

AssameseBackTranslit: Back Transliteration of Romanized Assamese Social Media Text

Hemanta Baruah, Sanasam Ranbir Singh, Priyankoo Sarmah

Centre for Linguistic Science & Technology
Indian Institute of Technology Guwahati

The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation

May 2, 2024



OUTLINE

- 1 Introduction
- 2 Types of Transliteration
- 3 Language Background: Assamese
- 4 Challenges of Transliteration in Social Media
- 5 Dataset Description
 - Data Collection
 - Data Annotation
 - Data Annotation Guidelines
 - Data Annotation Tool
 - Data Statistics
- 6 Experimental Setup
- 7 Experimental Dataset
- 8 Experimental Result and Discussion
- 9 Error Analysis
 - Identified Transliteration Output
- 10 Conclusion and Future Work

- **Transliteration** : Process of phonetic transformation of a word/token from source language script to a target language script.



Figure: Example of Transliteration

- Transliteration helps to pronounce words and names in foreign languages.

Types of Transliteration I

- **Forward Transliteration:** Process of transliterating native terms written using a non-native or foreign script.



Figure: Example of Forward Transliteration: Native Hindi language text written using Hindi(**Devanagari script**) converted to English(**Roman/Latin script**)

Types of Transliteration II

- **Backward Transliteration:** Process of transliterating a term written in non-native or foreign script back to its original native script.

E.g : **Gulabi aankhen
jo teri dekhi**

Script language - English

Underlying language - Hindi



गुलाबी आँखें जो तेरी देखीं

Script language - Hindi

Underlying language - Hindi

Figure: Example of Backward Transliteration: Hindi language text written in foreign language English with (**Roman/Latin script**) converted back to original native Hindi language text in (**Devanagari script**)

Language Background: Assamese



- Official language of Assam, a north-eastern state of India.
- Belongs to the Indo-Aryan language family.
- 15.2 million native Assamese speakers. [4]

Assamese Alphabets and Digits

Vowels and vowel symbols										
অ	আ	ই	ঈ	উ	ঊ	ঋ	এ	ঐ	ও	ঔ
	া	ি	ী	ু	ূ	্ৰ	ে	ৈ	ো	ৌ
[ɔ]	[a:]	[i]	[i:]	[u]	[u:]	[ri]	[e]	[oj]	[o]	[ou]
Consonants										
ক	খ	গ	ঘ	ঙ	চ	ছ	জ	ঝ	ঞ	
[k]	[k ^h]	[g]	[g ^h]	[ŋ]	[tʃ]	[tʃ ^h]	[dʒ]	[dʒ ^h]	[ɲ]	
ট	ঠ	ড	ঢ	ণ	ত	থ	দ	ধ	ন	
[t]	[t ^h]	[d]	[d ^h]	[ɳ]	[t̪]	[t̪ ^h]	[d̪]	[d̪ ^h]	[n]	
প	ফ	ব	ভ	ম	য	ৰ	ল	ৱ	শ	
[p]	[p ^h]	[b]	[b ^h]	[m]	[dʒ]	[r]	[l]	[ʋ]	[x]	
ষ	স	হ	ক্ষ	ড়	ঢ়	য়				
[x]	[s]	[h]	[k ^h ʃ]	[r̥]	[r̥]	[j]				
Digits										
০	১	২	৩	৪	৫	৬	৭	৮	৯	
xuinno	ek	dui	tini	sari	pas	soy	xat	ath	no	
Zero	One	Two	Three	Four	Five	six	seven	Eight	Nine	

Figure: Assamese Alphabets and Digits

Challenges of Transliteration in Social Media

← Tweet



@himantabiswa apunr bule students blkr krne bht mrom etya kot gol mrm 🤔?? Amk kio mitrur mukhr loi thli ase mama?? Youth mttrs
#CancelAssamBoardExams
#BoycottRanojPegu
#WeWantJusticeNotInjustice

1:07 PM · Jun 6, 2021 · Twitter for Android

5 Retweets 1 Quote Tweet 4 Likes



Text Documents Websites

GUJARATI - DETECTED ENGLISH GUJARATI ASSAMESE

ENGLISH ASSAMESE GUJARATI

@himantabiswa apunr bule students blkr krne bht mrom etya kot gol mrm 🤔?? Amk kio mitrur mukhr loi thli ase mama?? Youth mttrs
#CancelAssamBoardExams
#BoycottRanojPegu
#WeWantJusticeNotInjustice

@Himantabiswa apunr bule students blkr krne bht mrom etya kot gol mrm 🤔?? Amk kio mitrur mukhr loi thli ase mama?? Youth mttrs #CancelAssamBoardExams #BoycottRanojPegu #WeWantJusticeNotInjustice

