Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 6563–6570
Marseille, 20-25 June 2022

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

Abstract

Bodo is a scheduled Indian language spoken largely by the Bodo community 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 using Google Keep for OCR to generate a monolingual Bodo corpus from different books. In this work, a Bodo text corpus of 192,327 tokens and 32,268 unique tokens is generated using free, accessible, and daily-usable applications. Moreover, some essential characteristics of the Bodo language are discussed that are neglected by Natural Language Processing (NLP) researchers.

Keywords: Monolingual Corpus Creation, Low Resource Language, Bodo Language, Vulnerable Language

1. 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 (Census, 2011a)[1], it has nearly a million speakers. It is primarily spoken by the Bodo tribe in 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 (Census, 2011b) as shown in Table 1. It accounts for 0.12 % of India's overall population and ranks at 21 out of other 22 scheduled languages as shown in Table 2. As in Figure 1, 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.

NLP is an emerging field of study with rising applications in various domains. There is a huge demand for digital communication between Indian languages and the rest of the world. However, the amount of dataset available is significantly less

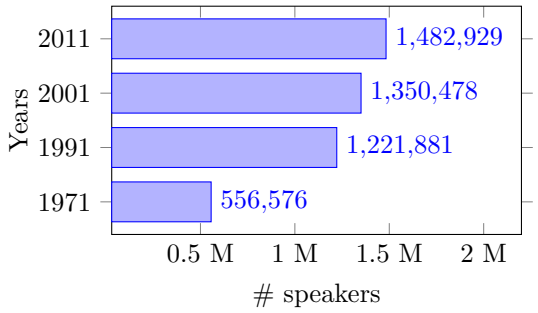

Figure 1: The year wise survey of number of Bodo speakers in millions. The census record for Bodo is not being conducted in year 1981.

Tribe	Number of Speakers
Bodo	1,454,547
Kachari	15,984
Mech/Mechhai	11,546
Others	852

Table 1: Distribution of number of tribes responding Bodo as Mother tongue in 2011 Census conducted by Government of India.

[1]Survey conducted by Government of India

Rank	Language	Percentage to total population
1	Hindi	43.63
2	Bengali	8.03
3	Marathi	6.86
4	Telugu	6.70
19	Konkani	0.19
20	Manipuri	0.15
21	**Bodo**	**0.12**
22	Sanskrit	N

Table 2: Percentage of top 4 and last 4 scheduled language to total population above along with its rank. Here rank represents the position of language in terms of language usage.

than that of other languages. Bodo is a resource limited language with a limited number of datasets publicly available. A good language model is required for Indic languages to have equal access to information and content, particularly in domains such as education, health, entertainment, judicial, etc. Building a dataset is a tedious process, especially for languages with little information or material in literature or on the internet. There also exists technical difficulties, such as the non-existence of updated dictionaries and glossaries. One possible solution to this challenge is to generate a dataset using existing resources such as books, magazines, and newspaper articles rather than generating fresh content. Text extraction from such a resource is a challenging task. As a result, there is a need for a free, accessible, and daily-usable system to build the corpus.

Hence, in this paper, we create a monolingual dataset from the old Bodo books by leveraging existing day-to-day usable applications like Google Docs[2], Google Drive[3], and Google Keep[4]. Google Keep is an OCR for text extraction. Furthermore, we discuss the various issues and challenges in the Bodo language. We considered translated Bodo books from English, motivated by the fact that they can be used to create a parallel corpus.

2. Background

India is a highly multilingual country, with 22 scheduled languages (Joshi et al., 2019) covering about 97% of the population. The Bodo language is one of the scheduled languages of India, spoken largely by the Bodo community. The Bodo community is the second-largest community in the North-Eastern region of India, with centuries of rich cultural history, heritage, and folklores. The Bodo language is one of the prominent languages of Northeast India. It belongs to the Sino-Tibetan language family under the subbranch of the Assam-Burmese group. Linguistically, the language shares some common features with the Dimasa, Garo, and Kokborok languages.

Bodo has a rich oral tradition but with no standard script for writing, although some scholars suggest the existence of the lost script *Deodhai* (Bhattacharya, 1964). After a long history of script movement, in 2003 the Bodo language was recognised by the Government of India, and the Devanagari script was officially adopted. Before the official adoption, various other scripts such as Assamese, Bengali, and Roman were used. The standardisation of written script led to a better-structured expression of the language in written literature. It is currently expanding, with a substantial number of publications created in genres such as novels, biographical works, poetry, children's literature, and fiction.

