Creating Corpus of Low Resource Indian Languages for Natural Language Processing: Challenges and Opportunities

Pratibha Dongare
The English and Foreign Languages University, Hyderabad, India.
pratibhaphdlandp22@efluniversity.ac.in

Abstract

Addressing tasks in Natural Language Processing requires access to sufficient and high-quality data. However, working with languages that have limited resources poses a significant challenge due to the absence of established methodologies, frameworks, and collaborative efforts. This paper intends to briefly outline the challenges associated with standardization in data creation, focusing on Indian languages, which are often categorized as low resource languages. Additionally, potential solutions and the importance of standardized procedures for low-resource language data are proposed. Furthermore, the critical role of standardized protocols in corpus creation and their impact on research is highlighted. Lastly, this paper concludes by defining what constitutes a corpus.

Keywords: Low resource languages, Corpus, Indian Languages

1. Introduction

Natural Language Processing (NLP) has witnessed unprecedented growth and advancements in machine learning, artificial intelligence and other allied fields. However, while NLP models have flourished in well-resourced languages, the landscape becomes markedly challenging when operating within low-resource language domains. Data plays a crucial role in any NLP task. The quality and quantity of the data has a huge impact on the performance of a system. The type of corpus may vary according to the tasks. For instance, spoken, textual, conversational, lexical, learner, and other types of corpora can be used while working on TTS, ASR, information extraction tasks, discourse corpus or conversational corpus can be used in creating chatbots or training LLMs, parallel corpus can be used in Machine translation.

India is a diverse country with many languages, but it lacks the necessary resources to adequately support even the most widely spoken Indian languages. When considering Asia as a whole, which is linguistically dense, similar challenges arise in representing these languages computationally. The absence of fundamental NLP tools for these languages has significant social implications (Singh, 2008).

Low-resource languages, commonly characterized by limited availability of linguistic data and tools, pose unique obstacles in developing effective NLP solutions. Low-resource languages are also known as less privileged languages (Singh, 2008), less advanced languages (Dash and Ramamoorthy, 2019), under-resourced, and resource-poor languages. Low resource languages can be understood as less studied, resource scarce, less computerized, less privileged, less commonly taught or low density among other denominations (as cited in Maguersse et al., 2020; Singh, 2008; Cieri et al., 2016; Tsvetkov, 2017).

When examining the definition of the concept, it becomes evident why Indian languages are classified as low-resource languages. Atkins et al. (1992) outlined several challenges experienced by languages with varying levels of development, from less advanced to more advanced languages. When exploring the lack of resources in languages, various factors come into play:

- **Digital presence:** Digital representation encompasses the online presence of a language, including its information in various domains, subjects, and diverse forms of data. It is crucial to gauge the extent of this data to address the ongoing debate about what constitutes an adequate amount of information.
- **User friendliness:** It is crucial to achieve a harmony between usability and linguistic representation. According to the findings of the KPMG- Google (2017) survey, it is projected that by 2021, 8 out of 10 Indian language users will access the Internet in regional Indian languages. This development significantly influences the accessibility and user-friendliness of these

languages. Therefore, it is imperative to invest in the development of language resources and NLP tools for low-resource languages to ensure equal opportunities and benefits for all linguistic communities.

- **Language Processing and Tools:** Enabling the development of language processing tools becomes feasible with increased availability of data. The foundation of NLP research relies on the presence of corpora; structured datasets (written, spoken, multimodal, etc.) carefully selected for training and evaluating language models. Corpora form the basis for various NLP applications such as machine translation, sentiment analysis, and information extraction. However, the creation and standardization of corpora poses complex challenges, particularly in languages with limited resources.

Corpus-based studies are incorporating new insights to investigate the cognitive areas of the human mind to understand the mysteries operating behind the cognitive process like receiving, processing, comprehending, and sharing linguistic signals (Winograd, 1983). Corpus can be used in wide applications. For instance, domains of social sciences, machine learning, sentiment analysis, dictionary compilation, grammar writing, wordnet design, word-sense disambiguation, translation, documentation, and other areas of linguistics like diachronic lexical semantics, pragmatic analysis of texts, sociolinguistic studies, and discourse analysis (Dash and Arulmozi, 2018; Leech and Fligestone, 1992).

Compared to other countries, India lags far behind not only in corpus generation but also in corpus-based linguistic studies and application Dash and Ramamoorthy (2019). The next section briefly outlines the challenges faced while working on resource-poor languages.

2. Challenges

Creating and compiling the corpus presents numerous challenges, some of which are briefly outlined in this section.