Did you mean: @himantabiswa apunr bule students blkr krne bht mrom etya kot gol mrm 🤔?? Amk kio mitrur **muke loi thali** ase mama?? Youth **matters**
#CancelAssamBoardExams #BoycottRanojPegu #WeWantJusticeNotInjustice

197 / 5,000

@himantabiswa apunr bule ছাৰু-ছাৰী সকল blkr krne bht mrom এত্যা কোট গোল mrm 🤔?? আমক কিও মিত্ৰৰ মুখৰ লয়ই থলি আসে নানা ?? যুৱক-যুৱতী mttrs
#অসমবৰ্ড পৰীক্ষা বাতিল কৰক
#বয়কট ৰাণোজপেগু
#আমি ন্যায় বিচাৰো অন্যায় নহয়

Data Collection I

- Collected data from 3 popular social media platforms: **Facebook**¹, **YouTube**² and **Twitter(now X)**³.
- Used 3 publicly available APIs: **Facebook Graph API**⁴, **YouTube Data API**⁵ and **Tweepy API**⁶ for collecting data.
- Few of the popular Assamese Facebook pages, Twitter handles and YouTube channels are selected for data collection.
 - **Facebook:** Extracted comments from 2 selected Facebook pages: **Gauhati University Confession**⁷ page, **Chief Minister of Assam**⁸ Official facebook page.
 - **YouTube:** Collected comments on videos posted by 3 popular Assamese YouTube channels, 2 popular YouTubers and 1 YouTube channel of a private media house: **Dimpu's Vlogs**⁹, **News Live**¹⁰ and **Assamese Mixture**¹¹.

- **Twitter(now X):** Extracted mainly the reply tweets from one of the prominent twitter(X) handle of the current Chief minister of the State of Assam: **@himantabiswa**¹².

¹<https://www.facebook.com>

²<https://www.youtube.com>

³<https://www.twitter.com>

⁴<https://developers.facebook.com/docs/graph-api/>

⁵<https://developers.google.com/youtube/v3>

⁶<https://docs.tweepy.org/en/stable/>

⁷<https://www.facebook.com/confessionsGauhatiUniversity>

⁸<https://www.facebook.com/cmofficeassam>

⁹<https://www.youtube.com/@DimpusVlogs>

¹⁰<https://www.youtube.com/channel/UCrQHRYuJG8jmpUVALIC9Gkw>

¹¹<https://www.youtube.com/c/AssameseMixture>

¹²<https://twitter.com/himantabiswa>

Data Annotation

- Engaged 24 annotators, 3 linguistic experts as validators to annotate and validate the dataset.
- Annotators were selected based on their proficiency in both English and Assamese.
- Linguistic experts validated annotations using two tags: accept and reject.
- Acceptance threshold set to 80% for each annotators for an initial test set of 100 samples.
- Annotators exceeding the acceptance threshold(80%) in the initial test set will proceed to the final annotation task.

Data Annotation Guidelines

- 3 main tasks for the annotators:
 - **Language Identification:** Identify the language of the post at sentence level: english, assamese, assamese-mixed, or other.
 - **Back Transliteration:** Identified Assamese words written in Roman script transliterated back to the native Assamese script. English-origin words were retained if spelled correctly, otherwise replaced with their accurate spelling.
 - **Entity tagging:** Recognize terms in the sentence as a person's name, a geographical location, or the name of an organization. Annotators are asked to tag the term or phrase as <person>, <place>, or <organization>, respectively, by clicking the appropriate label below the post.
- Validators reviewed and marked posts as accept or reject, providing reasons for rejections.
- Rejected posts and reasons were reflected in annotators accounts for accurate tagging in future attempts.

Data Annotation Tool

- Built our in-house data annotation tool¹³.



Figure: A snapshot of the annotation tool

- Both annotators and validators were required to register and log in to our system first.

¹³<https://www.iitg.ac.in/cseweb/osint/annotation/>

Data Statistics

Table: Statistics of the collected dataset from three major social media sources along with the duration of data collection

Social Media Sources	Duration of Data Collection	#posts collected	#posts annotated	#words Assamese (total)	#unique Assamese words
Facebook	Dec-2013 to Feb-2017	409,168	5,300	71,800	79,200
YouTube	Jun-2018 to Aug-2023	385,676	50,000	426,089	
Twitter	Mar-2021 to Aug-2021	285,676	5,012	91,400	

- **Sentence-level statistics:**

- Total 60,312 sentences.
- Sentence length ranges from single-word to 162 words.
- Average sentence length: 11.4 (std. deviation: 8.38).
- Average code-mixing(English and Assamese) percentage: 20.1%

- **Word-level statistics:**

- Total 671,921 words.
- English words: 67,131 and Assamese words: 589,289
- 15,501 mixed-script(both Roman and native Assamese script in a single term) words.
- Out of 589,289 Assamese words, only 79,200 are unique.
- Extracted 65,614 unique transliteration pairs for our experiments.

