Generating Monolingual Dataset for Low Resource Language Bodo from old books using Google Keep

Sanjib Narzary[a], Maharaj Brahma[b], Mwnthai Narzary[c], Gwmsrang Muchahary[d], Pranav Kumar Singh[e], Apurbalal Senapati[f], Sukumar Nandi[g], Bidisha Som[h]

a,b,c,d,e,f - Department of Computer Science and Engineering, Central Institute of Technology Kokrajhar, Kokrajhar, India

g,h - Center for Linguistic Science & Technology, Indian Institute of Technology Guwahati, Guwahati, India

{san, p20cse1012, p20cse1001, p20cse1011, p.singh, a.senapati}@cit.ac.in, {sukumar, bidishar}@iitg.ac.in

1. Abstract

Bodo is a scheduled Indian language spoken largely by the Bodo people of Assam and other northeastern Indian states. Due to a lack of resources, it is difficult for young languages to communicate more effectively with the rest of the world. This leads to a lack of research in low-resource languages. The creation of a dataset is a tedious and costly process, particularly for languages with no participatory research. This is more visible for languages that are young and have recently adopted standard writing scripts. In this paper, we present a methodology for generation of a monolingual Bodo corpus from different books using free and easily accessible resources like Google Keep. We generated a Bodo text corpus of 192,327 tokens and 32,268 unique tokens. Moreover, some essential characteristics of the Bodo language are discussed that are mostly neglected by Natural Language Processing (NLP) researchers.

2. Introduction

Bodo is one of the Indic languages and belongs to the Sino-Tibetan language family, one of the four language families widely spoken in India. According to the 2011 Census[1][a], it has nearly a million speakers. It is primarily spoken by the Bodo tribe in Indian state of Assam, as well as tribes such as the Kachari, Mech, and others. There are 1,454,547 native Bodo speakers and total of 1,482,929 Bodo speakers[2]. It accounts for 0.12 % of India's overall population and ranks at 21 out of other 22 scheduled languages. The number of Bodo speakers is rising steadily. Historically, Bodo has a rich oral tradition but no standard script for writing, and until recently, the Devanagari script is officially adopted. The Natural Language Processing (NLP) literature on Bodo language is relatively small, and the majority of research has been carried out just recently. This can be attributed to the following reasons: (a) the youngness of the language, (b) low data availability, (c) low NLP research interest, (d) unavailability of preliminary studies and benchmarks, and (e) lack of technical resources.

[a] Survey conducted by Government of India

3. Related Work

In the work [3], a Bodo corpus containing more than 1.5 million words, The resultant corpus contained a total of 1,577,750 Bodo words from three categories: learned materials, media, and literature.

8. References

[1] Census. Abstract of speakers' strength of languages and mother tongues. *Census 2011*, 2011.

[2] Census. Comparative speakers' strength of languages and mother tongues - 1971, 1981, 1991, 2001 and 2011. *Census of India 2011*, 2011.

