Bodo Resources for NLP - An Overview of Existing Primary Resources for Bodo

Mwnthai Narzary*, Gwmsrang Muchahary, Maharaj Brahma, Sanjib Narzary, Pranav Kumar Singh, Apurbalal Senapati

Department of Computer Science and Engineering, Central Institute of Technology, Kokrajhar, India.

*Corresponding author
doi: https://doi.org/10.21467/proceedings.115.12

ABSTRACT

With over 1.4 million Bodo speakers, there is a need for Automated Language Processing systems such as Machine translation, Part Of Speech tagging, Speech recognition, Named Entity Recognition, and so on. In order to develop such a system it requires a sufficient amount of dataset. In this paper we present a detailed description of the primary resources available for Bodo language that can be used as datasets to study Natural Language Processing and its applications. We have listed out different resources available for Bodo language: 8,005 Lexicon dataset collected from agriculture and health, Raw corpus dataset of 2,915,544 words, Tagged corpus consisting of 30,000 sentences, Parallel corpus of 28,359 sentences from tourism, agriculture and health and Tagged and Parallel corpus dataset of 37,768 sentences. We further discuss the challenges and opportunities present in Bodo language.

Keywords: Bodo language, Natural language Processing, Available Resources, Corpus, Text corpus, Parallel corpus

1 Introduction

Bodo as a language is largely spoken in Bodoland Territorial Region[1] and is the associate official language of the state of Assam, India. It is one of the recognized languages of India[2]. According to 2011 Census of India [1][2], there are 1,454,547 native speakers and a total of 1,482,929 Bodo speakers.

Table 1: *Year wise census of Bodo speakers collected by Government of India.*

Year	No. of Speakers
1971	556,576
1991	1,221,881
2001	1,350,478
2011	1,482,929

Natural Language Processing (NLP) is the process of data manipulation of human language or natural language to understand and analyze the text and speeches. NLP makes it possible for a computer to understand, read text, hear speech, and measure sentiments. With millions of Bodo speakers, the need of NLP for Bodo language is large. NLP for Bodo is an emerging field of study and not much work is available. There is lots of research

[1] Formerly known as Bodoland Territorial Autonomous District (BTAD)
[2] Scheduled languages of India: Bodo is one of the 22 scheduled languages, added in 2004

Proceedings DOI: 10.21467/proceedings.115; Series: AIJR Proceedings; ISSN: 2582-3922; ISBN: 978-81-947843-5-7

and work that needs to be done for the Bodo language in NLP. Applications like language translator, grammar checker, auto correction, auto completion, name entity recognition, language modeling, question answering, part of speech recognition, summarization, chunking and machine translation for Bodo language are yet to be fully studied. Although research for some of the applications have started it is still in an early stage.

This paper is written with the aim to enumerate the foundation of Bodo language resources for design and development of different NLP applications. In this paper we mainly focus on collecting and finding out existing resources and projects done in the field of natural language processing. The resources shown in the paper gives the present status of the publicly available dataset for research purposes. We have also listed out the domain specific corpuses, making it easy to find out the missing domains for which the corpus building can be taken up. Finally, it shows that Bodo is a low resource language, compared to high resource languages like English.

2 Background and Related works

Bodo pronounced as Boro is a language which is mainly spoken in the North-Eastern part of Brahmaputra valley of India. The language is the part of Sino Tibetan language family under the part of Assam-Burmese group. It has a very long history of evolution from inception till today. Bodo language has a rich oral history and no written scripts were present before the 20th century. However, some researchers suggested having its own script called Deodhai [3], which is lost at present times.

Bodo script movement was held in 1974-75, where many sacrificed their lives for Bodo script, ultimately leading to acceptance of Devanagari [4] script. From 1963, onwards Bodo language was introduced as a medium of instruction in Bodo dominant schools. Later 1997, post graduate degree was offered for Bodo language in Gauhati University. In 2003, it became one of the scheduled languages and recognized as the language of India. These lead to Government of India various initiatives for Bodo languages such as Sahitya Academy awards, used in official activities. In 2020, Bodo is declared as an associate official language in the entire state of Assam. From an NLP perspective building of corpus and tools was undertaken the application of NLP in Bodo language is relatively new and the field of study is just starting to grow.