In the following subsection, we provide existing issues in the Bodo language, which are the reasons for the low resources and less NLP research.

2.1. Youngness of language

Bodo is a language with rich oral literature, spoken mostly by the Bodo community in Assam, Northeast India, and its neighbouring states. It was introduced as a medium of instruction in schools in the year 1963. In 1966, post-graduate courses in the Bodo language were introduced at the state university. It was recognised as a scheduled language of India in 2003, with Devanagari as the official writing script.

Hence, Bodo is relatively young in terms of written literature compared to the other scheduled languages with standard scripts. The written literature has just grown in the last few decades. Despite the growth of literature, it is hampered by the lack of updated dictionaries, spelling variation issues (Brahma et al., 2012), and the inclusion of new words (Narzary et al., 2021).

2.2. Low data availability

Low-resource languages, in particular, experience data unavailability. The unavailability of data, according to (Nekoto et al., 2020), is significantly broader. For Indian languages, the lack of presence of the language usage on the internet also contributes to the data scarcity. This is more severe for languages that are young and have recently adopted standard written scripts. And this is very much visible to Bodo. Currently, Bodo has no available text corpus on well-known platforms used for dataset generation, such as *Wikipedia*. The Wikipedia for Bodo is currently in the incubation

[2] `https://docs.google.com`
[3] `https://drive.google.com`
[4] `https://keep.google.com`

phase[5].
The recent advancement of NLP uses deep neural networks for performing tasks such as machine translation (Sutskever et al., 2014) (Bahdanau et al., 2014) which require large amounts of data. Due to the availability of a dataset, such techniques cannot be exploited fully for the Bodo language.

2.3. Low NLP research interest

Most of the NLP techniques require a huge text corpus to be available, making it difficult for NLP researchers to perform experiments and research. This is particularly visible in case of Bodo language.

2.4. Unavailability of preliminary study and benchmarks

Bodo is a tonal language with two tones: *high* and *low*. The tonal characteristics of a word are not reflected in written text.

Tonal Word	Sentence
जा	आं ओंखाम जाबाय
Meaning - Eat	I ate rice
जा	बियो फोरोंगिरि जाबाय
Meaning - Become	He became teacher

Table 3: Example Bodo tonal word, meaning and its corresponding example sentence.

In Table 3, the word "जा" has two different meanings based on its pronunciation tone. By examining the word, we can't identify its meaning. It is because there is no identification of tone in written form.
Another key issue is the unavailability of publicly available benchmarks for NLP tasks such as NER, machine translation, text classification, etc., leading to low research contributions and a lack of participatory research.

2.5. Lack of technical resources

Due to the lack of study, NLP applications such as language modelling, spelling correction, machine translation, question answering, sentiment classification, etc., have not yet been either studied or fully explored, leading to no production-level applications. Bodo lacks technical resources for corpus development, such as web-based spell checkers and the availability of digitally accessible dictionaries and glossaries.

3. Related Work

The computational work on the Bodo language is not much, and only recently has it started. The majority of the corpus created for Bodo is dependent on financial support or research grants. In the work done by (Brahma et al., 2012), 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. In the paper (Islam et al., 2018), they constructed an English-Bodo parallel corpus in general and newspaper domains for the Bodo language. They developed the E-BPTC tool for typing the text of both English and Bodo languages that translated Bodo sentences into the corresponding English sentences collected from different sources. The general domain consists of English-Bodo parallel sentences that are commonly used in daily life. They collected from different sources such as monolingual corpus, dictionaries, books, and the web. The general domain contains over 6500 parallel sentences. The newspaper domain consists of 4000 English-Bodo parallel sentences related to news, important and general happenings. The newspapers they used for data collection are the English Newspaper (*The Assam Tribune and The Times of India*) and Bodo Newspaper (*Bodoland Sansri*). In the recent work done by (Narzary et al., 2019) the first neural machine translation baseline model for English → Bodo with BLEU (Papineni et al., 2002) score of 17.9 on a corpus size of 20,901 parallel sentences.