1. **Data scarcity:** As mentioned in the previous section, a significant challenge for Indian languages is the limited availability of resources, including corpora and tools. Despite an increase in internet and technology users, there has been no corresponding increase in resources for regional languages. Therefore, the development of resources for these languages presents a significant challenge. The accessibility of various domains and topics is also extremely important. The lack of diverse and representative data in the corpora for regional languages is another challenge. For example, there are several freely available datasets for download and use, such as those developed AI4 Bharat[1], datasets available on TDIL[2] and LDCIL[3] portals. However, these datasets primarily emphasize the scheduled languages and cover a limited range of domains like news articles. Furthermore, the lack of standardization and documentation poses difficulties in corpus compilation.
2. **Quantity of data:** Determining the appropriate amount of data is an important yet debatable question when it comes to building a corpus. The emergence of GPT models has recently generated considerable interest in large datasets, but gathering extensive data for languages with limited digital presence remains a challenge. Dash and Ramamoorthy (2019) have emphasized that the distribution of written and published texts is uneven, posing a challenge for corpus compilation. While Sinclair (19991) stated that containing around 1 million words may be sufficient for specific linguistic studies and research, Dash and Ramamoorthy argue that at least 10 million words are necessary for language description purposes.

3. **Linguistic and non-linguistic challenges:** Numerous Indian languages exhibit diverse varieties and dialects, with a significant number of speakers. Hence, it is imperative to address the need to support these various dialects and linguistic features. Furthermore, consideration must also be given to the shared linguistic attributes among these languages, their scripts, and variations. Additionally, certain tribal languages lack scripts altogether, necessitating representation using alternative available scripts while some languages use multiple scripts. For instance, Korku and Munda languages (languages belong to Austro Asiatic family of languages) have no regular scripts whereas Santali (one of the scheduled languages of India) uses five scripts:

[1] https://ai4bharat.iitm.ac.in/resources/datasets/

[2] https://tdildc.in/index.php?option=com_download&task=fsearch&Itemid=547&lang=en

[3] https://data.ldcil.org/index.php?route=common/home

Devanagari, Bengali, Odia, Alchiki and Roman scripts (census of. India, 2022).

4. **Code mixing:** The growing trend of code mixing, in which individuals alternate between two or more languages within a single conversation or sentence, presents challenges in compiling corpora and analyzing language. The use of script mixing (romanization) poses significant difficulties when dealing with data from social media platforms. Transliteration between roman and regional scripts, as well as glossing and annotation, becomes essential to account for code mixing during corpus development. For instance, the sentence, *'मी try करेन' (Mi try karen) (I will try)* uses Marathi and English with both Devanagari and roman script, requiring transliteration and annotation to accurately represent code mixing.

5. **Computational assistance:** Data must be represented, stored, and managed using computational devices. Advances in hardware and software technologies have facilitated data optimization. However, limited knowledge of these technologies and the affordability of such devices pose challenges. It is important for a wide range of researchers to have access to the latest hardware systems, updated software versions, and operating systems. For example, the availability of support for the OCR technique is also limited, which affects the digitization of many regional language texts. For example, old Marathi texts and manuscripts are written using a script called moḍi. Although this script is not commonly used today, it still holds significance in facilitating computational assistance for preserving languages and conducting diachronic research.

6. **Standardization challenges:** Supporting LR languages presents significant challenges such as transcription, transliteration, glossing, and encoding. Using a widely accepted standard such as Unicode facilitates consistent data representation, ensuring accurate display and processing across different software and hardware platforms. The absence of such standardization complicates the creation of a reliable corpus, which requires additional conversion efforts to integrate data from diverse sources into the Unicode-based corpus. Indic languages typically utilize 8-bit fonts for encoding. However, despite the existence of a standard 8-bit code table and layout for Devanagari in ISCII, varying keyboard layouts and non-standard character sets employed by font designers contribute to difficulties in standardization when gathering data from multiple sources (McEnery et al., 2000).

7. **Data revision and updates:** The data must undergo regular revisions and updates to ensure the relevance and accuracy of the information. This is particularly important for LR languages, as acquiring the initial data poses a significant challenge. Consequently, maintaining data quality and implementing updates presents an even greater challenge. For instance, machine learning models heavily rely on training data, regular updates are necessary to adapt to evolving language patterns and improve performance.

8. **Ethical considerations:** When conducting research on lesser-represented languages, particularly involving speech corpus, it is crucial to prioritize ethical considerations. This applies not only to multimodal data, but also when working with smaller language communities and non-mobile populations. It is essential to take steps to ethically collect and document high-quality data in these cases.

3. Potential Solutions

Working with LR languages presents various challenges, and the following section emphasizes the importance of standardization while offering potential solutions to these obstacles.

