IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 12, Issue 5, May 2023

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2023 ||

| DOI:10.15680/IJIRSET.2023.1205097 |

Natural Language Processing based on Kokborok Language: A Brief Survey

Enjula Uchoi [1], Kanika Debbarma [2]

Assistant Professor Department of School of Computer Science and Engineering, Lovely Professional University, Phagwara Punjab, India[1]

Department of Information Technology, Padmashree Dr.D.Y Patil Institute of Engineering & Technology , India[2]

ABSTRACT: Compared to other national languages, low resource languages like Kok-Borok are rarely explored. Bengali and Roman script are frequently used in Kokborok, which as of yet lacks even an official script. Because of this the overall evolution of the language as well as the enlargement of the literature have both been slowed down in part as a result of this. In the field of Kokborok Natural Language Processing, a few research articles based on morphological analysis, stemming, name entity recognition, part of speech (POS), code-mixing, and code-switching Kokborok have been published in recent years. This study aims especially to define the Kokborok language trends, methods pursued in Kokborok language employed, and special accuracy evaluation metrics for the researchers to get brief idea and can explore much deeper in Kokborok language.

KEYWORDS: Part of speech, Kokborok, Stemming, Code-mixing, Code-switching, Name entity recognition.

I. INTRODUCTION

In the discipline of natural language processing, computer programmes and machine learning methods try to comprehend and use massive amounts of text data. Indian is a democracy country speakers of English as a second language make up the country's largest population thanks to its rich linguistic past and close relationship with English. Native Indian language speakers rarely use Unicode while exchanging information on social media platforms. They use Roman letters with English words or phrases through code-mixing with their own tongue to convey themselves. To address these difficult fundamental difficulties, research is being conducted and is now necessary. Natural languages can include a wide variety of unstructured means of communication, such as signs, signals, or dialects. This makes it more difficult for someone looking for specific information that may only have a passing familiarity with the subject.

II. HISTORY OF KOKBOROK

"Kok-borok," which combines the words "Kok" and "Borok"—where "Kok" denotes a language and "Borok"—denotes a person. The word "Kok-borok" describes the language used by the Borok people who reside there and the Borok people who reside close to the adjacent states of Mizoram, Manipur, and Assam as well as Myanmar and Bangladesh. Kokborok is primarily spoken in Tripura, where it is written using Roman and Bengali scripts. Roman script was preferred over Bengali script by most Kokborok speakers. Koloma was the name of the original Kokborok script. The only state in India with 184 rulers, Tripura, merged with the rest of the country on October 15, 1949.Rajratnakar's book contains examples of the Koloma script, which was employed to record the historical scripts. After translating the language into Sanskrit, the two Brahmins known as Sukreswar and Vaneswar repeated the process by translating the history into Bengali in the fourteenth century. The Kokborok language contains stem homophony, phonemic tone, word order, verb morphology, and verb-derived suffixes. English, Arabic, Bengali, Persian, and other languages have finally been heard by the Kokborok. The Kokborok language has a variety of phonological characteristics, including word order; stem homophony, verb morphology, and verb derivational suffixes.

.

.

|e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2023 ||

| DOI:10.15680/IJIRSET.2023.1205097 |

III. DIFFERENT APPROACHES AND METHODS

The study of natural language processing, a subfield of computer science, focuses on creating tools that can comprehend and analyse written or spoken human language. NLP sub-tasks vary, and may include different types of approaches. As per our knowledge and through information we have specified those methods which are already done in this low resources Kokborok language in the year 2012 to 2023 in the field of natural language processing.