Data Statistics: Variations in Social Media Dataset I

Social media data allows for multiple representations of a single source token, with a single token potentially exhibiting multiple transliteration variations.

- **Romanization variations for native words:** A maximum of 127 Roman transliteration variations observed for a native Assamese word in the dataset.

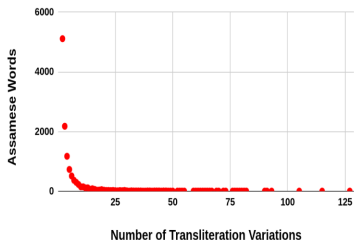


Figure: Distribution of Roman Transliteration Variations for Native Assamese Words

Data Statistics: Variations in Social Media Dataset II

- **Native Assamese word variations for Roman words:** Based on context or pronunciation similarity, one Roman word represents a maximum of 31 native Assamese words as observed from our dataset.

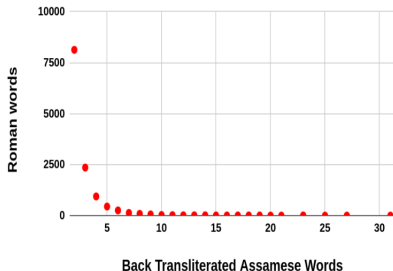


Figure: Distribution of Back Transliterated Native Assamese words represented by the Roman words

Data Statistics: Variations in Social Media Dataset

Table: Examples of Roman and native variations with frequencies in the dataset

Term	Script Language	Underlying Language	English Meaning	Total Variations	Variations with Frequencies
“বহুত”	Assamese	Assamese	Many	27	bohut: 2915, bhut: 1047, bht: 851, bahut: 651, bohhot: 126, buhut: 105, bhout: 47, bhot: 46, bahot: 24, bhtt: 13, boht: 12, bohud: 12, bhoot: 8, bout: 7, buhot: 7, bhohut: 6, bohout: 6, bhaut: 6, vohut: 6, bohoot: 5, bohoooooooouttttt: 1, bohoooooooout: 1, vohot: 1, bohoooooooout: 1, bohoooooooo: 1, bohoooooooo: 1, bhhhuut: 1
“gai”	Roman	Assamese	To Sing	5	গাই: 65, গায়: 18, গৈ: 3, যায়: 1, গান: 1

Experimental Setup I

- 1 **Statistical Machine Transliteration Setup:** Utilized two state-of-the-art statistical models for transliteration. Both systems were trained up to 4-grams for the language model.
 - **Phrase-based Statistical Transliteration model:** Using Moses¹⁴ decoder[7], employing GIZA++ [9] for character alignment and KenLM [5] for language modeling.
 - **Joint Source Channel based model:** Using SEQUITUR¹⁵, a joint n-gram-based string transduction system [3].
- 2 **Neural Network Based Transliteration Setup:** Employed two state-of-the-art neural network-based sequence-to-sequence transliteration models.
 - **BiLSTM encoder-decoder model with attention[1]:** Implemented using the OpenNMT toolkit [6].
 - **Neural Transformer model:** With the help of Fairseq [10] implementation of neural transformer [11] model.

3 Pre-trained Transliteration Model Evaluation Setup:

- Pre-trained multi-lingual transliteration model IndicXlit [8].
- Google Transliteration API for Google Input Tools¹⁶.
- Publicly available transliteration API, indictrans [2].

4 Transfer Learning Setup: Utilized three pre-trained models for transfer learning. One existing state-of-the-art multilingual transliteration model along with two pre-trained large language models.

- **IndicXlit model:** Transformer-based multilingual pre-trained transliteration model, IndicXlit [8], trained on 22 Indic languages, including Assamese.

Experimental Setup III

- **mT5 model [13]:** Multilingual transformer-based pre-trained large language model.
 - Trained on 101 languages from the mC4 [13] dataset, where Assamese is not included.
 - Pre-training objective is “**Span Corruption**”.
 - Employs pre-trained SentencePiece embedding for vocabulary creation.
- **ByT5 model [12]:** Similar pre-training objective like mT5 model but it includes heavy encoder and light decoder.
 - A tokenizer-free extension of the mT5 model.
 - ByT5 model processes UTF-8 bytes directly, enabling it to handle text in any language without a fixed vocabulary size.