- In order to create uniformity in generation, compilation, and maintenance of the corpus, we need a standardized procedure or common guidelines. For instance, in Baker et al. (2003) mentioned that in their study, ISCII was an attempt to standardize 8-bit encodings for Indian writing systems, but the paper notes that this standard is largely ignored by developers of TTF fonts for Indic scripts and so is mostly absent from the web. This leads to a significant challenge in corpus creation, as many different incompatible glyph encodings exist for Indic fonts compared to a standardized approach, like the hexadecimal code 42 always representing "B" in English fonts. ASCII is a character encoding standard for electronic communication that represents text in computers, and UTF-8 is a variable width character encoding that can represent all characters in the Unicode character set. Both play crucial roles in text processing and data exchange, with UTF-8 being particularly important as a way to encode Unicode characters efficiently while preserving backward compatibility with ASCII. Baker et al (2003 a) conducted research on the EMILLE corpus. They emphasized that transforming numerous 8-bit based texts into a uniform format such as Unicode was challenging and time-consuming, mainly because of the absence of consistent 8-bit font encoding standards across various creators of electronic texts in the respective languages. This proved to be a substantial technical obstacle in compiling the corpora. In the proposed solution, McEnery and colleagues (2000) used a 16-bit universal character set.

- The standardized process used in creating and managing the corpus, along with encoding, will assist linguists and annotators by providing a clear framework for collaboration. This will facilitate the development of consistent guidelines for data annotation, preprocessing, and analysis to ensure high-quality results.
- Working with a standardized approach for low-resource languages is crucial as it would not only support the computational advancement of these languages but also enable more widespread contributions. Additionally, this approach would make it possible to have a comparative analysis of the data, leading to valuable insights and progress in linguistic research.
- **Interoperability:** Interoperability can be improved through the establishment of standard and uniform procedures. This would lead to better data transfer and usage, benefiting researchers worldwide. Moreover, this improvement in interoperability would ensure greater convenience regardless of costly hardware or software upgrades. Furthermore, adopting a standardized approach in text processing and data exchange would promote accessibility and inclusivity.
- **Quality of data:** When creating a corpus, it is crucial to take into account different linguistic and statistical factors like the size of the data, its manner, intended users or tool usage (in relation to task- specific and domain-specific tools), multilingual and monolingual data, preprocessing and cleaning procedures, as well as data storage and management. Ethical considerations are necessary to ensure the authenticity of the data by obtaining prior consent from participants or informants. It is important to also consider potential biases in the collection process that might affect the overall quality of the corpus.
- **Collaborative efforts:** Collaboration and contribution from researchers with diverse expertise in linguistic, statistical, and computational fields are essential for the development and advancement of LR languages. Through their combined efforts, we can achieve more accurate, dynamic, and impactful results. Advancements in the field will be achieved through community efforts.

4. Conclusion

The paper aimed to briefly outline the practical obstacles encountered when working with Indian language corpora. The compatibility, accessibility, and interoperability of the data can be improved using the standard practices and efforts. The potential solutions could be improved. It is necessary to consider multidimensional and multilingual corpus development, which can have applications in various related fields such as language description, comparative analysis, documentation, tool development, and more. A corpus is not simply a collection of data; rather, it is a curated and processed collection of information tailored for specific research purposes because, while gathering data may not be challenging, transforming it into a corpus is.

5. Bibliographical References

Atkins, S., Clear, J., & Ostler, N. (1992). Corpus design criteria. *Literary and linguistic computing*.

Baker, P., Hardie, A., McEnery, T., & Jayaram, B. D. (2003a). Corpus data for South Asian language processing. In *Proceedings of the 10th Annual Workshop for South Asian Language Processing, EACL*.

Baker, P., Hardie, A., McEnery, T., & Jayaram, B. D. (2003a). Constructing corpora of South Asian languages. https://ucrel.lancs.ac.uk/publications/cL2003/papers/baker.pdf

Census of India. (2022). Census of India 2011—*Language Atlas- India.* Office of the Registrar General & Census Commissioner, India. https://censusindia.gov.in/nada/index.php/catalog/42561

Dash, N. S., & Arulmozi, S. (2018). *History, features, and typology of language corpora*. Springer Singapore.

Dash, N. S., & Ramamoorthy, L. (2019). *Utility and application of language corpora.* Singapore: Springer.

KPMG-Google. (2017). *Indian languages defining India's internet.* https://assets.kpmg/content/dam/kpmg/in/pdf/2017/04/Indian-languages-Defining- Indias-Internet.pdf

Leech, G. and S. Fligestone. (1992). *Computers and corpus analysis*. Oxford: Blackwell publishers.

Magueresse, A., Carles, V., & Heetderks, E. (2020). Low-resource languages: a review of past work and future challenges. *arxiv preprint arxiv:2006.07264*.

McEnery, A., Baker, P., Gaizauskas, R., & Cunningham, H. (2000). EMILLE: Building a corpus of South Asian languages. In *Proceedings of the International Conference on Machine Translation and Multilingual Applications in the new Millennium: MT 2000*.

Sinclair, J. (1991). *Corpus, concordance, collocation.* Oxford University Press.

Singh, A. K. (2008). Natural Language Processing for Less Privileged Languages: Where do we come from? Where are we going?. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*.

Tsvetkov, Y. (2017). Opportunities and challenges in working with low-resource languages.

http://www.cs cmu.edu/~ytsvetko/jsalt-part1.pdf.

Winograd, T. (1983). *Language as a cognitive process.* Vol. I. Mass: Addison-Wesley publication.