TABLE 1: FEW RECENT SURVEY REFERENCES OF KOKBOROK

Ref.	Year	Task	Features Extracted	Algorithm and objectives	Accuracy (%)
1.	2012	Morphological Analysis of Kokborok for Universal Networking Language Dictionary	Morphological analysis	In order to create a Kokborok dictionary and Machine Translator, this paper focuses on integrating efforts on morphological analysis	NA
2.	2012	A Light Weight Stemmer in Kokborok A Light Weight Stemmer in Kokborok	Stemmer	It explains how to create a straightforward rule-based stemmer for Kokborok by removing the affixes and adding boundary rules where necessary.	minimum suffix stripping algorithm =80.02% and maximum suffix stripping algorithm =85.13%
3.	2012	Part of Speech (POS) Tagger for Kokborok	POS	It explains how Conditional Random Field (CRF) and Support Vector Machines (SVM) were used in Kokborok to construct a POS tagger utilising rule-based and supervised approaches.	Rule based =70% and supervised POS tagging = 84%.
4.	2012	Morphological Analyzer for Kokborok Morphological Analyzer for Kokborok	Morphological analysis	This article describes the creation and use of a database-driven affix stripping technique that was used to create the Morphological Analyzer.	Accuracy =80%
5.	2014	Stemmer for resource scarce language using string similarity measure	Stemmer	In this study, a Kokborok stemmer is built using a statistical method based on string measure.	Correct analysis of 78%.
6.	2014	Frequency based named entity recognition system for under resource language	Named entity recognition	This study examines the problems and difficulties associated with creating a frequency-based Named Entity Recognition (NER) system for Kokborok.	Success rate of 78%
7.	2015	Morphological Analyzer in the Development of Bilingual Dictionary (Kokborok-English) - An Analysis for Appropriate Method and Approach	Morphological Analyzer	This paper explained and analyse the morphology of the Kokborok language using a variety of NLP techniques.	NA

|e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2023 ||

| DOI:10.15680/IJIRSET.2023.1205097 |

8.	2019	Machine Translator of Kokborok Language for Machine Learning Using Natural Language Processing, An Universal Networking Language and Dictionary.	Universal Networking language and Dictionary	The prefix stripping approach was used in this research to create a straightforward rule-based stemmer for Kokborok.	Min-Suffix stripping algorithm= 80.02% and max-suffix= 85.13%
9	2020	An Unsupervised Word Level Language Identification of English and Kokborok Code-Mixed and Code-Switched Sentences	Dictionary based, N-gram Model for Kokborok mixed sentences	In order to accurately categorise each word, the proposed model combines a character-n-gram language model based on frequency lexicons with a languagedependent morphological dictionary-based model.	accuracy =84%,
10	2021	Efficient POS Tagger for Kokborok Language	POS Tagger	The rule-based POS tagger for Kokborok, an Indian language with few resources and no digitised literature, is the topic of this study. Naive Bayes (NB) and decision trees are two machine learning techniques they use for supervised learning.	Accuracy = 79%, 70% and 71% respectively.
11.	2022	Adoption of Natural Language Processing in Education: Implications for students	NLP	This paper emphasize a number of languages, especially tribal languages like Santhali and Kokborok, are mostly not taught in schools and are connected to limited economic prospects.	NA
12.	2022	English to Kokborok Bilingual Thesaurus for Natural Language Processing	Bilingual Thesaurus	The focus of this work is on creating a bilingual thesaurus in English and Kokborok, which will preserve the language's existing vocabulary and provide further contributions to Kokborok's NLP.	NA
13.	2023	Modeling a Hybrid Stemmer for Kokborok	Hybrid Stemmer	In this paper, they provide a hybrid stemming algorithm model for the resource-poor and susceptible Kokborok language, and we evaluate its performance using a number of different techniques.	Recall accuracy = 89.28%. Precision = 89.61%, F1-measure = 89.44 %

|e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2023 ||

| DOI:10.15680/IJIRSET.2023.1205097 |

Table 2: YEAR WISE PUBLICATIONS

Sl. No	Year of Publication	Total research on NLP Kokborok
1	2012	4
2.	2014	2
3.	2015	1
4	2019	1
5	2020	1
6	2021	1
7.	2022	2
8.	2023	1

A. Part of Speech (POS)

A necessary prerequisite for the majority of NLP applications is POS tagging. According to the amount of ongoing research for English social media material, POS tagging of text is now an established technology. Studies on POS tagging for code-mixed languages have, however, not gotten much attention in the Indian context.

B. Named Entity Recognition

One of the most important information extraction tasks, NER focuses on identifying the names of entities, including people, places, organizations, and products. Entity typing and entity detection, or the classification, are the two fundamental NER processes.

C. Sentiment Analysis

Identifying whether a sentence, a passage of text, or a phrase is good, negative, or neutral is the core goal of the research area known as sentiment analysis (SA). One of the most difficult challenges is performing SA activities for performance evaluation in a multilingual setting

D. Language Identification:

One of the most crucial tasks in a code-mixed environment is language identification. While it performs well with monolingual datasets, multilingual datasets are more complicated and draw in high-quality research. To address the issue of language identification in code-mixing and code-switching, researchers apply straightforward dictionary-based approaches to create new tools.