¹⁴<https://github.com/moses-smt/mosesdecoder>

¹⁵<https://www-i6.informatik.rwth-aachen.de/web/Software/g2p.html>

¹⁶<https://www.google.com/intl/en/inputtools/try/>

Experimental Dataset

- In our word-level transliteration task, we maintained uniformity in the training, validation, and test sets across all ten different setups.
- We employed a 70-10-20 (train, validation and test set) split.
- Out of the total 65,614 unique transliteration pairs:
 - The training data includes 45,934 pairs.
 - The validation set contains 6,560 pairs.
 - We evaluated the models with 13,120 pairs in the testing set.

Experimental Result and Discussion I

Table: Transliteration result in terms of Word Error Rate(WER), Character Error Rate(CER), Number of Substitution, Insertion, Deletion errors and the BLEU (up to 4 gram) score for all the 10 setups (Experimental setups with the lowest Word Error Rate(WER), lowest Character Error Rate(CER) and the highest BLEU (up to 4 gram) score are highlighted in **bold**)

	Statistical Model Setup		Neural Model Setup		Pre-trained Model Evaluation Setup			Transfer Learning Setup		
	Phrase-based Statistical Transliteration model using Moses	Joint Source Channel based model using Sequitur	BiLSTM model with Attention	Neural Transformer model	IndicXlit model	Goggle Transliteration API	indic-trans model	IndicXlit model	Google's mT5 model	Google's ByT5 model
		setup 1								
WORD ERROR RATE (WER)	66.78	63.99	58.90	55.05	73.38	58.79	95.11	63.53	76.38	66.36
CHARACTER ERROR RATE (CER)	23.04	21.34	19.76	19.44	29.91	24.01	54.62	21.10	31.69	22.94
SUBSTITUTION ERROR	9146	8593	9728	8129	12265	9625	24507	8924	12633	8996
INSERTION ERROR	3093	3038	3311	2889	2946	3289	4021	2868	4340	3085
DELETION ERROR	3825	3258	3911	3289	5639	3821	9550	2921	5119	3911
BLEU (up to 4 gram) SCORE	64.41	67.33	68.60	69.15	54.03	67.61	24.04	66.48	55.50	65.93

Experimental Result and Discussion II

- 3 different evaluation matrices: **WER (word error rate)**, **CER (character error rate)** and **BLEU(upto 4 gram)** score are used for performance evaluation.
- Among statistical model setups, the Sequitur implementation of the joint n-gram-based transliteration model (**setup 2**) exhibited superior performance.
- The neural transformer model (**setup 4**) surpassed its neural counterpart, the BiLSTM model with attention (**setup 3**).
- Among pre-trained models, Google Transliteration API (**setup 6**) achieves comparable results with the best-performing neural transformer model (**setup 4**), demonstrating the highest performance among all three pre-trained model setups.
- The fine-tuned IndicXlit model (**setup 8**) achieves the highest performance among the three transfer learning setups: IndicXlit, mT5 and ByT5.
- Overall, Neural transformer model (**setup 4**), outperformed other baselines, achieving the lowest WER and CER values, along with the highest BLEU score.

Error Analysis

We have identified a few of the errors from the transliteration outputs of 10 different transliteration setups.

- 1 Errors due to multiple character mapping.
- 2 Errors due to short form representation.
- 3 Errors due to long form representation.
- 4 Errors due to alphanumeric word.

Identified Errors from Transliteration Output

Table: Comparison between the outputs of ten different transliteration model setups with the same Roman input (output of the setups that match with the ground truths are highlighted in blue)