Various efforts have been put into developing corpus and different tools. Here we try to provide a brief overview of related works done in the field of NLP for Bodo Language. In the paper [5] authors mention the need for more computerized information of English-Bodo language pairs. They also claimed to use Statistical Machine Translation and got messy results by using collected 6,000 general English-Bodo parallel corpus. In the paper [6] authors mentions about developing machine translation through transliteration. The system developed used hybrid technique to improve accuracy of translation results in the multi domain English-Bodo translation system. In the paper [7], an algorithm is proposed to syllabify Bodo words into syllables with upto 95.5%.

The foundation for developing a Bodo Wordnet [8] started with a sponsored project by the Department of Information Technology, Ministry of Communication and Information Technology, Government of India. The paper mentions the expansion of an already existing English-Hindi wordnet to English-Bodo wordnet. In this paper [9] authors shared an experience while building a corpus of containing 1.5 million words of Bodo language. Author mentioned difficult problems while entering the documents in the format. The paper mentions various problems like spelling variation, word split, joined sentences/grammatical error, incomplete sentences while building Bodo corpus.

Ministry of Information Technology, Department of Electronics and Information Technology (DeitY), the Central Institute of Indian Languages (CIIL), Mysore and Centre for Development of Advanced Computing

(C-DAC) took some essentials steps to develop Bodo language in digital media and executed project named **Language Technology Development Project** (LTDP) in collaboration with C-DAC, Pune and Gauhati University. In this project a script grammar for Bodo language is developed, compiled a Bodo corpus and a spell checker tool. Recently projects like UNICODE compliant font sets, keyboard drivers, word processors, CLDS projects are initiated and sponsored by the Director of Information Technology, Government of India. However, these projects are stopped half way and are not completed. A project named **Indian Language Corpora Initiative** project phase-II has been running since January, 2013 in Gauhati University Bodo department[3]. Monolingual corpora has a total 20 files which contains 100 raw words each. Similarly, parallel corpus are collected from the Health and Tourism domain.

Recently, in paper English-Bodo Neural Machine Translation for Tourism Domain [10] machine translation using deep neural networks is taken up and is the first known work of creating a baseline model in Neural Machine Translation. The paper uses two layer bidirectional Long Short Term Memory (LSTM), achieving Bilingual Evaluation Understudy Score [11] (BLEU) score of 11.8 (baseline model). The baseline model is improved by introducing attention mechanisms resulting in improvement of BLEU score from 11.8 to 16.71, which is further improved to 17.9 with beam search.

3 Resources

Natural Language Processing (NLP) tasks depend on the availability of corpus text. In recent years, the use of deep learning in NLP is increasing. The need of a dataset in deep learning techniques is huge for testing set to generalize well. The NLP research for Bodo languages is relatively new, making the availability of the corpus in the public domain is rare. Different Government institutions in India have put forward various consortia to collect and build corpus. The two such programmes are *English to Indian Language Machine Translation* (EILMT) and *Indian Languages Corpora Initiative* (ILCI) consortia. In the subsequent sub-sections we have listed down the publicly available resources for various domains: Literature (L), Science (S), Media (M), Art (A), Aesthetics (Ae), Commerce (C), Social Sciences (So), Agriculture (Ag), Health (H), Tourism (T), Entertainment (E) and General (G). The Tourism domain corpus have missing sentences which is reported in the works by [10] and number of sentences in the dataset will be lower than the reported figures.

Lexicon

Lexicon resources for Agriculture with 2,384 words and Health domain with 5,621 words shown in Table 2 is available in Indian Language Technology Proliferation and Deployment Centre (TDIL-DC) (http://www.tdil-dc.in/)[4] contributed by EILMT Consortia, CDAC Pune.