E. Parsing

The process of parsing involves tagging word sequences with the structural grammatical descriptions of sentences. The fundamental reason for parsing is that when the grammatical structure of a Natural Language word sequence is discovered, it aids in determining the sequence's true meaning. Information extraction, semantic analysis, identifying proper name recognition , and machine translation are a few examples of the many uses of parsing.

F. Stemmer

Stemming is the method of reducing a word to its stem, which affixes to suffixes, prefixes, or the "lemmas" (or roots) of words. In both natural language understanding (NLU) and natural language processing (NLP), stemming is essential.

IV. CONCLUSION AND DISCUSSION

Due to the massive rise in undiscovered constructs brought on by the blending of the two or more languages' syntax and thesaurus, Code Mixing, code-switching, Part of Speech, NER, Stemming, is a difficult phenomenon to analyse especially in Kokborok language.

|e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2023 ||

| DOI:10.15680/IJIRSET.2023.1205097 |

This survey report is the first of its type to provide a glance into a variety of applications connected to POS, NER, Morphological analyzer, Stemmer, Code-Mixing and Code-switching, etc datasets in related to low resources languages like Kokborok. It talks about the issues and difficulties that the researchers in this field have to deal with. It highlights Machine Learning (ML) algorithms and NLP strategies for processing the dataset. A few publications that deal with issues that have not yet been fully resolved for the bilingual dataset, such as machine translation, dialect identification, parsing, and automatic speech recognition, are also discussed. Unique assessment metrics are tabulated.

If we see the Table 2 year wish publication in the field of Natural language Processing (NLP) Kokborok language, the research area is reducing day by day and to make it more impact in the field of NLP research, we have to do more research in Kokborok language. This survey papers will help any researchers to get brief highlight in exploring much more deeper in the field of NLP Kokborok language in the coming days in the field of research areas.

REFERENCES

[1] Debbarma, Khumbar, et al. "Morphological analysis of Kokborok for universal networking language dictionary." 2012 1st International Conference on Recent Advances in Information Technology (RAIT). IEEE, 2012

[2] Patra, Braja Gopal, et al. "A light weight stemmer in kokborok." Proceedings of the 24th Conference on Computational Linguistics and Speech Processing (ROCLING 2012). 2012

[3] Patra, Braja Gopal, et al. "Part of Speech (POS) tagger for Kokborok." Proceedings of COLING 2012: Posters. 2012

[4] Debbarma, Khumbar, et al. "Morphological Analyzer for Kokborok." Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing.

[5] Debbarma, Abhijit, B. S. Purkayastha, and Paritosh Bhattacharya. "Stemmer for resource scarce language using string similarity measure." 2014 International Conference on Reliability Optimization and Information Technology (ICROIT). IEEE, 2014

[6] Debbarma, Abhijit, Paritosh Bhattacharya, and B. S. Purkayastha. "Frequency based named entity recognition system for under resource language." 2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT). IEEE, 2014

[7] Sarkar, Partha, and Bipul Syam Purkayastha. "Morphological analyzer in the development of bilingual dictionary (Kokborok-English): an analysis for appropriate method and approach." International Journal of Engineering and Innovative Technology 4.10 (2015)

[8] "Machine translator of Kokborok Language for machine learning using Natural Language Processing, an Universal Networking Language and Dictionary ", International Journal of Emerging Technologies and Innovative Research (www.jetir.org), ISSN:2349-5162, Vol.6, Issue 4, page no.170-174,

[9] Uchoi, Enjula and Lenin Laitonjam. "An Unsupervised Word Level Language Identification of English and Kokborok Code-Mixed and Code-Switched Sentences." *Journal of emerging technologies and innovative research* (2020): n. pag..

[10] Patra, Braja Gopal, et al. "Part of Speech (POS) tagger for Kokborok." *Proceedings of COLING 2012: Posters*. 2012

[11] Dey, Joyeeta. "Adoption of Natural Language Processing in Education: Implications for students."

[12] Debbarma, Pankaj, Pusparwng Hrangkhawl, and Swapan Debbarma. "Modeling a Hybrid Stemmer for Kokborok." 2023 International Conference on Intelligent Systems, Advanced Computing and Communication (ISACC). IEEE, 2023

[13] Debbarma, Pankaj, and Pusparwng Hrangkhawl. "English to Kokborok Bilingual Thesaurus for Natural Language Processing."

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.423

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details