4. Problem Statement

In this section, we discuss the data unavailability issue of the Bodo language. As 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. In the context of machine translation for English-Bodo, the majority of the parallel corpus is available for three domains such as tourism, agriculture, and health. The total corpus for tourism, agriculture, and health domain is 11977, 4000, and 12382 sentences respectively[6]. There are no general domain parallel datasets, thus restricting the quality of the translation.

5. Methodology

In the following section, we describe our methodology for generating datasets from Bodo books. We followed a four-step process.

- Book Collection
- Scanning
- Text Extraction
- Manual Cleaning

[5] https://incubator.wikimedia.org/wiki/Wp/brx

[6] Available at TDIL-DC https://tdil-dc.in/

Book Title	Genre
[B] एत'वा मुन्दाया दावहायाव देरहाबाय [T] Etoa Mundaya Dauhayau Derhabai	Fiction
[B] दाय आरो साजा [T] Dai arw Saja	Novel
[B] सेक्सपीयारनि जुलियास सीजार [T] Shakespeareni Julius Caesar	Play & Drama
[B] आरबनि मोनानि सल [T] Aborni mwnani solo	Children
[B] लेंडि-चाटार्लिनि गोसो थोनाय [T] Lady Chatterleyni gwsw twnai	Novel
[B] दोंसे खैफोदनि सल' [T] Dwngse kwipwdni solo	Fiction
[B] खेबसे गंसे गामियाव [T] Kebse gongse gamiau	Children
[B] जाय जुनारखौ नों बावनो हाया [T] Jai junarkou nwng baounw haya	Children
[B] थिखिर गैयि दावबायारि [T] Kiter gwiywi daubayari	Children
[B] बिफाङालाइ मा? [T] Bipangalai ma ?	Children
[B] बिरजु आरो बिरग्रा गराइ [T] Birju arw birgra gorai	Children
[B] आंनि दावबायनाय [T] Angni daobainai	Biography
[B] लैथोनि हानन्थियारिफोर [T] Lwitwni hantiaripwr	Play

Table 4: Collected Books with its corresponding genre. A total of 13 books were collected, are in the hard copy. Here B represents book title in Bodo and T represents the English transliteration of book title.

5.1. Book Collection

Book collection is a tedious process. We collected 13 books that are translated into Bodo from English books. This is done so that a parallel corpus can be created. However, this method of dataset building from books applies to all kinds of datasets. The collected books, along with their corresponding genre, are shown in Table 4.

5.2. Scanning

Books can be scanned in two ways: with a laptop or desktop (with an external scanner) and with a smartphone (mobile device). Scanning using external scanner devices requires extra hardware, and the scanned images of the books are then stored on a laptop (desktop). Using a smartphone to scan books is more straightforward and accessible. We used Android smartphones with moderate-to-good camera quality. The pictures of the pages of the collected books were taken. Special care was taken to exclude cover pages, headers, and footer margin lines from the scanning process. The scanned quality and text extraction process are mostly dependent on each other.

5.2.1. Text Extraction

The extraction process is performed using Google Keep (Android Version), which is a note-keeping application with optical character recognition capabilities. The *"Grab text"* functionality of Google Keep allows for text extraction. The extracted texts are then copied to another file like the raw text format (TXT) or document files formats such as DOCX or ODF. Due to the required manual preprocessing, the texts are saved in document format in Google Docs. All the page-wise extracted texts are copied into a single document for each book. The cover pages, table of contents, and preface were excluded from the extraction process. The extracted number of pages from the corresponding books is shown in Table 5.

5.3. Manual Cleaning

The extracted text is cleaned using the free cloud-based document editor Google Doc. This is primarily due to its support of multiple collaborators and integration with Google Drive for easy access. The texts are checked manually word by word and matched with the actual book text. This step is the most tedious and time-consuming, but it must