Table 2: *Available Lexicon*

Corpus Name	Domain	Source
English-Bodo Agriculture Lexicon - EILMT	Ag	TDIL-DC
English-Bodo Health Lexicon – EILMT	H	TDI-DC

[3]http://www.igntu.ac.in/eContent/IGNTU-eContent-816960917521-MA-Linguistics-4-HarjitSingh-ComputationalLinguistics-5.pdf

[4] http://www.tdil-dc.in/

Proceedings DOI: 10.21467/proceedings.115
ISBN: 978-81-947843-5-7

Series: AIJR Proceedings
ISSN: 2582-3922

Raw Corpus

Bodo raw corpus Bodo General Text Corpus for domain of Literature, Science, Media, Art is available and contributed by Gauhati University. The Bodo Raw Text Corpus available at National Platform for Language Technology (NPLT)[5] consists of 2,915,544 words, 80 titles and 5 domains shown in Table 3.

Table 3: *Raw Corpus*

Corpus Name	Domain	Source
Bodo General Text Corpus	L, S, M, A	TDIL-DC
A Gold Standard Bodo Raw Text Corpus	Ae, C, M, S, So	NPLT

Tagged Corpus

Bodo Monolingugal Text Corpus shown in Table 4 is created under Indian Languages Corpora Initiative phase–II (ILCI Phase-II) project initiated by the Government of India and Jawaharlal Nehru University, New Delhi. The corpus has 30,000 sentences of General domain. The available corpus is Part Of Speech (POS) tagged according to Bureau of Indian Standards (BIS) tagset.

Table 4: *Tagged Corpus*

Corpus Name	Domain	Source
Bodo Monolingual Text Corpus ILCI-II	G	TDIL-DC

Parallel Corpus

Parallel corpus for English-Bodo language pair is available for Tourism, Agriculture and Health domain shown in Table 5. Tourism, Agriculture and Health corpus consists of 11977, 4000 and 12382 sentences respectively. English-Bodo Parallel Tourism Text corpus consists of 11,977 sentences in excel format, contributed by Gauhati University under EILMT consortium. The vocabulary of the corpus consists of various cultures and civilizations, destinations, visiting places, names, art and architecture. Largely, the corpus contains text related to tourism and tourist destinations. English-Bodo Parallel Agriculture Text corpus consists of 4,000 sentences in excel format, developed under English to Indian Language Machine Translation consortium. English-Bodo Parallel Health Text corpus is built under English to Indian Language Machine Translation consortium. The corpus is created in excel format and contains 12,382 parallel text sentences in English and Bodo.

Table 5: *Parallel Corpus*

Corpus Name	Domain	Source
English-Bodo Sentences of Tourism Domain	T	TDIL-DC
English-Bodo Agriculture Text Corpus – EILMT	Ag	TDIL-DC
English-Bodo Health Text Corpus – EILMT	H	TDIL-DC

[5] https://nplt.in

Tagged and Parallel Corpus

According to our findings, two resources shown in Table 6, are both tagged and consists of parallel text are available for Bodo. Both of the corpus is created under the Indian Languages Corpora Initiative phase-II, initiated by the Ministry of Electronics and Information Technology (MeitY), Government of India and contributed by Jawaharlal Nehru University, New Delhi. The parallel corpus are for Hindi-Bodo language pairs. Hindi-Bodo Agriculture & Entertainment text corpus is built with Hindi as the source language which is translated into Bodo. The 37,768 sentences are related to Agriculture & Entertainment. Sentences are POS tagged according to the Bureau of Indian Standards (BIS) tagset.

