual models are facing the issue of dialect; as well, it cannot differentiate the continuous speech from dialects.

Performance issues of ASR systems are noticeable when speech content becomes less predictable [16]. Pronunciation difference can exist due to speakers' health condition as well such as stuttering in pronunciation, voice and children's limited speech ability.

Currently, there is limited datasets for ASR research, which is a significant limitation in speech-to-text research domain [19]. Therefore, this model would be considering ASR as suitable source of text data for the dataset. Moreover, the limitations can be avoided with massive collection for preparing parallel corpora of low-resource and endangered languages.

IV. Conclusion

This proposed model can conclude that all primary low-resource language sources are considered for making this parallel corpus. The collection of speech data and convert it into text dataset would be a bigger challenge in this research context. Web scraping, OCR-based text data extraction from online and printed sources respectively would be effortless and automated. These techniques could be faster compared to ASR-based text data collection. Human volunteers can contribute to the parallel corpora in suitable way and their manual data entry would take longer duration. However, the noise reduction in ASR would be primary concern in the implementation. It should be recommended to collect speech data from speakers when they are in confined space so that noise can be avoided. Individual speaker should be considered for collecting their speech to avoid overlapping issue and high-quality microphone should be used to reduce noise as well.

References

[1] Anderson, Stephen R., 2010. How many languages are there in the world? Linguistic Society of America Brochure Series: Frequesntly Asked Questions. Pp-1-6.

[2] Austin, Peter, K. & Salabank Julia. 2011. The Cambrige handbook of endangered languages. Cambridge University Press.

[3] Ladefoged, Peter. 1992. Another view of endangered languages. Language, 68(4). Pp-809-811.

[4] Whalen, Douglas H., Margaret Moss, and Daryl Baldwin. 2016. Healing through language: Positive physical effects of indigenous language use. F1000Research. 5 (852) pp-852.

[5] 2011 Census Data of India https://censusindia.gov.in/census.website/.

[6] Joshi, Pratik., Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp- 6282–6293.

[7] Blasi, B., Antonios Anastasopoulos, and Graham Neubig. 2021. Systematic inequalities in language technology performance across the world's languages. arXiv preprint arXiv:2110.06733.

[8] Khidwai, A. 2019. The people's linguistic survey of India volumes: neither linguistcs, nor a successor to Grierson's LSI, but still a point of reference. Social Change. SAGE publication.

[9] Agrawal, P. and Ganapathy, S., 2019. Modulation filter learning using deep variational networks for robust speech recognition. IEEE journal of selected topics in signal processing, 13(2), pp.244-253.

[10] Cai, C., Xu, Y., Ke, D. and Su, K., 2015. A fast learning method for multilayer perceptrons in automatic speech recognition systems. Journal of Robotics, 2015.

[11] Chen, Z., Droppo, J., Li, J. and Xiong, W., 2017. Progressive joint modeling in unsupervised single-channel overlapped speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26(1), pp.184-196.

[12] Ganapathy, S., 2017. Multivariate autoregressive spectrogram modeling for noisy speech recognition. IEEE signal processing letters, 24(9), pp.1373-1377.

[13] Gosztolya, G. and Grósz, T., 2016. Domain adaptation of deep neural networks for automatic speech recognition via wireless sensors. Journal of Electrical Engineering, 67(2), p.124.

[14] Hirayama, N., Yoshino, K., Itoyama, K., Mori, S. and Okuno, H.G., 2015. Automatic speech recognition for mixed dialect utterances by mixing dialect language models. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23(2), pp.373-382.

[15] Kumar, M., Kim, S.H., Lord, C., Lyon, T.D. and Narayanan, S., 2020. Leveraging linguistic context in dyadic interactions to improve automatic speech recognition for children. Computer speech & language, 63, p.101101.

[16] Lee, H.Y., Cho, J.W., Kim, M. and Park, H.M., 2016. DNN-based feature enhancement using DOA-constrained ICA for robust speech recognition. IEEE Signal Processing Letters, 23(8), pp.1091-1095.

[17] Lee, S.C., Wang, J.F. and Chen, M.H., 2018. Threshold-based noise detection and reduction for automatic speech recognition system in human-robot interactions. Sensors, 18(7), p.2068.

[18] Ming, J. and Crookes, D., 2017. Speech enhancement based on full-sentence correlation and clean speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(3), pp.531-543.
[19] Wang, D., Wang, X. and Lv, S., 2019. An overview of end-to-end automatic speech recognition. Symmetry, 11(8), p.1018.
[20] Cruz, J.C.B. and Cheng, C., 2020. Establishing baselines for text classification in low-resource languages. arXiv preprint arXiv:2005.02068.
[21] Linder, L., Jungo, M., Hennebert, J., Musat, C. and Fischer, A., 2019. Automatic creation of text corpora for low-resource languages from the internet: The case of swiss german. arXiv preprint arXiv:1912.00159.
[22] Chaudhuri, A., Mandaviya, K., Badelia, P. and Ghosh, S.K., 2017. Optical character recognition systems. In Optical Character Recognition Systems for Different Languages with Soft Computing (pp. 9-41). Springer, Cham.