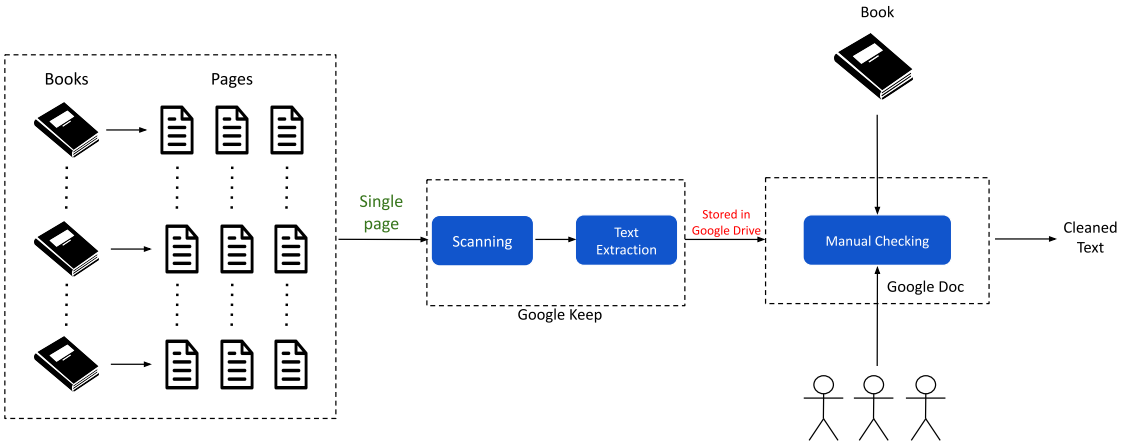

Figure 2: Monolingual Corpus extraction process. The single page is scanned and text are extracted using Google Keep. The raw extracted text is then manually cleaned.

Book	Pages
एत'वा मुन्दाया दावहायाव देरहाबाय Etoa Mundaya Dauhayau Derhabai	54
दाय आरो साजा Dai arw Saja	128
सेक्सपीयारनि जुलियास सीजार Shakespeareni Julius Caesar	131
आरबनि मोनानि सल Aborni mwnani solo	120
लेडि-चाटार्लिनि गोसो थोनाय Lady Chatterleyni gwsw twnai	179
दोंसे खैफोदनि सल' Dwngse kwipwdni solo	26
खेबसे गंसे गामियाव Kebse gongse gamiau	19
जाय जुनारखौ नों बावनो हाया Jai junarkou nwng baounw haya	38
थिखिर गैयि दावबायारि Kiter gwiywi daubayari	18
बिफाङालाइ मा? Bipangalai ma ?	12
बिरजु आरो बिरग्रा गराइ Birju arw birgra gorai	9
आंनि दावबायनाय Angni daobainai	157
लैथोनि हानन्थियारिफोर Lwitwni hantiaripwr	29

Table 5: Collected Book along with its corresponding number of pages scanned, extracted and cleaned manually.

be done with the utmost care. The text extracted by Google Keep is not perfect, and some characters from the Bodo language are not correctly extracted. In some cases, it is completely missed by OCR. Any word or character that is found incorrect is fixed by correcting it in the documents manually. The correction is made using the Devanagari script keyboard in Windows. The cleaning and validation of the extracted text were done by native Bodo speakers. The extraction and manual cleaning process takes time and requires an internet connection. The use of Google Docs allowed us to perform collaborative cleaning.

6. Results

We extracted raw text from 13 books from children, fiction, novel, play, biography, and drama genres. The raw text is extracted and cleaned manually and stored in Google Drive. The books files are then downloaded as TXT file and then empty lines were removed, and the text files were combined into a single TXT file[7]. The resulted corpus consists of 192,327 tokens and 32,268 unique tokens as shown in Table 6. The corpus created have a type-token ratio of 0.16 suggesting that the corpus have substantial lexical richness.

Tokens	Unique Tokens	Type-Token Ratio
192327	32268	16.77%

Table 6: Corpus Statistics.

6.1. Evaluation

For the evaluation of the dataset, we performed manual cross-checking and language modelling.