Different Setups	Roman Input and Actual Native ground truth		Statistical Model Setups		Neural Model Setups		Pre-trained Model Evaluation Setups			Transfer Learning Setups		
	Roman Input	Assamese Native	Moses output (setup1)	Sequitur output (setup2)	BiLSTM output (setup3)	Transformer output (setup4)	IndicXlit model output (setup5)	Google Transliteration API output (setup6)	indic-trans model output (setup7)	Fine-tune IndicXlit model output (setup8)	Fine-tune mT5 model output (setup9)	Fine-tune ByT5 model output (setup10)
Multiple Character Mapping	hani	শনি	হানি	হানি	শনি	সানি	হানি	সানি	হানী	হানি	সানি	হানি
	xuola	শুৱলা	শূলা	সোলা	খুৱলা	খোৱালা	সোৱিলা	শুৱলা	জুওলা	শুৱলা	সোলা	শুৱলা
Short form Representation	jrhtt	যোৰহাটত	যোৰহাতত	যোৰহাতত	যোৰহাতত	যোৰহাটত	জাট	হৰ্ষদ	জরন্ত	যোৰহাট	যোৰহাতত	যোৰহাতত
	bhtor	বহুতৰ	বহঁতৰ	বহঁতৰ	বহঁতৰ	বহুতৰ	ভটৰ	ভাতৰ	ভটৰ	ভাতৰ	ভিতৰ	ভটৰ
Long form Representation	aaauuu	আওঁ	আওঁ	আ	আওঁ	আওঁ	আওঁওও	আআওওও	আউ	আও	আও	আআও
	bappppaaooiii	বাপ্পাঐ	বাপপায়	বাপ্পাও	বাপ্পাওঁ	বাপ্পাওঁ	বাপ্পাআওই	বাপ্পাপায়ী	বাপ্পাওঈ	বাপ্পায়	বাপ্পাঐ	বাপ্পাঐ
Alphanumeric Word	ai2e	এইটোৱে	এইটো	এইটো	এইটোৱে	এইটোৱে	এআইআইচে	অং২ৱে	আই2এ	এইটোৱে	এইটোৱে	এইটোএ
	kn2	কিন্তু	কিন্তু	কোনতো	কিন্তু	কিন্তু	কেএনচি	কঁ২	ন2	কিন্তু	কিন্তু	কিন্তু

Conclusion and Future Work

- Introduced pioneering Assamese back transliteration dataset from 3 major social media platforms.
- Neural Transformer model demonstrated superior performance in terms of WER, CER, and BLEU score.
- Future efforts target sentence-level transliteration and code-mixed text.
- Expand dataset diversity to cover diverse social media platforms and languages.
- Address challenges such as informal representations and slang prevalent in social media text.
- Investigate additional mainstream Large Language Models (LLMs).
- Fine-tune these models with task-specific augmented Assamese datasets due to Assamese's limited representation in existing pre-trained LLMs.

Thank you

Contact me in:

✉ `hemanta.b@iitg.ac.in`

🐦 `@hbdestine`

🌐 <https://in.linkedin.com/in/hemanta-baruah/>

🔍 https://scholar.google.com/citations?hl=en&user=CT_NUQMAAAAJ

References I



D. Bahdanau, K. H. Cho, and Y. Bengio.

Neural machine translation by jointly learning to align and translate.

In 3rd International Conference on Learning Representations, ICLR 2015, 2015.



I. A. Bhat, V. Mujadia, A. Tammewar, R. A. Bhat, and M. Shrivastava.

liit-h system submission for fire2014 shared task on transliterated search.

In Proceedings of the 6th Annual Meeting of the Forum for Information Retrieval Evaluation, FIRE '14, page 48–53, New York, NY, USA, 2014. Association for Computing Machinery.



M. Bisani and H. Ney.

Joint-sequence models for grapheme-to-phoneme conversion.

Speech communication, 50(5):434–451, 2008.



C. Chandramouli and R. General.

Census of india.

Rural Urban Distribution of Population, Provisional Population Total. New Delhi: Office of the Registrar General and Census Commissioner, India, 2011.



K. Heafield.

Kenlm: Faster and smaller language model queries.

In Proceedings of the sixth workshop on statistical machine translation, pages 187–197, 2011.



G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush.

Opennmt: Open-source toolkit for neural machine translation.

arXiv preprint arXiv:1701.02810, 2017.

References II



P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst.

Moses: Open source toolkit for statistical machine translation.

In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, pages 177–180, Prague, Czech Republic, June 2007. Association for Computational Linguistics.



Y. Madhani, S. Parthan, P. Bedekar, R. Khapra, V. Seshadri, A. Kunchukuttan, P. Kumar, and M. M. Khapra.

Aksharantar: Towards building open transliteration tools for the next billion users, 2022.



F. J. Och and H. Ney.

A systematic comparison of various statistical alignment models.

Computational linguistics, 29(1):19–51, 2003.



M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli.

fairseq: A fast, extensible toolkit for sequence modeling.

In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), pages 48–53, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.



A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin.

Attention is all you need.

Advances in neural information processing systems, 30, 2017.



L. Xue, A. Barua, N. Constant, R. Al-Rfou, S. Narang, M. Kale, A. Roberts, and C. Raffel.

Byt5: Towards a token-free future with pre-trained byte-to-byte models, 2021.

References III



L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel.

mT5: A massively multilingual pre-trained text-to-text transformer.

In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 483–498, Online, June 2021. Association for Computational Linguistics.