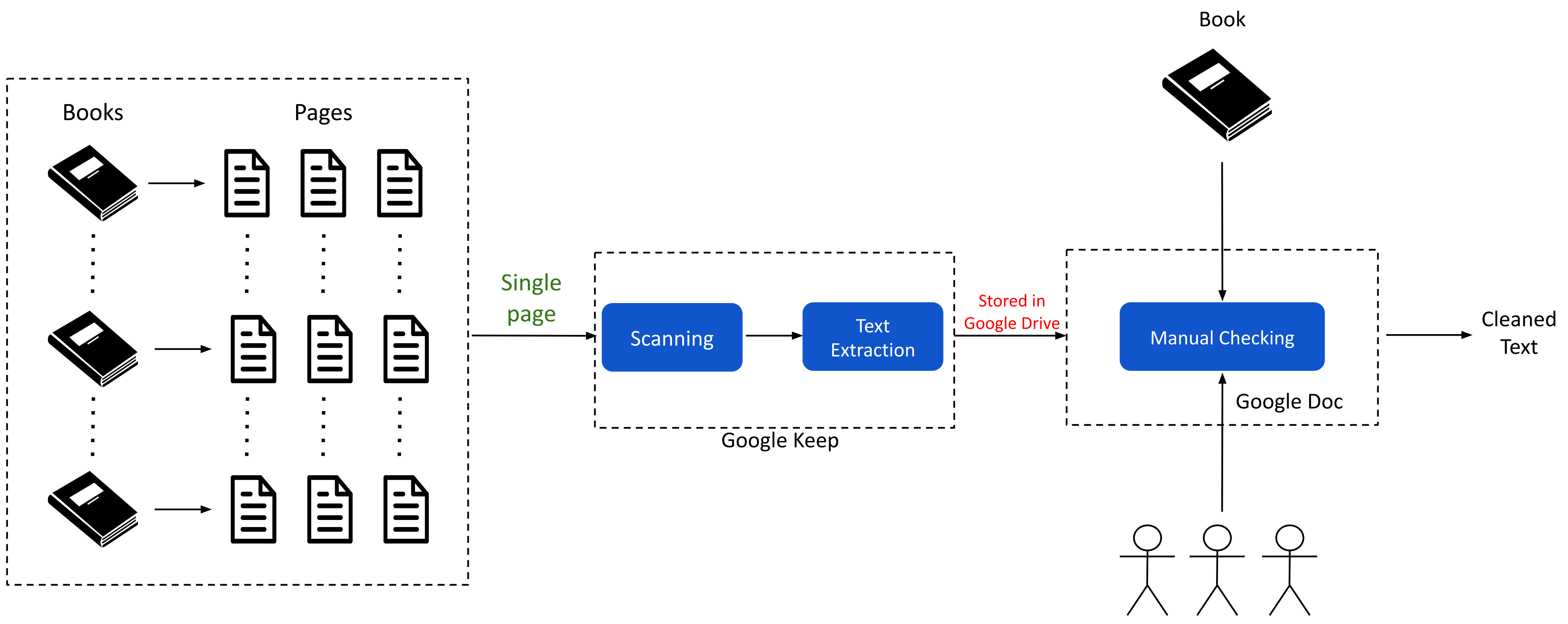
[3] Biswajit Brahma, Anup Kr. Barman, Shikhar Kr. Sarma, and Bhatima Boro. Corpus building of literary lesser rich language-Bodo: Insights and challenges. In *Proceedings of the 10th Workshop on Asian Language Resources*, pages 29–34, Mumbai, India, December 2012. The COLING 2012 Organizing Committee.

4. Problem Statement

Bodo is a relatively young language with a recent standard writing format, the availability of corpora for natural language processing tasks such as machine translation is low. There are no general domain parallel datasets, thus restricting the quality of the translation.

5. Methodology

Our methodology comprises 4 steps: Book Collection, Scanning, Text Extraction, and Manual Cleaning. We collected 13 books from children, fiction, novel, play, biography, and drama genres. Google Keep *"Grab text"* functionality is used for Scanning and Text Extraction. The extracted texts are manually cleaned using Google Docs.



6. Results

The resulted corpus consists of 192,327 tokens, 32,268 unique tokens and 0.16 type-token ratio. We evalute the dataset through manual cross-checking and language modelling.

1. Manual Cross-Checking: To evaluate the correctness of the cleaned text we randomly distributed samples to five different people for cross-checking and asked them to rate the samples between 1-5. Each sample received a rating of 5, suggesting that the quality of the post-extraction cleaning is good.

2. Language Modelling: We trained[a] tokenized version of the Bodo Monolingual Text Corpus ILCI-II[b], a general domain corpus containing 31,026 sentences and 1,029,408 words, on n-grams of 1, 2, and 3, and achieved a perplexity of 3210.6, 38.11, and 3.56 respectively. After training, we evaluated the cleaned text with an n-gram of 3 by randomly sampling paragraphs from the extracted text. We evaluate the extracted text genre-wise, and we achieved a perplexity of 31127.55, 24553.62, 37168.39, and 62051.7 for Fiction, Novel, Play, and Children genre respectively. Results suggest that the existing monolingual corpus still lacks generality of data and that the created corpus can be used to improve the monolingual data for the Bodo language.

[a] Performed using Stanford Research Institute Language Modeling Toolkit http://www.speech.sri.com/projects/srilm/

[b] Provided by TDIL-DC

7. Conclusions

In our work, we show that already existing day-to-day usable applications such as Google Keep, Google Doc, and Google Drive can be used to build monolingual datasets for the Bodo language. This approach of free and easily accessible applicable can substantially make the dataset creation process easy and accessible for other low-resource languages. Often, a language remains a low resource due to a lack of researchers and technical information required in the community to contribute to the corpus creation process. Hence, this process can substantially improve the status of resource scarcity. It is scalable in situations where the language does not have many technical resources or NLP researchers.