[7]The empty lines and file combination was done using Python script

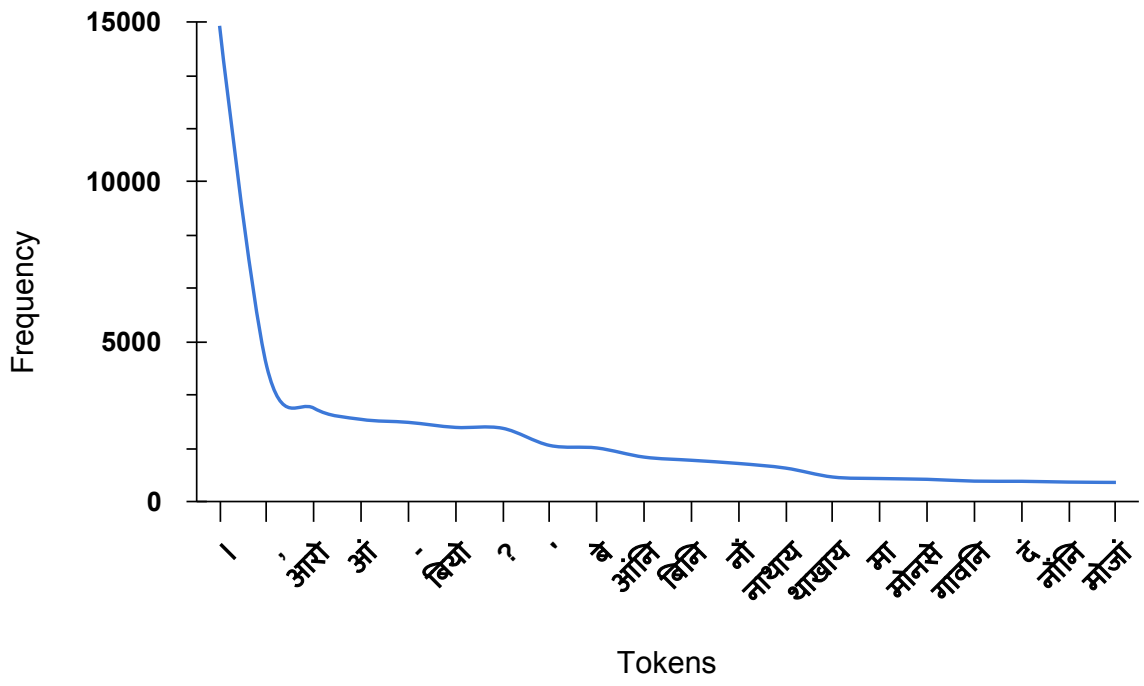

Figure 3: Top 20 frequently occurring tokens of combined TXT files.

6.1.1. Manual Cross-Checking

To evaluate the correctness of the cleaned text mentioned in Section 5.3 we randomly distributed samples to five (5) 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.

6.1.2. Language Modelling

We performed statistical language modelling using the Stanford Research Institute Language Modeling Toolkit (SRILM)[8]. To perform the evaluation of the language model, we measure its perplexity. Perplexity is an intrinsic evaluation that measures performance of a language model. It is calculated as the probability of the test set normalized by the number of words. Lower perplexity, better is the language model. A tokenized version of the Bodo Monolingual Text Corpus ILCI-II[9], a general domain corpus containing 31,026 sentences and 1,029,408 words, is used for training. We trained on n-gram 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, as shown in Table 7. The perplexity of random samples is very high. However, based on the results of manual cross-checking and the results of language modelling, it suggests 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.

Sample	Genre	Perplexity	OOV
1	Fiction	31127.55	63
2	Novel	24553.62	28
3	Play	37168.39	40
4	Children	62051.7	21

Table 7: Random samples number, genre, perplexity, and No. of Out of Vocabulary tokens. The samples for each genre were chosen randomly from the extracted text. N-gram of 3 is used because of its lower perplexity while training.

7. Conclusions

In this paper, 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.

[8] http://www.speech.sri.com/projects/srilm/
[9] Provided by TDIL-DC

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. This process is under-represented among the low-resource language community and can be used to generate a significant amount of data in zero-resource settings where the data may be available in the print book only. Our proposed methodology may also be carried out using the services provided by Microsoft Office 365 suite, which can be further explored by researchers.

Although this process works for zero-resource setting languages, it is highly dependent on the quality of the OCR provided by Google Keep. Furthermore, the work done can be extended to build a parallel corpus for the English-Bodo language pair for machine translation tasks.

The language model of the created corpus is made available for the research community at GitHub[10]. Keeping in mind the copyright issues, the dataset will be made available after taking permission from its respective author.

8. Discussions

8.1. OCR Problem

Since Bodo uses the same Devanagari script as Hindi and Nepali languages. The OCR used by Google Keep performs quite well, although Bodo doesn't have its own OCR. Despite, its good performance, it can't recognize characters/syllables such as ङ, न्व',छै , that exist in Bodo but otherwise do not exist in other languages. Hence, having an OCR specifically for the Bodo language can help in the text extraction process. However, most of the existing research is focused on languages with high resources as compared to low-resource languages.