Similarly, Hindi-Bodo Parallel Chunked text corpus is created with Hindi as source language and the corresponding Bodo sentences are translated. The parallel chunked corpus consists of 70,000 sentences from 4 domains: *Health, Tourism, Agriculture and Entertainment*. The Bodo sentences are POS tagged according to BIS tagset and chunked

Table 6: *Tagged and Parallel Corpus*

Corpus Name	Domain	Chunked	Source
Hindi-Bodo Agriculture & Entertainment Text	E, Ag	No	TDIL-DC
Hindi-Bodo Parallel Chunked Text	H, T, Ag, E	Yes	TDIL-DC

4 Challenges and Opportunities

Addition of new vocabulary is one of the major challenges. Since, new words need to be added in the corpus. There are no existing system that can incorporate addition of new vocabulary to corpus. Such a system can be useful from the perspective of language processing. Secondly, there is no research community dedicated to Bodo language making it difficult for existing researchers to share datasets and results. Making the reproducibility of results very difficult. Third, there are no publicly available benchmarks and baseline models making it hard to compare new NLP techniques.

5 Conclusions

We have divided the resource list based on corpus type Lexicon, Raw corpus, Tagged corpus, Parallel corpus, Tagged and Parallel corpus. The available resources will help the research community or anyone in general to perform experiments. Raw text corpus can be used for developing language models, spelling checker, creating dataset and building of other tagged corpus is possible. POS tagged corpus particularly will help in developing automatic POS tagging systems, and in developing other NLP systems. Parallel corpus can be particularly used for studying and developing Machine translation systems. The resources described are in public domain corpus accessible for research purposes from TDIL-DC, which requires to be associated with institutions and organizations to get access. Making it less open for the research community to grow and restricts the innovations and progress. The open source community for Bodo is particularly missing, making the contribution of data from internet sources like Wikipedia negligible. We believe resource lists presented in this paper will help the research community to work on Bodo language for NLP without having to search for available resources and speed up the dataset finding process.

Proceedings DOI: 10.21467/proceedings.115
ISBN: 978-81-947843-5-7

Series: AIJR Proceedings
ISSN: 2582-3922

6 Acknowledgements

The resources presented in the paper is possible due to Government of India initiatives: EILMT and ILCI. TDIL-DC (http://www.tdil-dc.in) played a crucial role in indexing the corpus created by such initiatives and making it accessible over the internet.

References

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

[2] Census 2011, "Comparative speakers' strength of languages and mother tongues – 1971, 1981, 1991, 2001 and 2011"

[3] Laishri Mahilary, "Scripts used in missionary period of bodo literature: Discussion from linguistic point of view," *International Journal of Research in Humanities, Arts and Literature (IMPACT: IJRHAL)*

[4] M.S. Prabhakar, "The politics of a script: Demand for acceptance of roman script for bodo language". *Economic and Political Weekly*, 9(51): 20972102

[5] Saiful Islam and Bipul Syam Purkayastha, "Implementation of english to bodo machine translation using smt approach,"

[6] Saiful Islam and Bipul Syam Purkayastha, "Bodo to english machine translation through transliteration"

[7] Jyotismita Talukdar, Chandan Sarma, and Prof. P.H. Talukdar. Automatic syllabification rules for bodo language. *International Journal Of Computational Engineering Research*, pages 110-114.

[8] Shikhar Kr Sarma, M. Gogoi, B. Brahma, and Mane Bala Ramchiary. A wordnet for Bodo language: Structure and development. *Global Wordnet Conference (GWC10), Mumbai, India*.

[9] Biswajit Brahma, Anup Kr. Barma, Shikar Kr. Sarma, and Bhatima Boro, "Corpus building of literary lesser rich bodo: Insights and challenges," *In Proceedings of the 10th Workshop on Asian Language Resources*, pp. 29-34, 2012.

[10] Sanjib Narzary, Maharaj Brahma, Bobita Singha, Rangjali Brahma, Bonali Dibragede, Sunita Barman, Sukumar Nandi, and Bidisha Som, "Attention based english-bodo neural machine translation system for tourism domain", *3rd International Conference on Computing Methodologies and Communication (ICCMC)* pages 335-343

[11] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. "Bleu: a method for automatic evaluation of machine translation.", *Proceedings of the 40th Annual Meeting of the Association for Computional Linguistics*