8.2. Spelling variation Problem

For a young language, it is very difficult to manage the use of different spellings, as reported by (Brahma et al., 2012). Therefore, there is a need to make digitally accessible dictionaries and glossaries that can be used in dataset preparation and processing. Hence, the existence of general domain monolingual dataset can help in this process.

8.3. Tonal Characteristics

Bodo as a tonal language, has two tones: *high* and *low*. However, the tonal characteristics of such words are not included in the written format. This is a concerning issue with most natural language processing research focusing on considering unique words as vocabulary and having a learned vector representation (Mikolov et al., 2013) (Bojanowski et al., 2017). Therefore, it is important to further study and address the effect of the tonality of the Bodo language and the created monolingual corpus can be used as the dataset for study in the future.

8.4. Lack of participatory research

Bodo is spoken by the tribal community of India, and the lack of participatory research in the field of natural language processing is the root cause of data scarcity and low research contribution. This is visible with the non-existence of Wikipedia contributors, language translators, and open source contributors. Hence, it is important to have more inclusive and community-based participatory research to accelerate the NLP studies of the Bodo language, considering its linguistic features.

9. Acknowledgements

We would like to thank Ajwli Basumatary, Mr. Kunal Kumar Basumatary, Mwnabili Narzary, Jeenu Rani Brahma, Krishna Wary, Gangashree Narzary, Phulung Borgayari, and Ansumai Brahma for their help in extraction and manual cleaning progress. We also extend our thanks to the Indian Language Technology Proliferation and Deployment Centre (TDIL-DC) for providing us Bodo Monolingual Text Corpus ILCI-II dataset.

10. Bibliographical References

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bhattacharya, P. C. (1964). *A descriptive analysis of the Boro language*. Ph.D. thesis, Gauhati University.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Brahma, B., Barman, A. K., Sarma, S. K., and Boro, B. (2012). 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. The COLING 2012 Organizing Committee.

Census. (2011a). Abstract of speakers' strength of languages and mother tongues. *Census 2011*.

Census. (2011b). Comparative speakers' strength of languages and mother tongues - 1971, 1981, 1991, 2001 and 2011. *Census of India 2011*.

Islam, S., Paul, A., Purkayastha, B. S., and Hussain, I. (2018). Construction of english-bodo parallel text corpus for statistical machine translation. *International Journal on Natural Language Computing (IJNLC) Vol*, 7.

[10] https://github.com/bodonlp/bodolm-2022

Joshi, P., Barnes, C., Santy, S., Khanuja, S., Shah, S., Srinivasan, A., Bhattamishra, S., Sitaram, S., Choudhury, M., and Bali, K. (2019). Unsung challenges of building and deploying language technologies for low resource language communities. In *Proceedings of the 16th International Conference on Natural Language Processing*, pages 211–219, International Institute of Information Technology, Hyderabad, India, December. NLP Association of India.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, et al., editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Narzary, S., Brahma, M., Singha, B., Brahma, R., Dibragede, B., Barman, S., Nandi, S., and Som, B. (2019). Attention based english-bodo neural machine translation system for tourism domain. In *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, pages 335–343.

Narzary, M., Muchahary, G., Brahma, M., Narzary, S., Singh, P. K., and Senapati, A. (2021). Bodo resources for nlp-an overview of existing primary resources for bodo. *AIJR Proceedings*, pages 96–101.

Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Fagbohungbe, T., Akinola, S. O., Muhammad, S., Kabongo Kabenamualu, S., Osei, S., Sackey, F., Niyongabo, R. A., Macharm, R., Ogayo, P., Ahia, O., Berhe, M. M., Adeyemi, M., Mokgesi-Selinga, M., Okegbemi, L., Martinus, L., Tajudeen, K., Degila, K., Ogueji, K., Siminyu, K., Kreutzer, J., Webster, J., Ali, J. T., Abbott, J., Orife, I., Ezeani, I., Dangana, I. A., Kamper, H., Elsahar, H., Duru, G., Kioko, G., Espoir, M., van Biljon, E., Whitenack, D., Onyefuluchi, C., Emezue, C. C., Dossou, B. F. P., Sibanda, B., Bassey, B., Olabiyi, A., Ramkilowan, A., Öktem, A., Akinfaderin, A., and Bashir, A. (2020). Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online, November. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.